Mathematics for Physics II

A set of lecture notes by

Michael Stone

PIMANDER-CASAUBON
Alexandria • Florence • London

Preface

These notes cover the material from the second half of a two-semester sequence of mathematical methods courses given to first year physics graduate students at the University of Illinois. They consist of three loosely connected parts: i) an introduction to modern "calculus on manifolds", the exterior differential calculus, and algebraic topology; ii) an introduction to group representation theory and its physical applications; iii) a fairly standard course on complex variables.

Contents

Chapter 1

Tensors in Euclidean Space

In this chapter we explain how a vector space V gives rise to a family of associated tensor spaces, and how mathematical objects such as linear maps or quadratic forms should be understood as being elements of these spaces. We then apply these ideas to physics. We make extensive use of notions and notations from the appendix on linear algebra, so it may help to review that material before we begin.

1.1 Covariant and Contravariant Vectors

When we have a vector space V over $\mathbb{R}$, and $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ and $\{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n\}$ are both bases for V, then we may expand each of the basis vectors $\mathbf{e}_\mu$ in terms of the $\mathbf{e}'_\mu$ as

$$\mathbf{e}_\nu = a^\mu_\nu \mathbf{e}'_\mu. \tag{1.1}$$

We are here, as usual, using the Einstein summation convention that repeated indices are to be summed over. Written out in full for a three-dimensional space, the expansion would be

$$\begin{aligned} \mathbf{e}_1 &= a^1_1 \mathbf{e}'_1 + a^2_1 \mathbf{e}'_2 + a^3_1 \mathbf{e}'_3, \\ \mathbf{e}_2 &= a^1_2 \mathbf{e}'_1 + a^2_2 \mathbf{e}'_2 + a^3_2 \mathbf{e}'_3, \\ \mathbf{e}_3 &= a^1_3 \mathbf{e}'_1 + a^2_3 \mathbf{e}'_2 + a^3_3 \mathbf{e}'_3. \end{aligned}$$

We could also have expanded the $\mathbf{e}'_\mu$ in terms of the $\mathbf{e}_\mu$ as

$$\mathbf{e}'_\nu = (a^{-1})^\mu_\nu \mathbf{e}'_\mu. \tag{1.2}$$

As the notation implies, the matrices of coefficients a^μ_ν and $(a^{-1})^\mu_\nu$ are inverses of each other:

$$a^\mu_\nu (a^{-1})^\nu_\sigma = (a^{-1})^\mu_\nu a^\nu_\sigma = \delta^\mu_\sigma. \tag{1.3}$$

If we know the components x^μ of a vector $\mathbf{x}$ in the $\mathbf{e}_\mu$ basis then the components x'^μ of $\mathbf{x}$ in the $\mathbf{e}'_\mu$ basis are obtained from

$$\mathbf{x} = x'^\mu \mathbf{e}'_\mu = x^\nu \mathbf{e}_\nu = (x^\nu a^\mu_\nu)\, \mathbf{e}'_\mu \tag{1.4}$$

by comparing the coefficients of $\mathbf{e}'_\mu$. We find that $x'^\mu = a^\mu_\nu x^\nu$. Observe how the $\mathbf{e}_\mu$ and the x^μ transform in "opposite" directions. The components x^μ are therefore said to transform *contravariantly*.

Associated with the vector space V is its *dual space* V^*, whose elements are *covectors*, *i.e.* linear maps $\mathbf{f} : V \to \mathbb{R}$. If $\mathbf{f} \in V^*$ and $\mathbf{x} = x^\mu \mathbf{e}_\mu$, we use the linearity property to evaluate $\mathbf{f}(\mathbf{x})$ as

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x^\mu \mathbf{e}_\mu) = x^\mu \mathbf{f}(\mathbf{e}_\mu) = x^\mu\, f_\mu. \tag{1.5}$$

Here, the set of numbers $f_\mu = \mathbf{f}(\mathbf{e}_\mu)$ are the components of the covector $\mathbf{f}$. If we change basis so that $\mathbf{e}_\nu = a^\mu_\nu \mathbf{e}'_\mu$ then

$$f_\nu = \mathbf{f}(\mathbf{e}_\nu) = \mathbf{f}(a^\mu_\nu \mathbf{e}'_\mu) = a^\mu_\nu \mathbf{f}(\mathbf{e}'_\mu) = a^\mu_\nu f'_\mu. \tag{1.6}$$

We conclude that $f_\nu = a^\mu_\nu f'_\mu$. The f_μ components transform in the same manner as the basis. They are therefore said to transform *covariantly*. In physics it is traditional to call the the set of numbers x^μ with upstairs indices (the components of) a *contravariant vector*. Similarly, the set of numbers f_μ with downstairs indices is called (the components of) a *covariant vector*. Thus, contravariant vectors are elements of V and covariant vectors are elements of V^*.

The relationship between V and V^* is one of mutual duality, and to mathematicians it is only a matter of convenience which space is V and which space is V^*. The evaluation of $\mathbf{f} \in V^*$ on $\mathbf{x} \in V$ is therefore often written as a "pairing" $(\mathbf{f}, \mathbf{x})$, which gives equal status to the objects being put togther to get a number. A physics example of such a mutually dual pair is provided by the space of displacements $\mathbf{x}$ and the space of wave-numbers $\mathbf{k}$. The units of $\mathbf{x}$ and $\mathbf{k}$ are different (meters *versus* meters^{-1}). There is therefore no meaning to "$\mathbf{x} + \mathbf{k}$," and $\mathbf{x}$ and $\mathbf{k}$ are not elements of the same vector space. The "dot" in expressions such as

$$\psi(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}} \tag{1.7}$$

cannot be a true inner product (which requires the objects it links to be in the same vector space) but is instead a pairing

$$(\mathbf{k}, \mathbf{x}) \equiv \mathbf{k}(\mathbf{x}) = k_\mu x^\mu. \tag{1.8}$$

In describing the physical world we usually give priority to the space in which we live, breathe and move, and so treat it as being "V". The displacement vector $\mathbf{x}$ then becomes the contravariant vector, and the Fourier-space wave-number $\mathbf{k}$, being the more abstract quantity, becomes the covariant covector.

Our vector space may come equipped with a *metric* that is derived from a non-degenerate inner product. We regard the inner product as being a bilinear form $\mathbf{g} : V \times V \to \mathbb{R}$, so the length $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is $\sqrt{\mathbf{g}(\mathbf{x}, \mathbf{x})}$. The set of numbers

$$g_{\mu\nu} = \mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) \tag{1.9}$$

comprises the (components of) the *metric tensor*. In terms of them, the inner of product $\langle \mathbf{x}, \mathbf{y} \rangle$ of pair of vectors $\mathbf{x} = x^\mu \mathbf{e}_\mu$ and $\mathbf{y} = y^\mu \mathbf{e}_\mu$ becomes

$$\langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{g}(\mathbf{x}, \mathbf{y}) = g_{\mu\nu} x^\mu y^\nu. \tag{1.10}$$

Real-valued inner products are always symmetric, so $\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{g}(\mathbf{y}, \mathbf{x})$ and $g_{\mu\nu} = g_{\nu\mu}$. As the product is non-degenerate, the matrix $g_{\mu\nu}$ has an inverse, which is traditionally written as $g^{\mu\nu}$. Thus

$$g_{\mu\nu} g^{\nu\lambda} = g^{\lambda\nu} g_{\nu\mu} = \delta^\lambda_\mu. \tag{1.11}$$

The additional structure provided by the metric permits us to identify V with V^*. The identification is possible, because, given any $\mathbf{f} \in V^*$, we can find a vector $\widetilde{\mathbf{f}} \in V$ such that

$$\mathbf{f}(\mathbf{x}) = \langle \widetilde{\mathbf{f}}, \mathbf{x} \rangle. \tag{1.12}$$

We obtain $\widetilde{\mathbf{f}}$ by solving the equation

$$f_\mu = g_{\mu\nu} \widetilde{f}^\nu \tag{1.13}$$

to get $\widetilde{f}^\nu = g^{\nu\mu} f_\mu$. We may now drop the tilde and identify $\mathbf{f}$ with $\widetilde{\mathbf{f}}$, and hence V with V^*. When we do this, we say that the covariant components f_μ are related to the contravariant components f^μ by *raising*

$$f^\mu = g^{\mu\nu} f_\nu, \tag{1.14}$$

or *lowering*

$$f_\mu = g_{\mu\nu} f^\nu, \tag{1.15}$$

the index μ using the metric tensor. Bear in mind that this $V \cong V^*$ identification depends crucially on the metric. A different metric will, in general, identify an $\mathbf{f} \in V^*$ with a completely different $\widetilde{\mathbf{f}} \in V$.

We may play this game in the Euclidean space $\mathbb{E}^n$ with its "dot" inner product. Given a vector $\mathbf{x}$ and a basis $\mathbf{e}_\mu$ for which $g_{\mu\nu} = \mathbf{e}_\mu \cdot \mathbf{e}_\nu$, we can define two sets of components for the same vector. Firstly the coefficients x^μ appearing in the basis expansion

$$\mathbf{x} = x^\mu \mathbf{e}_\mu, \tag{1.16}$$

and secondly the "components"

$$x_\mu = \mathbf{e}_\mu \cdot \mathbf{x} = \mathbf{g}(\mathbf{e}_\mu, \mathbf{x}) = \mathbf{g}(\mathbf{e}_\mu, x^\nu \mathbf{e}_\nu) = \mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) x^\nu = g_{\mu\nu} x^\nu \tag{1.17}$$

of $\mathbf{x}$ along the basis vectors. These two set of numbers are then respectively called the contravariant and covariant components of the vector $\mathbf{x}$. If the $\mathbf{e}_\mu$ constitute an orthonormal basis, where $g_{\mu\nu} = \delta_{\mu\nu}$, then the two sets of components (covariant and contravariant) are numerically coincident. In a non-orthogonal basis they will be different, and we must take care never to add contravariant components to covariant ones.

1.2 Tensors

We now introduce tensors in two ways: firstly as sets of numbers labelled by indices and equipped with transformation laws that tell us how these numbers change as we change basis; and secondly as basis-independent objects that are elements of a vector space constructed by taking multiple tensor products of the spaces V and V^*.

1.2.1 Transformation rules

After we change basis $\mathbf{e}_\mu \to \mathbf{e}'_\mu$, where $\mathbf{e}_\nu = a^\mu_\nu \mathbf{e}'_\mu$, the metric tensor will be represented by a new set of components

$$g'_{\mu\nu} = \mathbf{g}(\mathbf{e}'_\mu, \mathbf{e}'_\nu). \tag{1.18}$$

These are be related to the old components by

$$g_{\mu\nu} = \mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) = \mathbf{g}(a^\rho_\mu \mathbf{e}'_\rho, a^\sigma_\nu \mathbf{e}'_\sigma) = a^\rho_\mu a^\sigma_\nu \mathbf{g}(\mathbf{e}'_\rho, \mathbf{e}'_\sigma) = a^\rho_\mu a^\sigma_\nu \, g'_{\rho\sigma}. \tag{1.19}$$

This transformation rule for $g_{\mu\nu}$ has both of its subscripts behaving like the downstairs indices of a covector. We therefore say that $g_{\mu\nu}$ transforms as a *doubly covariant tensor*. Written out in full, for a two-dimensional space, the transformation law is

$$\begin{aligned}
g_{11} &= a^1_1 a^1_1 g'_{11} + a^1_1 a^2_1 g'_{12} + a^2_1 a^1_1 g'_{21} + a^2_1 a^2_1 g'_{22}, \\
g_{12} &= a^1_1 a^1_2 g'_{11} + a^1_1 a^2_2 g'_{12} + a^2_1 a^1_2 g'_{21} + a^2_1 a^2_2 g'_{22}, \\
g_{21} &= a^1_2 a^1_1 g'_{11} + a^1_2 a^2_1 g'_{12} + a^2_2 a^1_1 g'_{21} + a^2_2 a^2_1 g'_{22}, \\
g_{22} &= a^1_2 a^1_2 g'_{11} + a^1_2 a^2_2 g'_{12} + a^2_2 a^1_2 g'_{21} + a^2_2 a^2_2 g'_{22}.
\end{aligned}$$

In three dimensions each row would have nine terms, and sixteen in four dimensions. We see why Einstein was driven to invent his summation convention!

A set of numbers $Q^{\alpha\beta}{}_{\gamma\delta\epsilon}$, whose indices range from 1 to the dimension of the space and that transforms as

$$Q^{\alpha\beta}{}_{\gamma\delta\epsilon} = (a^{-1})^\alpha_{\alpha'} (a^{-1})^\beta_{\beta'} \, a^{\gamma'}_\gamma \, a^{\delta'}_\delta \, a^{\epsilon'}_\epsilon Q'^{\alpha'\beta'}{}_{\gamma'\delta'\epsilon'}, \tag{1.20}$$

or conversely as

$$Q'^{\alpha\beta}{}_{\gamma\delta\epsilon} = a^\alpha_{\alpha'} a^\beta_{\beta'} (a^{-1})^{\gamma'}_\gamma (a^{-1})^{\delta'}_\delta (a^{-1})^{\epsilon'}_\epsilon Q^{\alpha'\beta'}{}_{\gamma'\delta'\epsilon'}, \tag{1.21}$$

comprises the components of a *doubly contravariant, triply covariant* tensor. More compactly, the $Q^{\alpha\beta}{}_{\gamma\delta\epsilon}$ are the components of a tensor of type $(2, 3)$. Tensors of type (p, q) are defined analogously. The total number of indices $p + q$ is called the *rank* of the tensor.

Note how the indices are wired up in the transformation rules (1.20) and (1.21): free (not summed over) upstairs indices on the left hand side of the equations match to free upstairs indices on the right hand side, similarly for the downstairs indices. Also upstairs indices are summed only with downstairs ones.

Similar conditions apply to equations relating tensors in any particular basis. If they are violated you do not have a valid tensor equation — meaning that an equation valid in one basis will not be valid in another basis. Thus an equation

$$A^\mu{}_{\nu\lambda} = B^{\mu\tau}{}_{\nu\lambda\tau} + C^\mu{}_{\nu\lambda} \tag{1.22}$$

is fine, but

$$A^{\mu}{}_{\nu\lambda} \stackrel{?}{=} B^{\nu}{}_{\mu\lambda} + C^{\mu}{}_{\nu\lambda\sigma\sigma} + D^{\mu}{}_{\nu\lambda\tau} \tag{1.23}$$

has something wrong in each term.

Incidentally, although not illegal, it is a good idea not to write tensor indices directly underneath one another — *i.e.* do not write Q^{ij}_{kjl} — because if you raise or lower indices using the metric tensor, and some pages later in a calculation try to put them back where they were, they might end up in the wrong order.

Tensor algebra

The sum of two tensors of a given type is also a tensor of that type. The sum of two tensors of different types is not a tensor. Thus each particular type of tensor constitutes a distinct vector space, but one derived from the common underlying vector space whose change-of-basis formula is being utilized.

Tensors can be combined by multiplication: if $A^{\mu}{}_{\nu\lambda}$ and $B^{\mu}{}_{\nu\lambda\tau}$ are tensors of type $(1,2)$ and $(1,3)$ respectively, then

$$C^{\alpha\beta}{}_{\nu\lambda\rho\sigma\tau} = A^{\alpha}{}_{\nu\lambda} B^{\beta}{}_{\rho\sigma\tau} \tag{1.24}$$

is a tensor of type $(2,5)$.

An important operation is *contraction*, which consists of setting one or more contravariant index index equal to a covariant index and summing over the repeated indices. This reduces the rank of the tensor. So, for example,

$$D_{\rho\sigma\tau} = C^{\alpha\beta}{}_{\alpha\beta\rho\sigma\tau} \tag{1.25}$$

is a tensor of type $(0,3)$. Similarly $\mathbf{f}(\mathbf{x}) = f_\mu x^\mu$ is a type $(0,0)$ tensor, *i.e.* an *invariant* — a number that takes the same value in all bases. Upper indices can only be contracted with lower indices, and *vice versa*. For example, the array of numbers $A_\alpha = B_{\alpha\beta\beta}$ obtained from the type $(0,3)$ tensor $B_{\alpha\beta\gamma}$ is *not* a tensor of type $(0,1)$.

The contraction procedure outputs a tensor because setting an upper index and a lower index to a common value μ and summing over μ, leads to the factor $\ldots (a^{-1})^{\mu}_{\alpha} a^{\beta}_{\mu} \ldots$ appearing in the transformation rule. Now

$$(a^{-1})^{\mu}_{\alpha} a^{\beta}_{\mu} = \delta^{\beta}_{\alpha}, \tag{1.26}$$

and the Kronecker delta effects a summation over the corresponding pair of indices in the transformed tensor.

Although often associated with general relativity, tensors occur in many places in physics. They are used, for example, in elasticity theory, where the word "tensor" in its modern meaning was introduced by Woldemar Voigt in 1898. Voigt, following Cauchy and Green, described the infinitesimal deformation of an elastic body by the *strain tensor* $e_{\alpha\beta}$, which is a tensor of type (0,2). The forces to which the strain gives rise are described by the *stress tensor* $\sigma^{\lambda\mu}$. A generalization of Hooke's law relates stress to strain via a tensor of elastic constants $c^{\alpha\beta\gamma\delta}$ as

$$\sigma^{\alpha\beta} = c^{\alpha\beta\gamma\delta} e_{\gamma\delta}. \tag{1.27}$$

We study stress and strain in more detail later in this chapter.

Exercise 1.1: Show that $g^{\mu\nu}$, the matrix inverse of the metric tensor $g_{\mu\nu}$, is indeed a doubly contravariant tensor, as the position of its indices suggests.

1.2.2 Tensor character of linear maps and quadratic forms

As an illustration of the tensor concept and of the need to distinguish between upstairs and downstairs indices, we contrast the properties of matrices representing linear maps and those representing quadratic forms.

A linear map $M : V \to V$ is an object that exists independently of any basis. Given a basis, however, it is represented by a matrix $M^{\mu}{}_{\nu}$ obtained by examining the action of the map on the basis elements:

$$M(\mathbf{e}_\mu) = \mathbf{e}_\nu M^{\nu}{}_{\mu}. \tag{1.28}$$

Acting on $\mathbf{x}$ we get a new vector $\mathbf{y} = M(\mathbf{x})$, where

$$y^\nu \mathbf{e}_\nu = \mathbf{y} = M(\mathbf{x}) = M(x^\mu \mathbf{e}_\mu) = x^\mu M(\mathbf{e}_\mu) = x^\mu M^{\nu}{}_{\mu} \mathbf{e}_\nu = M^{\nu}{}_{\mu} x^\mu \, \mathbf{e}_\nu. \tag{1.29}$$

We therefore have

$$y^\nu = M^{\nu}{}_{\mu} x^\mu, \tag{1.30}$$

which is the usual matrix multiplication $\mathbf{y} = \mathbf{M}\mathbf{x}$. When we change basis, $\mathbf{e}_\nu = a^\mu_\nu \mathbf{e}'_\mu$, then

$$\mathbf{e}_\nu M^{\nu}{}_{\mu} = M(\mathbf{e}_\mu) = M(a^\rho_\mu \mathbf{e}'_\rho) = a^\rho_\mu M(\mathbf{e}'_\rho) = a^\rho_\mu \mathbf{e}'_\sigma M'^{\sigma}{}_{\rho} = a^\rho_\mu (a^{-1})^\nu_\sigma \mathbf{e}_\nu M'^{\sigma}{}_{\rho}. \tag{1.31}$$

Comparing coefficients of $\mathbf{e}_\nu$, we find

$$M'^{\nu}{}_{\mu} = a^{\rho}_{\mu}(a^{-1})^{\nu}_{\sigma} M'^{\sigma}{}_{\rho}, \tag{1.32}$$

or, conversely,

$$M'^{\nu}{}_{\mu} = (a^{-1})^{\rho}_{\mu} a^{\nu}_{\sigma} M^{\sigma}{}_{\rho}. \tag{1.33}$$

Thus a matrix representing a linear map has the tensor character suggested by the position of its indices, *i.e.* it transforms as a type $(1,1)$ tensor. We can derive the same formula in matrix notation. In the new basis the vectors $\mathbf{x}$ and $\mathbf{y}$ have new components $\mathbf{x}' = \mathbf{A}\mathbf{x}$, and $\mathbf{y}' = \mathbf{A}\mathbf{y}$. Consequently $\mathbf{y} = \mathbf{M}\mathbf{x}$ becomes

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{M}\mathbf{x} = \mathbf{A}\mathbf{M}\mathbf{A}^{-1}\mathbf{x}', \tag{1.34}$$

and the matrix representing the map M has new components

$$\mathbf{M}' = \mathbf{A}\mathbf{M}\mathbf{A}^{-1}. \tag{1.35}$$

Now consider the quadratic form $Q : V \to \mathbb{R}$ that is obtained from a symmetric bilinear form $Q : V \times V \to \mathbb{R}$ by setting $Q(\mathbf{x}) = Q(\mathbf{x}, \mathbf{x})$. We can write

$$Q(\mathbf{x}) = Q_{\mu\nu} x^\mu x^\nu = x^\mu Q_{\mu\nu}\, x^\nu = \mathbf{x}^T \mathbf{Q} \mathbf{x}, \tag{1.36}$$

where $Q_{\mu\nu} \equiv Q(\mathbf{e}_\mu, \mathbf{e}_\nu)$ are the entries in the symmetric matrix $\mathbf{Q}$, the suffix T denotes transposition, and $\mathbf{x}^T\mathbf{Q}\mathbf{x}$ is standard matrix-multiplication notation. Just as does the metric tensor, the coefficients $Q_{\mu\nu}$ transform as a type $(0,2)$ tensor:

$$Q_{\mu\nu} = a^{\alpha}_{\mu} a^{\beta}_{\nu} Q'_{\alpha\beta}. \tag{1.37}$$

In matrix notation the vector $\mathbf{x}$ again transforms to have new components $\mathbf{x}' = \mathbf{A}\mathbf{x}$, but $\mathbf{x}'^T = \mathbf{x}^T\mathbf{A}^T$. Consequently

$$\mathbf{x}'^T \mathbf{Q}' \mathbf{x}' = \mathbf{x}^T \mathbf{A}^T \mathbf{Q}' \mathbf{A} \mathbf{x}. \tag{1.38}$$

Thus

$$\mathbf{Q} = \mathbf{A}^T \mathbf{Q}' \mathbf{A}. \tag{1.39}$$

The message is that linear maps and quadratic forms can both be represented by matrices, but these matrices correspond to distinct types of tensor and transform differently under a change of basis.

A matrix representing a linear map has a basis-independent determinant. Similarly the *trace* of a matrix representing a linear map

$$\operatorname{tr} \mathbf{M} \stackrel{\text{def}}{=} M^{\mu}{}_{\mu} \tag{1.40}$$

is a tensor of type $(0,0)$, i.e. a scalar, and therefore basis independent. On the other hand, while you can certainly compute the determinant or the trace of the matrix representing a quadratic form in some particular basis, when you change basis and calculate the determinant or trace of the transformed matrix, you will get a different number.

It *is* possible to make a quadratic form out of a linear map, but this requires using the metric to lower the contravariant index on the matrix representing the map:

$$Q(\mathbf{x}) = x^\mu g_{\mu\nu} Q^\nu{}_\lambda x^\lambda = \mathbf{x} \cdot \mathbf{Q}\mathbf{x}. \tag{1.41}$$

Be careful, therefore: the matrices "$\mathbf{Q}$" in $\mathbf{x}^T\mathbf{Q}\mathbf{x}$ and in $\mathbf{x}\cdot\mathbf{Q}\mathbf{x}$ are representing different mathematical objects.

Exercise 1.2: In this problem we will use the distinction between the transformation law of a quadratic form and that of a linear map to resolve the following "paradox":

- In quantum mechanics we are taught that the matrices representing two operators can be simultaneously diagonalized only if they commute.
- In classical mechanics we are taught how, given the Lagrangian

$$L = \sum_{ij} \left(\frac{1}{2} \dot{q}_i M_{ij} \dot{q}_j - \frac{1}{2} q_i V_{ij} q_j \right),$$

to construct normal co-ordinates Q_i such that L becomes

$$L = \sum_i \left(\frac{1}{2} \dot{Q}_i^2 - \frac{1}{2} \omega_i^2 Q_i^2 \right).$$

We have apparantly managed to simultaneously diagonize the matrices $M_{ij} \to \mathrm{diag}\,(1,\ldots,1)$ and $V_{ij} \to \mathrm{diag}\,(\omega_1^2,\ldots,\omega_n^2)$, even though there is no reason for them to commute with each other!

Show that when $\mathbf{M}$ and $\mathbf{V}$ are a pair of symmetric matrices, with $\mathbf{M}$ being positive definite, then there exits an invertible matrix $\mathbf{A}$ such that $\mathbf{A}^T\mathbf{M}\mathbf{A}$ and $\mathbf{A}^T\mathbf{V}\mathbf{A}$ are simultaneously diagonal. (Hint: Consider $\mathbf{M}$ as defining an inner product, and use the Gramm-Schmidt procedure to first find a orthonormal frame in which $M'_{ij} = \delta_{ij}$. Then show that the matrix corresponding to $\mathbf{V}$ in this frame can be diagonalized by a further transformation that does not perturb the already diagonal M'_{ij}.)

1.2.3 Tensor product spaces

We may regard the set of numbers $Q^{\alpha\beta}{}_{\gamma\delta\epsilon}$ as being the components of an object $\mathbf{Q}$ that is element of the vector space of type $(2,3)$ tensors. We denote this vector space by the symbol $V \otimes V \otimes V^* \otimes V^* \otimes V^*$, the notation indicating that it is derived from the original V and its dual V^* by taking *tensor products* of these spaces. The tensor $\mathbf{Q}$ is to be thought of as existing as an element of $V \otimes V \otimes V^* \otimes V^* \otimes V^*$ independently of any basis, but given a basis $\{\mathbf{e}_\mu\}$ for V, and the dual basis $\{e^{*\nu}\}$ for V^*, we expand it as

$$\mathbf{Q} = Q^{\alpha\beta}{}_{\gamma\delta\epsilon}\, \mathbf{e}_\alpha \otimes \mathbf{e}_\beta \otimes \mathbf{e}^{*\gamma} \otimes \mathbf{e}^{*\delta} \otimes \mathbf{e}^{*\epsilon}. \tag{1.42}$$

Here the tensor product symbol "$\otimes$" is distributive

$$\begin{aligned} \mathbf{a} \otimes (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \otimes \mathbf{b} + \mathbf{a} \otimes \mathbf{c}, \\ (\mathbf{a} + \mathbf{b}) \otimes \mathbf{c} &= \mathbf{a} \otimes \mathbf{c} + \mathbf{b} \otimes \mathbf{c}, \end{aligned} \tag{1.43}$$

and associative

$$(\mathbf{a} \otimes \mathbf{b}) \otimes \mathbf{c} = \mathbf{a} \otimes (\mathbf{b} \otimes \mathbf{c}), \tag{1.44}$$

but is not commutative

$$\mathbf{a} \otimes \mathbf{b} \neq \mathbf{b} \otimes \mathbf{a}. \tag{1.45}$$

Everything commutes with the field, however,

$$\lambda(\mathbf{a} \otimes \mathbf{b}) = (\lambda \mathbf{a}) \otimes \mathbf{b} = \mathbf{a} \otimes (\lambda \mathbf{b}). \tag{1.46}$$

If we change basis $\mathbf{e}_\alpha = a^\beta_\alpha \mathbf{e}'_\beta$ then these rules lead, for example, to

$$\mathbf{e}_\alpha \otimes \mathbf{e}_\beta = a^\lambda_\alpha a^\mu_\beta\, \mathbf{e}'_\lambda \otimes \mathbf{e}'_\mu. \tag{1.47}$$

From this change-of-basis formula, we deduce that

$$T^{\alpha\beta} \mathbf{e}_\alpha \otimes \mathbf{e}_\beta = T^{\alpha\beta} a^\lambda_\alpha a^\mu_\beta\, \mathbf{e}'_\lambda \otimes \mathbf{e}'_\mu = T'^{\lambda\mu}\, \mathbf{e}'_\lambda \otimes \mathbf{e}'_\mu, \tag{1.48}$$

where

$$T'^{\lambda\mu} = T^{\alpha\beta} a^\lambda_\alpha a^\mu_\beta. \tag{1.49}$$

The analogous formula for $\mathbf{e}_\alpha \otimes \mathbf{e}_\beta \otimes \mathbf{e}^{*\gamma} \otimes \mathbf{e}^{*\delta} \otimes \mathbf{e}^{*\epsilon}$ reproduces the transformation rule for the components of $\mathbf{Q}$.

The meaning of the tensor product of a collection of vector spaces should now be clear: If $\mathbf{e}_\mu$ consititute a basis for V, the space $V \otimes V$ is, for example,

the space of all linear combinations[1] of the abstract symbols $\mathbf{e}_\mu \otimes \mathbf{e}_\nu$, which we declare by *fiat* to constitute a basis for this space. There is no geometric significance (as there is with a vector product $\mathbf{a} \times \mathbf{b}$) to the tensor product $\mathbf{a} \otimes \mathbf{b}$, so the $\mathbf{e}_\mu \otimes \mathbf{e}_\nu$ are simply useful place-keepers. Remember that these are *ordered* pairs, $\mathbf{e}_\mu \otimes \mathbf{e}_\nu \neq \mathbf{e}_\nu \otimes \mathbf{e}_\mu$.

Although there is no *geometric* meaning, it is possible, however, to give an *algebraic* meaning to a product like $\mathbf{e}^{*\lambda} \otimes \mathbf{e}^{*\mu} \otimes \mathbf{e}^{*\nu}$ by viewing it as a multilinear form $V \times V \times V :\rightarrow \mathbb{R}$. We define

$$\mathbf{e}^{*\lambda} \otimes \mathbf{e}^{*\mu} \otimes \mathbf{e}^{*\nu} \left(\mathbf{e}_\alpha, \mathbf{e}_\beta, \mathbf{e}_\gamma\right) = \delta^\lambda_\alpha \, \delta^\mu_\beta \, \delta^\nu_\gamma. \tag{1.50}$$

We may also regard it as a linear map $V \otimes V \otimes V :\rightarrow \mathbb{R}$ by defining

$$\mathbf{e}^{*\lambda} \otimes \mathbf{e}^{*\mu} \otimes \mathbf{e}^{*\nu} \left(\mathbf{e}_\alpha \otimes \mathbf{e}_\beta \otimes \mathbf{e}_\gamma\right) = \delta^\lambda_\alpha \, \delta^\mu_\beta \, \delta^\nu_\gamma \tag{1.51}$$

and extending the definition to general elements of $V \otimes V \otimes V$ by linearity. In this way we establish an isomorphism

$$V^* \otimes V^* \otimes V^* \cong (V \otimes V \otimes V)^*. \tag{1.52}$$

This multiple personality is typical of tensor spaces. We have already seen that the metric tensor is simultaneously an element of $V^* \otimes V^*$ and a map $\mathbf{g} : V \rightarrow V^*$.

Tensor products and quantum mechanics

When we have two quantum-mechanical systems having Hilbert spaces $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, the Hilbert space for the combined system is $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$. Quantum mechanics books usually denote the vectors in these spaces by the Dirac "bra-ket" notation in which the basis vectors of the separate spaces are denoted by[2] $|n_1\rangle$ and $|n_2\rangle$, and that of the combined space by $|n_1, n_2\rangle$. In this notation, a state in the combined system is a linear combination

$$|\Psi\rangle = \sum_{n_1, n_2} |n_1, n_2\rangle\langle n_1, n_2|\Psi\rangle, \tag{1.53}$$

[1] Do not confuse the tensor-product space $V \otimes W$ with the Cartesian product $V \times W$. The latter is the set of all ordered pairs $(\mathbf{x}, \mathbf{y})$, $\mathbf{x} \in V$, $\mathbf{y} \in W$. The tensor product includes also *formal sums* of such pairs. The Cartesian product of two vector spaces can be given the structure of a vector space by defining an addition operation $\lambda(\mathbf{x}_1, \mathbf{y}_1) + \mu(\mathbf{x}_2, \mathbf{y}_2) = (\lambda\mathbf{x}_1 + \mu\mathbf{x}_2, \lambda\mathbf{y}_1 + \mu\mathbf{y}_2)$, but this construction does not lead to the tensor product. Instead it defines the *direct sum* $V \oplus W$.

[2] We assume for notational convenience that the Hilbert spaces are finite dimensional.

This is the tensor product in disguise. To unmask it, we simply make the notational translation

$$\begin{aligned} |\Psi\rangle &\rightarrow \mathbf{\Psi} \\ \langle n_1, n_2|\Psi\rangle &\rightarrow \psi^{n_1,n_2} \\ |n_1\rangle &\rightarrow \mathbf{e}^{(1)}_{n_1} \\ |n_2\rangle &\rightarrow \mathbf{e}^{(2)}_{n_2} \\ |n_1, n_2\rangle &\rightarrow \mathbf{e}^{(1)}_{n_1} \otimes \mathbf{e}^{(2)}_{n_2}. \end{aligned} \tag{1.54}$$

Then (1.53) becomes

$$\mathbf{\Psi} = \psi^{n_1,n_2}\, \mathbf{e}^{(1)}_{n_1} \otimes \mathbf{e}^{(2)}_{n_2}. \tag{1.55}$$

Entanglement: Suppose that $\mathcal{H}^{(1)}$ has basis $\mathbf{e}^{(1)}_1, \ldots, \mathbf{e}^{(1)}_m$ and $\mathcal{H}^{(2)}$ has basis $\mathbf{e}^{(2)}_1, \ldots, \mathbf{e}^{(2)}_n$. The Hilbert space $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$ is then nm dimensional. Consider a state

$$\mathbf{\Psi} = \psi^{ij} \mathbf{e}^{(1)}_i \otimes \mathbf{e}^{(2)}_j \in \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}. \tag{1.56}$$

If we can find vectors

$$\begin{aligned} \mathbf{\Phi} &\equiv \phi^i \mathbf{e}^{(1)}_i \in \mathcal{H}^{(1)}, \\ \mathbf{X} &\equiv \chi^j \mathbf{e}^{(2)}_j \in \mathcal{H}^{(2)}, \end{aligned} \tag{1.57}$$

such that

$$\mathbf{\Psi} = \mathbf{\Phi} \otimes \mathbf{X} \equiv \phi^i \chi^j \mathbf{e}^{(1)}_i \otimes \mathbf{e}^{(2)}_j \tag{1.58}$$

then the tensor $\mathbf{\Psi}$ is said to be *decomposable* and the two quantum systems are said to be *unentangled*. If there are no such vectors then the two systems are *entangled* in the sense of the Einstein-Podolski-Rosen (EPR) paradox.

Quantum states are really in one-to-one correspondence with *rays* in the Hilbert space, rather than vectors. If we denote the n dimensional vector space over the field of the complex numbers as $\mathbb{C}^n$, the space of rays, in which we do not distinguish between the vectors $\mathbf{x}$ and $\lambda\mathbf{x}$ when $\lambda \neq 0$, is denoted by $\mathbb{C}P^{n-1}$ and is called *complex projective space*. Complex projective space is where *algebraic geometry* is studied. The set of decomposable states may be thought of as a subset of the complex projective space $\mathbb{C}P^{nm-1}$, and, since, as the following excercise shows, this subset is defined by a finite number of homogeneous polynomial equations, it forms what algebraic geometers call a *variety*. This particular subset is known as the *Segre variety*.

Exercise 1.3: The Segre conditions for a state to be decomposable:

i) By counting the number of independent components that are at our disposal in $\mathbf{\Psi}$, and comparing that number with the number of free parameters in $\mathbf{\Phi} \otimes \mathbf{X}$, show that the coefficients ψ^{ij} must satisfy $(n-1)(m-1)$ relations if the state is to be decomposable.

ii) If the state is decomposable, show that

$$0 = \begin{vmatrix} \psi^{ij} & \psi^{il} \\ \psi^{kj} & \psi^{kl} \end{vmatrix}$$

for all sets of indices i, j, k, l.

iii) Assume that ψ^{11} is not zero. Using your count from part (i) as a guide, find a subset of the relations from part (ii) that constitute a necessary and sufficient set of conditions for the state Ψ to be decomposable. Include a proof that your set is indeed sufficient.

1.2.4 Symmetric and skew-symmetric tensors

By examining the transformation rule you may see that if a pair of upstairs or downstairs indices is *symmetric* (say $Q^{\mu\nu}{}_{\rho\sigma\tau} = Q^{\nu\mu}{}_{\rho\sigma\tau}$) or *skew-symmetric* ($Q^{\mu\nu}{}_{\rho\sigma\tau} = -Q^{\nu\mu}{}_{\rho\sigma\tau}$) in one basis, it remains so after the basis has been changed. (This is **not** true of a pair composed of one upstairs and one downstairs index.) It makes sense, therefore, to define symmetric and skew-symmetric tensor product spaces. Thus skew-symmetric doubly-contravariant tensors can be regarded as belonging to the space denoted by $\bigwedge^2 V$ and expanded as

$$\mathbf{A} = \frac{1}{2} A^{\mu\nu}\, \mathbf{e}_\mu \wedge \mathbf{e}_\nu, \tag{1.59}$$

where the coefficients are skew-symmetric, $A^{\mu\nu} = -A^{\nu\mu}$, and the *wedge product* of the basis elements is associative and distributive, as is the tensor product, but in addition obeys $\mathbf{e}_\mu \wedge \mathbf{e}_\nu = -\mathbf{e}_\nu \wedge \mathbf{e}_\mu$. The "1/2" (replaced by $1/p!$ when there are p indices) is convenient in that each independent component only appears once in the sum. For example, in three dimensions,

$$\frac{1}{2} A^{\mu\nu}\, \mathbf{e}_\mu \wedge \mathbf{e}_\nu = A^{12}\, \mathbf{e}_1 \wedge \mathbf{e}_2 + A^{23}\, \mathbf{e}_2 \wedge \mathbf{e}_3 + A^{31}\, \mathbf{e}_3 \wedge \mathbf{e}_1. \tag{1.60}$$

Symmetric doubly-contravariant tensors can be regarded as belonging to the space $\mathrm{sym}^2 V$ and expanded as

$$\mathbf{S} = S^{\alpha\beta}\, \mathbf{e}_\alpha \odot \mathbf{e}_\beta \tag{1.61}$$

where $\mathbf{e}_\alpha \odot \mathbf{e}_\beta = \mathbf{e}_\beta \odot \mathbf{e}_\alpha$ and $S^{\alpha\beta} = S^{\beta\alpha}$. (We do not insert a "1/2" here because including it leads to no particular simplification in any consequent equations.)

We can treat these symmetric and skew-symmetric products as symmetric or skew multilinear forms. Define, for example,

$$\mathbf{e}^{*\alpha} \wedge \mathbf{e}^{*\beta}\,(\mathbf{e}_\mu, \mathbf{e}_\nu) = \delta^\alpha_\mu \delta^\beta_\nu - \delta^\alpha_\nu \delta^\beta_\mu, \tag{1.62}$$

and

$$\mathbf{e}^{*\alpha} \wedge \mathbf{e}^{*\beta}\,(\mathbf{e}_\mu \wedge \mathbf{e}_\nu) = \delta^\alpha_\mu \delta^\beta_\nu - \delta^\alpha_\nu \delta^\beta_\mu. \tag{1.63}$$

We need two terms on the right-hand-side of these examples because the skew-symmetry of $\mathbf{e}^{*\alpha} \wedge \mathbf{e}^{*\beta}(\ ,\ )$ in its slots does not allow us the luxury of demanding that the $\mathbf{e}_\mu$ be inserted in the exact order of the $\mathbf{e}^{*\alpha}$ to get a non-zero answer. Because the p-th order analogue of (1.62) form has $p!$ terms on its right-hand side, some authors like to divide the right-hand-side by $p!$ in this definition. We prefer the one above, though. With our definition, and with $\mathbf{A} = \frac{1}{2} A_{\mu\nu} \mathbf{e}^{*\mu} \wedge \mathbf{e}^{*\nu}$ and $\mathbf{B} = \frac{1}{2} B^{\alpha\beta} \mathbf{e}_\alpha \wedge \mathbf{e}_\beta$, we have

$$\mathbf{A}(\mathbf{B}) = \frac{1}{2} A_{\mu\nu} B^{\mu\nu} = \sum_{\mu<\nu} A_{\mu\nu} B^{\mu\nu}, \tag{1.64}$$

so the sum is only over independent terms.

The wedge ($\wedge$) product notation is standard in mathematics wherever skew-symmetry is implied.[3] The "sym" and $\odot$ are not. Different authors use different notations for spaces of symmetric tensors. This reflects the fact that skew-symmetric tensors are extremely useful and appear in many different parts of mathematics, while symmetric ones have fewer special properties (although they are common in physics). Compare the relative usefulness of determinants and permanents.

Exercise 1.4: Show that in d dimensions:

i) the dimension of the space of skew-symmetric covariant tensors with p indices is $d!/p!(d-p)!$;
ii) the dimension of the space of symmetric covariant tensors with p indices is $(d+p-1)!/p!(d-1)!$.

[3] Skew products, along with the first formulation of the idea of an abstract vector space, were introduced in Hermann Grassmann's *Ausdehnungslehre* (1844). Grassmann's mathematics was not appreciated in his lifetime. In his disappointment he turned to other fields, making significant contributions to the theory of colour mixtures (Grassmann's law), and to the philology of Indo-European languages (another Grassmann's law).

Bosons and fermions

Spaces of symmetric and skew-symmetric tensors appear whenever we deal with the quantum mechanics of many indistinguishable particles possessing Bose or Fermi statistics. If we have a Hilbert space $\mathcal{H}$ of single-particle states with basis $\mathbf{e}_i$ then the N-boson space is $\mathrm{Sym}^N\mathcal{H}$ which consists of states

$$\mathbf{\Phi} = \Phi^{i_1 i_2 \ldots i_N} \mathbf{e}_{i_1} \odot \mathbf{e}_{i_2} \odot \cdots \odot \mathbf{e}_{i_{i_N}}, \tag{1.65}$$

and the N-fermion space is $\bigwedge^N\mathcal{H}$, which contains states

$$\mathbf{\Psi} = \frac{1}{N!} \Psi^{i_1 i_2 \ldots i_N} \, \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N}. \tag{1.66}$$

The symmetry of the Bose wavefunction

$$\Phi^{i_1 \ldots i_\alpha \ldots i_\beta \ldots i_N} = \Phi^{i_2 \ldots i_\beta \ldots i_\alpha \ldots i_N}, \tag{1.67}$$

and the skew-symmetry of the Fermion wavefunction

$$\Psi^{i_1 \ldots i_\alpha \ldots i_\beta \ldots i_N} = -\Psi^{i_2 \ldots i_\beta \ldots i_\alpha \ldots i_N}, \tag{1.68}$$

under the interchange of the particle labels α, β is then natural.
Slater Determinants and the Plücker Relations: Some N-fermion states can be decomposed into a product of single-particle states

$$\begin{aligned} \mathbf{\Psi} &= \boldsymbol{\psi}_1 \wedge \boldsymbol{\psi}_2 \wedge \cdots \wedge \boldsymbol{\psi}_N \\ &= \psi_1^{i_1} \psi_2^{i_2} \cdots \psi_N^{i_N} \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N}. \end{aligned} \tag{1.69}$$

Comparing the coefficients of $\mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N}$ in (1.66) and (1.69) shows that the many-body wavefunction can then be written as

$$\Psi^{i_1 i_2 \ldots i_N} = \begin{vmatrix} \psi_1^{i_1} & \psi_1^{i_2} & \cdots & \psi_1^{i_N} \\ \psi_2^{i_1} & \psi_2^{i_2} & \cdots & \psi_2^{i_N} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_N^{i_1} & \psi_N^{i_2} & \cdots & \psi_N^{i_N} \end{vmatrix}. \tag{1.70}$$

The wavefunction is therefore given by a single *Slater determinant*. Such wavefunctions correspond to a very special class of states. The general many-fermion state is not decomposable, and its wavefunction can only be expressed as a sum of many Slater determinants. The Hartree-Fock method

of quantum chemistry is a variational approximation that takes such a single Slater determinant as its trial wavefunction and varies only the one-particle wavefunctions $\langle i|\psi_a\rangle \equiv \psi_a^i$. It is a remarkably successful approximation, given the very restricted class of wavefunctions it explores.

As with the Segre condition for two distinguishable quantum systems to be unentangled, there is a set of necessary and sufficient conditions on the $\Psi^{i_1 i_2 \ldots i_N}$ for the state $\mathbf{\Psi}$ to be decomposable into single-particle states. The conditions are that

$$\Psi^{i_1 i_2 \ldots i_{N-1}[j_1} \Psi^{j_1 j_2 \ldots j_{N+1}]} = 0 \tag{1.71}$$

for any choice of indices $i_1, \ldots i_{N-1}$ and $j_1, \ldots, j_{N+1}$. The square brackets $[\ldots]$ indicate that the expression is to be antisymmetrized over the indices enclosed in the brackets. For example, a three-particle state is decomposable if and only if

$$\Psi^{i_1 i_2 j_1} \Psi^{j_2 j_3 j_4} - \Psi^{i_1 i_2 j_2} \Psi^{j_1 j_3 j_4} + \Psi^{i_1 i_2 j_3} \Psi^{j_1 j_2 j_4} - \Psi^{i_1 i_2 j_4} \Psi^{j_1 j_2 j_3} = 0. \tag{1.72}$$

These conditions are called the *Plücker relations* after Julius Plücker who discovered them long before before the advent of quantum mechanics.[4] It is easy to show that Plücker's relations are necessary conditions for decomposability. It takes more sophistication to show that they are sufficient. We will therefore defer this task to the exercises as the end of the chapter. As far as we are aware, the Plücker relations are not exploited by quantum chemists, but, in disguise as the *Hirota bilinear equations*, they constitute the geometric condition underpinning the many-soliton solutions of the Korteweg-de-Vries and other soliton equations.

1.2.5 Kronecker and Levi-Civita tensors

Suppose the tensor δ^μ_ν is defined, with respect to some basis, to be unity if $\mu = \nu$ and zero otherwise. In a new basis it will transform to

$$\delta'^\mu_{\ \nu} = a^\mu_\rho (a^{-1})^\sigma_\nu \delta^\rho_\sigma = a^\mu_\rho (a^{-1})^\nu_\rho = \delta^\mu_\nu. \tag{1.73}$$

In other words the Kronecker delta symbol of type $(1,1)$ has the same numerical components in all co-ordinate systems. This is not true of the Kroneker delta symbol of type $(0,2)$, *i.e.* of $\delta_{\mu\nu}$.

[4] As well as his extensive work in algebraic geometry, Plücker (1801-68) made important discoveries in experimental physics. He was, for example, the first person to observe the deflection of cathode rays — beams of electrons — by a magnetic field, and the first to point out that each element had its characteristic emission spectrum.

Now consider an n-dimensional space with a tensor $\eta_{\mu_1\mu_2\ldots\mu_n}$ whose components, in some basis, coincides with the Levi-Civita symbol $\epsilon_{\mu_1\mu_2\ldots\mu_n}$. We find that in a new frame the components are

$$\begin{aligned}\eta'_{\mu_1\mu_2\ldots\mu_n} &= (a^{-1})^{\nu_1}_{\mu_1}(a^{-1})^{\nu_2}_{\mu_2}\cdots(a^{-1})^{\nu_n}_{\mu_n}\epsilon_{\nu_1\nu_2\ldots\nu_n} \\ &= \epsilon_{\mu_1\mu_2\ldots\mu_n}\,(a^{-1})^{\nu_1}_{1}(a^{-1})^{\nu_2}_{2}\cdots(a^{-1})^{\nu_n}_{n}\epsilon_{\nu_1\nu_2\ldots\nu_n} \\ &= \epsilon_{\mu_1\mu_2\ldots\mu_n}\det\mathbf{A}^{-1} \\ &= \eta_{\mu_1\mu_2\ldots\mu_n}\det\mathbf{A}^{-1}. \end{aligned} \tag{1.74}$$

Thus, unlike the δ^μ_ν, the Levi-Civita symbol is not quite a tensor.

Consider also the quantity

$$\sqrt{g} \stackrel{\text{def}}{=} \sqrt{\det[g_{\mu\nu}]}. \tag{1.75}$$

Here we assume that the metric is positive-definite, so that the square root is real, and that we have taken the positive square root. Since

$$\det[g'_{\mu\nu}] = \det[(a^{-1})^\rho_\mu(a^{-1})^\sigma_\nu g_{\rho\sigma}] = (\det\mathbf{A})^{-2}\det[g_{\mu\nu}], \tag{1.76}$$

we see that

$$\sqrt{g'} = |\det\mathbf{A}|^{-1}\sqrt{g} \tag{1.77}$$

Thus $\sqrt{g}$ is also not quite an invariant. This is only to be expected, because $\mathbf{g}(\ ,\ )$ is a quadratic form and we know that there is no basis-independent meaning to the determinant of such an object.

Now define

$$\varepsilon_{\mu_1\mu_2\ldots\mu_n} = \sqrt{g}\,\epsilon_{\mu_1\mu_2\ldots\mu_n}, \tag{1.78}$$

and assume that $\varepsilon_{\mu_1\mu_2\ldots\mu_n}$ has the type $(0,n)$ tensor character implied by its indices. When we look at how this transforms, and restrict ourselves to *orientation preserving* changes of of bases, *i.e.* ones for which $\det\mathbf{A}$ is positive, we see that factors of $\det\mathbf{A}$ conspire to give

$$\varepsilon'_{\mu_1\mu_2\ldots\mu_n} = \sqrt{g'}\,\epsilon_{\mu_1\mu_2\ldots\mu_n}. \tag{1.79}$$

A similar exercise indictes that if we define $\epsilon^{\mu_1\mu_2\ldots i_n}$ to be numerically equal to $\epsilon_{i_1i_2\ldots\mu_n}$ then

$$\varepsilon^{\mu_1\mu_2\ldots\mu_n} = \frac{1}{\sqrt{g}}\,\epsilon^{\mu_1\mu_2\ldots\mu_n} \tag{1.80}$$

also transforms as a tensor — in this case a type $(n, 0)$ contravariant one — provided that the factor of $1/\sqrt{g}$ is always calculated with respect to the current basis.

If the dimension n is even and we are given a skew-symetric tensor $F_{\mu\nu}$, we can therefore construct an invariant

$$\varepsilon^{\mu_1\mu_2\ldots\mu_n} F_{\mu_1\mu_2}\cdots F_{\mu_{n-1}\mu_n} = \frac{1}{\sqrt{g}}\epsilon^{\mu_1\mu_2\ldots\mu_n} F_{\mu_1\mu_2}\cdots F_{\mu_{n-1}\mu_n}. \tag{1.81}$$

Similarly, given an skew-symmetric covariant tensor $F_{\mu_1\ldots\mu_m}$ with m $(\leq n)$ indices we can form its *dual*, denoted by F^*, a $(n-m)$-contravariant tensor with components

$$(F^*)^{\mu_{m-1}\ldots\mu_n} = \frac{1}{m!}\varepsilon^{\mu_1\mu_2\ldots\mu_n} F_{\mu_1\ldots\mu_m} = \frac{1}{\sqrt{g}}\frac{1}{m!}\epsilon^{\mu_1\mu_2\ldots\mu_n} F_{\mu_1\ldots\mu_m}. \tag{1.82}$$

We meet this "dual" tensor again, when we study differential forms.

1.3 Cartesian Tensors

If we restrict ourselves to Cartesian co-ordinate systems having orthonormal basis vectors, so that $g_{ij} = \delta_{ij}$, then there are considerable simplifications. In particular, we do not have to make a distinction between co- and contravariant indices. We shall usually write their indices as roman-alphabet suffixes.

A change of basis from one orthogonal n-dimensional basis $\mathbf{e}_i$ to another $\mathbf{e}'_i$ will set

$$\mathbf{e}'_i = O_{ij}\mathbf{e}_j, \tag{1.83}$$

where the numbers O_{ij} are the entries in an *orthogonal* matrix $\mathbf{O}$, *i.e.* a real matrix obeying $\mathbf{O}^T\mathbf{O} = \mathbf{O}\mathbf{O}^T = \mathbf{I}$, where T denotes the transpose. The set of n-by-n orthogonal matrices constitutes the *orthogonal group* $\mathrm{O}(n)$.

1.3.1 Isotropic tensors

The Kronecker δ_{ij} with both indices downstairs is unchanged by $\mathrm{O}(n)$ transformations,

$$\delta'_{ij} = O_{ik}O_{jl}\delta_{kl} = O_{ik}O_{jk} = O_{ik}O^T_{kj} = \delta_{ij}, \tag{1.84}$$

and has the same components in any Cartesian frame. We say that its components are *numerically invariant*. A similar property holds for tensors made up of products of δ_{ij}, such as

$$T_{ijklmn} = \delta_{ij}\delta_{kl}\delta_{mn}. \tag{1.85}$$

It is possible to show[5] that any tensor whose components are numerically invariant under all orthogonal transformations is a sum of products of this form. The most general O(n) invariant tensor of rank four is, for example.

$$\alpha\delta_{ij}\delta_{kl} + \beta\delta_{ik}\delta_{lj} + \gamma\delta_{il}\delta_{jk}. \tag{1.86}$$

The determinant of an orthogonal transformation must be ± 1. If we only allow orientation-preserving changes of basis then we restrict ourselves to orthogonal transformations O_{ij} with $\det \mathbf{O} = 1$. These are the *proper* orthogonal transformations. In n dimensions they constitute the group SO(n). Under SO(n) transformations, both δ_{ij} and $\epsilon_{i_1 i_2 \ldots i_n}$ are numerically invariant and the most general SO(n) invariant tensors consist of sums of products of δ_{ij}'s and $\epsilon_{i_1 i_2 \ldots i_n}$'s. The most general SO(4)-invariant rank-four tensor is, for example,

$$\alpha\delta_{ij}\delta_{kl} + \beta\delta_{ik}\delta_{lj} + \gamma\delta_{il}\delta_{jk} + \lambda\epsilon_{ijkl}. \tag{1.87}$$

Tensors that are numerically invariant under SO(n) are known as *isotropic tensors*.

As there is no longer any distinction between co- and contravariant indices, we can now contract any pair of indices. In three dimensions, for example,

$$B_{ijkl} = \epsilon_{nij}\epsilon_{nkl} \tag{1.88}$$

is a rank-four isotropic tensor. Now $\epsilon_{i_1 \ldots i_n}$ is *not* invariant when we transform via an orthogonal transformation with $\det \mathbf{O} = -1$, but the product of two ϵ's *is* invariant under such transformations. The tensor B_{ijkl} is therefore numerically invariant under the larger group O(3) and must be expressible as

$$B_{ijkl} = \alpha\delta_{ij}\delta_{kl} + \beta\delta_{ik}\delta_{lj} + \gamma\delta_{il}\delta_{jk} \tag{1.89}$$

for some coefficients α, β and γ. The following exercise explores some consequences of this and related facts.

[5] The proof is surprisingly complicated. See, for example, M. Spivak, *A Comprehensive Introduction to Differential Geometry* (second edition) Vol. V, pp. 466-481.

Exercise 1.5: We defined the n-dimensional Levi-Civita symbol by requiring that $\epsilon_{i_1 i_2 \ldots i_n}$ be antisymmetric in all pairs of indices, and $\epsilon_{12\ldots n} = 1$.

a) Show that $\epsilon_{123} = \epsilon_{231} = \epsilon_{312}$, but that $\epsilon_{1234} = -\epsilon_{2341} = \epsilon_{3412} = -\epsilon_{4123}$.
b) Show that
$$\epsilon_{ijk}\epsilon_{i'j'k'} = \delta_{ii'}\delta_{jj'}\delta_{kk'} + \text{five other terms},$$
where you should write out all six terms explicitly.
c) Show that $\epsilon_{ijk}\epsilon_{ij'k'} = \delta_{jj'}\delta_{kk'} - \delta_{jk'}\delta_{kj'}$.
d) For dimension $n = 4$, write out $\epsilon_{ijkl}\epsilon_{ij'k'l'}$ as a sum of products of δ's similar to the one in part (c).

Exercise 1.6: *Vector Products.* The vector product of two three-vectors may be written in Cartesian components as $(\mathbf{a} \times \mathbf{b})_i = \epsilon_{ijk} a_j b_k$. Use this and your results about ϵ_{ijk} from the previous exercise to show that

i) $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$,
ii) $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$,
iii) $(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c})$.
iv) If we take $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{d}$, with $\mathbf{d} \equiv \mathbf{b}$, to be unit vectors, show that the identities (i) and (iii) become the sine and cosine rule, respectively, of spherical trigonometry. (Hint: for the spherical sine rule, begin by showing that $\mathbf{a} \cdot [(\mathbf{a} \times \mathbf{b}) \times (\mathbf{a} \times \mathbf{c})] = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.)

1.3.2 Stress and strain

As an illustration of the utility of Cartesian tensors, we consider their application to elasticity.

Suppose that an elastic body is slightly deformed so that the particle that was originally at the point with Cartesian co-ordinates x_i is moved to $x_i + \eta_i$. We define the (infinitesimal) *strain tensor* e_{ij} by

$$e_{ij} = \frac{1}{2}\left(\frac{\partial \eta_j}{\partial x_i} + \frac{\partial \eta_i}{\partial x_j}\right). \tag{1.90}$$

It is automatically symmetric: $e_{ij} = e_{ji}$. We will leave for later (exercise 2.3) a discussion of why this is the natural definition of strain, and also the modifications necessary were we to employ a non-Cartesian co-ordinate system.

To define the *stress tensor* σ_{ij} we consider the portion Ω of the body in figure 1.1, and an element of area $dS = \mathbf{n}\, d|S|$ on its boundary. Here, $\mathbf{n}$ is

the unit normal vector pointing out of Ω. The force $\mathbf{F}$ exerted on this surface element by the parts of the body exterior to Ω has components

$$F_i = \sigma_{ij} n_j \, d|S|. \tag{1.91}$$

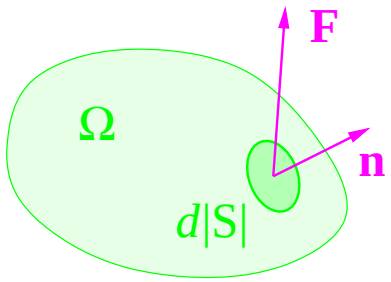

Figure 1.1: *Stress forces.*

That $\mathbf{F}$ is a linear function of $\mathbf{n}\, d|S|$ can be seen by considering the forces on an small tetrahedron, three of whose sides coincide with the co-ordinate planes, the fourth side having $\mathbf{n}$ as its normal. In the limit that the lengths of the sides go to zero as ϵ, the mass of the body scales to zero as ϵ^3, but the forces are proprtional to the areas of the sides and go to zero only as ϵ^2. Only if the linear relation holds true can the acceleration of the tetrahedron remain finite. A similar argument applied to torques and the moment of intertia of a small cube shows that $\sigma_{ij} = \sigma_{ji}$.

A generalization of Hooke's law,

$$\sigma_{ij} = c_{ijkl} e_{kl}, \tag{1.92}$$

relates the stress to the strain via the tensor of *elastic constants* c_{ijkl}. This rank-four tensor has the symmetry properties

$$c_{ijkl} = c_{klij} = c_{jikl} = c_{ijlk}. \tag{1.93}$$

In other words, the tensor is symmetric under the interchange of the first and second pairs of indices, and also under the interchange of the individual indices in either pair.

For an isotropic material — a material whose properties are invariant under the rotation group SO(3) — the tensor of elastic constants must be an

isotropic tensor. The most general such tensor with the required symmetries is

$$c_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}). \tag{1.94}$$

As isotropic material is therefore characterized by only two independent parameters, λ and μ. These are called the *Lamé* constants after the mathematical engineer Gabriel Lamé. In terms of them the generalized Hooke's law becomes

$$\sigma_{ij} = \lambda\delta_{ij}e_{kk} + 2\mu e_{ij}. \tag{1.95}$$

By considering particular deformations, we can express the more directly measurable *bulk modulus*, *shear modulus*, *Young's modulus* and *Poisson's ratio* in terms of λ and μ.

The bulk modulus κ is defined by

$$\frac{dV}{V} = -\kappa dP, \tag{1.96}$$

where an infinitesimal isotropic external pressure dP causes a change $V \to V + dV$ in the volume of the material. This applied pressure corresponds to a surface stress of $\sigma_{ij} = -\delta_{ij}\, dP$. An isotropic expansion displaces points in the material so that

$$\eta_i = \frac{1}{3}\frac{dV}{V}x_i. \tag{1.97}$$

The strains are therefore given by

$$e_{ij} = \frac{1}{3}\delta_{ij}\frac{dV}{V}. \tag{1.98}$$

Inserting this strain into the stress-strain relation gives

$$\sigma_{ij} = \delta_{ij}(\lambda + \frac{2}{3}\mu)\frac{dV}{V} = -\delta_{ij}dP. \tag{1.99}$$

Thus

$$\kappa = \lambda + \frac{2}{3}\mu. \tag{1.100}$$

To define the shear modulus, we assume a deformation $\eta_1 = \theta x_2$, so $e_{12} = e_{21} = \theta/2$, with all other e_{ij} vanishing.

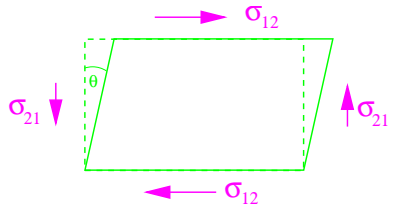

Figure 1.2: *Shear strain. The arrows show the direction of the applied stresses. The σ_{21} on the vertical faces are necessary to stop the body rotating.*

The applied shear stress is $\sigma_{12} = \sigma_{21}$. The shear modulus, is defined to be σ_{12}/θ. Inserting the strain components into the stress-strain relation gives

$$\sigma_{12} = \mu\theta, \tag{1.101}$$

and so the shear modulus is equal to the Lamé constant μ. We can therefore write the generalized Hooke's law as

$$\sigma_{ij} = 2\mu(e_{ij} - \tfrac{1}{3}\delta_{ij}e_{kk}) + \kappa e_{kk}\delta_{ij}, \tag{1.102}$$

which reveals that the shear modulus is associated with the traceless part of the strain tensor, and the bulk modulus with the trace.

Young's modulus Y is measured by stretching a wire of initial length L and square cross section of side W under a tension $T = \sigma_{33}W^2$.

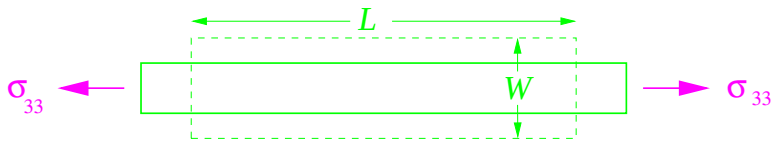

Figure 1.3: *Forces on a stretched wire.*

We define Y so that

$$\sigma_{33} = Y\frac{dL}{L}. \tag{1.103}$$

At the same time as the wire stretches, its width changes $W \to W + dW$. Poisson's ratio σ is defined by

$$\frac{dW}{W} = -\sigma\frac{dL}{L}, \tag{1.104}$$

so that σ is positive if the wire gets thinner as it gets longer. The displacements are

$$\begin{aligned} \eta_3 &= z\left(\frac{dL}{L}\right), \\ \eta_1 &= x\left(\frac{dW}{W}\right) = -\sigma x\left(\frac{dL}{L}\right), \\ \eta_2 &= y\left(\frac{dW}{W}\right) = -\sigma y\left(\frac{dL}{L}\right), \end{aligned} \tag{1.105}$$

so the strain components are

$$e_{33} = \frac{dL}{L}, \quad e_{11} = e_{22} = \frac{dW}{W} = -\sigma e_{33}. \tag{1.106}$$

We therefore have

$$\sigma_{33} = (\lambda(1-2\sigma) + 2\mu)\left(\frac{dL}{L}\right), \tag{1.107}$$

leading to

$$Y = \lambda(1-2\sigma) + 2\mu. \tag{1.108}$$

Now, the side of the wire is a free surface with no forces acting on it, so

$$0 = \sigma_{22} = \sigma_{11} = (\lambda(1-2\sigma) - 2\sigma\mu)\left(\frac{dL}{L}\right). \tag{1.109}$$

This tells us that[6]

$$\sigma = \frac{1}{2}\frac{\lambda}{\lambda+\mu}, \tag{1.110}$$

and

$$Y = \mu\left(\frac{3\lambda+2\mu}{\lambda+\mu}\right). \tag{1.111}$$

Other relations, following from those above, are

$$\begin{aligned} Y &= 3\kappa(1-2\sigma), \\ &= 2\mu(1+\sigma). \end{aligned} \tag{1.112}$$

[6]Poisson and Cauchy believed that $\lambda = \mu$, and hence that $\sigma = 1/4$.

Exercise 1.7: Show that the symmetries

$$c_{ijkl} = c_{klij} = c_{jikl} = c_{ijlk}$$

imply that a general homogeneous material has 21 independent elastic constants. (This result was originally obtained by George Green, of Green function fame.)

Exercise 1.8: A steel beam is forged so that its cross section has the shape of a region $\Gamma \in \mathbb{R}^2$. When undeformed, it lies along the z axis. The centroid O of each cross section is defined so that

$$\int_\Gamma x \, dxdy = \int_\Gamma y \, dxdy = 0,$$

when the co-ordinates x, y are taken with the centroid O as the origin. The beam is slightly bent away from the z axis so that the line of centroids remains in the y, z plane. At a particular cross section with centroid O, the line of centroids has radius of curvature R.

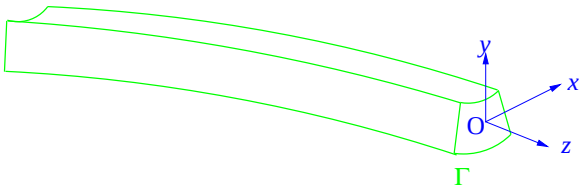

Figure 1.4: *Bent beam.*

Assume that the deformation in the vicinity of O is such that

$$\begin{aligned} \eta_x &= -\frac{\sigma}{R} xy, \\ \eta_y &= \frac{1}{2R} \left\{ \sigma(x^2 - y^2) - z^2 \right\}, \\ \eta_z &= \frac{1}{R} yz. \end{aligned}$$

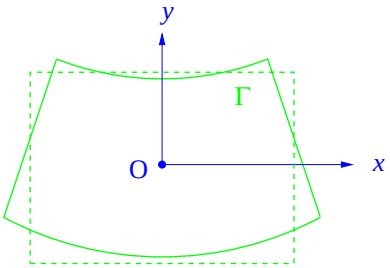

Figure 1.5: *The original (dashed) and anticlastically deformed (full) cross-section.*

For positive Poisson ratio, the cross section deforms *anticlastically* — the sides bend *up* as the beam bends *down*.

Compute the strain tensor resulting from the given deformation, and show that its only non-zero components are

$$e_{xx} = -\frac{\sigma}{R}y, \qquad e_{yy} = -\frac{\sigma}{R}y, \qquad e_{zz} = \frac{1}{R}y.$$

Next, show that

$$\sigma_{zz} = \left(\frac{Y}{R}\right) y,$$

and that all other components of the stress tensor vanish. Deduce from this vanishing that the assumed deformation satisfies the free-surface boundary condition, and so is indeed the way the beam responds when it is bent by forces applied at its ends.

The work done in bending the beam

$$\int_{\text{beam}} \frac{1}{2} e_{ij} c_{ijkl} e_{kl} \, d^3x$$

is stored as elastic energy. Show that for our bent rod this energy is equal to

$$\int \frac{YI}{2} \left(\frac{1}{R^2}\right) ds \approx \int \frac{YI}{2} (y'')^2 dz,$$

where s is the arc-length taken along the line of centroids of the beam,

$$I = \int_\Gamma y^2 \, dxdy$$

is the moment of inertia of the region Γ about the x axis, and y'' denotes the second derivative of the deflection of the beam with respect to z (which

approximates the arc-length). This last formula for the strain energy has been used in a number of our calculus-of-variations problems.

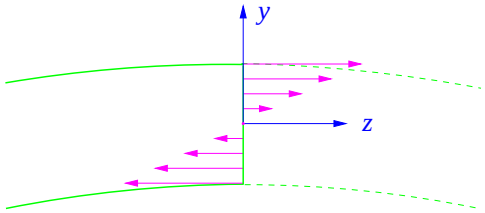

Figure 1.6: *The distribution of forces σ_{zz} exerted on the left-hand part of the bent rod by the material to its right.*

1.3.3 Maxwell stress tensor

Consider a small cubical element of an elastic body. If the stress tensor were position independent, the external forces on each pair of opposing faces of the cube would be equal in magnitude but pointing in opposite directions. There would therefore be no net external force on the cube. When σ_{ij} is *not* constant then we claim that the total force acting on an infinitesimal element of volume dV is

$$F_i = \partial_j \sigma_{ij}\, dV. \tag{1.113}$$

To see that this assertion is correct, consider a finite region Ω with boundary $\partial\Omega$, and use the divergence theorem to write the total force on Ω as

$$F_i^{\rm tot} = \int_{\partial\Omega} \sigma_{ij} n_j d|S| = \int_\Omega \partial_j \sigma_{ij} dV. \tag{1.114}$$

Whenever the force-per-unit-volume f_i acting on a body can be written in the form $f_i = \partial_j \sigma_{ij}$, we refer to σ_{ij} as a "stress tensor," by analogy with stress in an elastic solid. As an example, let $\mathbf{E}$ and $\mathbf{B}$ be electric and magnetic fields. For simplicity, initially assume them to be static. The force per unit volume exerted by these fields on a distribution of charge ρ and current $\mathbf{j}$ is

$$\mathbf{f} = \rho\mathbf{E} + \mathbf{j} \times \mathbf{B}. \tag{1.115}$$

From Gauss' law $\rho = \operatorname{div} \mathbf{D}$, and with $\mathbf{D} = \epsilon_0 \mathbf{E}$, we find that the force per unit volume due the electric field has components

$$\rho E_i = (\partial_j D_j) E_i \;\; = \;\; \epsilon_0 \Big(\partial_j (E_i E_j) - E_j\, \partial_j E_i \Big)$$

$$\begin{aligned} &= \epsilon_0\Big(\partial_j(E_iE_j) - E_j\,\partial_iE_j\Big) \\ &= \epsilon_0\partial_j\left(E_iE_j - \frac{1}{2}\delta_{ij}|E|^2\right). \end{aligned} \tag{1.116}$$

Here, in passing from the first line to the second, we have used the fact that $\operatorname{curl}\mathbf{E}$ is zero for static fields, and so $\partial_jE_i = \partial_iE_j$. Similarly, using $\mathbf{j} = \operatorname{curl}\mathbf{H}$, together with $\mathbf{B} = \mu_0\mathbf{H}$ and $\operatorname{div}\mathbf{B} = 0$, we find that the force per unit volume due the magnetic field has components

$$(\mathbf{j}\times\mathbf{B})_i = \mu_0\partial_j\left(H_iH_j - \frac{1}{2}\delta_{ij}|H|^2\right). \tag{1.117}$$

The quantity

$$\sigma_{ij} = \epsilon_0\left(E_iE_j - \frac{1}{2}\delta_{ij}|E|^2\right) + \mu_0\left(H_iH_j - \frac{1}{2}\delta_{ij}|H|^2\right) \tag{1.118}$$

is called the *Maxwell stress tensor*. Its utility lies in in the fact that the total electromagnetic force on an isolated body is the integral of the Maxwell stress over its surface. We do not need to know the fields within the body.

Michael Faraday was the first to intuit a picture of electromagnetic stresses and attributed both a longitudinal tension and a mutual lateral repulsion to the field lines. Maxwell's tensor expresses this idea mathematically.

Exercise 1.9: Allow the fields in the preceding calculation to be time dependent. Show that Maxwell's equations

$$\begin{aligned} \operatorname{curl}\mathbf{E} &= -\frac{\partial\mathbf{B}}{\partial t}, \quad \operatorname{div}\mathbf{B} = 0, \\ \operatorname{curl}\mathbf{H} &= \mathbf{j} + \frac{\partial\mathbf{D}}{\partial t}, \quad \operatorname{div}\mathbf{D} = \rho, \end{aligned}$$

with $\mathbf{B} = \mu_0\mathbf{H}$, $\mathbf{D} = \epsilon_0\mathbf{E}$, and $c = 1/\sqrt{\mu_0\epsilon_0}$, lead to

$$(\rho\mathbf{E} + \mathbf{j}\times\mathbf{B})_i + \frac{\partial}{\partial t}\left\{\frac{1}{c^2}(\mathbf{E}\times\mathbf{H})_i\right\} = \partial_j\sigma_{ij}.$$

The left-hand side is the time rate of change of the mechanical (first term) and electromagnetic (second term) momentum density. Observe that we can equivalently write

$$\frac{\partial}{\partial t}\left\{\frac{1}{c^2}(\mathbf{E}\times\mathbf{H})_i\right\} + \partial_j(-\sigma_{ij}) = -(\rho\mathbf{E} + \mathbf{j}\times\mathbf{B})_i,$$

and think of this a local field-momentum conservation law. In this interpretation $-\sigma_{ij}$ is thought of as the *momentum flux* tensor, its entries being the flux in direction j of the component of field momentum in direction i. The term on the right-hand side is the rate at which momentum is being supplied to the electro-magnetic field by the charges and currents.

1.4 Further Exercises and Problems

Exercise 1.10: *Quotient theorem.* Suppose that you have come up with some recipe for generating an array of numbers T^{ijk} in any co-ordinate frame, and want to know whether these numbers are the components of a triply contravariant tensor. Suppose further that you know that, given the components a_{ij} of an arbitrary doubly covariant tensor, the numbers

$$T^{ijk}a_{jk} = v^i$$

transform as the components of a contravariant vector. Show that T^{ijk} does indeed transform as a triply contravariant tensor. (The natural generalization of this result to arbitrary tensor types is known as the *quotient theorem*.)

Exercise 1.11: Let $T^i{}_j$ be the 3-by-3 array of components of a tensor. Show that the quantities

$$a = T^i{}_i, \quad b = T^i{}_j T^j{}_i, \quad c = T^i{}_j T^j{}_k T^k{}_i$$

are invariant. Further show that the eigenvalues of the linear map represented by the matrix $T^i{}_j$ can be found by solving the cubic equation

$$\lambda^3 - a\lambda^2 + \frac{1}{2}(a^2 - b)\lambda - \frac{1}{6}(a^3 - 3ab + 2c) = 0.$$

Exercise 1.12: Let the covariant tensor R_{ijkl} possess the following symmetries:

i) $R_{ijkl} = -R_{jikl}$,
ii) $R_{ijkl} = -R_{ijlk}$,
iii) $R_{ijkl} + R_{iklj} + R_{iljk} = 0$.

Use the properties i),ii), iii) to show that:

a) $R_{ijkl} = R_{klij}$.
b) If $R_{ijkl}x^i y^j x^k y^l = 0$ for all vectors x^i, y^i, then $R_{ijkl} = 0$.

c) If B_{ij} is a symmetric covariant tensor and set we $A_{ijkl} = B_{ik}B_{jl} - B_{il}B_{jk}$, then A_{ijkl} has the same symmetries as R_{ijkl}.

Exercise 1.13: Write out Euler's equation for fluid motion

$$\dot{\mathbf{v}} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla h$$

in Cartesian tensor notation. Transform it into

$$\dot{\mathbf{v}} - \mathbf{v} \times \boldsymbol{\omega} = -\nabla \left(\frac{1}{2}\mathbf{v}^2 + h \right),$$

where $\boldsymbol{\omega} = \nabla \times \mathbf{v}$ is the vorticity. Deduce Bernoulli's theorem, that for steady ($\dot{\mathbf{v}} = 0$) flow the quantity $\frac{1}{2}\mathbf{v}^2 + h$ is constant along streamlines.

Exercise 1.14: *Symmetric integration.* Show that the n-dimensional integral

$$I_{\alpha\beta\gamma\delta} = \int \frac{d^n k}{(2\pi)^n} (k_\alpha k_\beta k_\gamma k_\delta)\, f(k^2),$$

is equal to

$$A(\delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma})$$

where

$$A = \frac{1}{n(n+2)} \int \frac{d^n k}{(2\pi)^n} (k^2)^2 f(k^2).$$

Similarly evaluate

$$I_{\alpha\beta\gamma\delta\epsilon} = \int \frac{d^n k}{(2\pi)^n} (k_\alpha k_\beta k_\gamma k_\delta k_\epsilon)\, f(k^2).$$

Exercise 1.15: Write down the most general three-dimensional isotropic tensors of rank two and three.

In piezoelectric materials, the application of an electric field E_i induces a mechanical strain that is described by a rank-two symmetric tensor

$$e_{ij} = d_{ijk} E_k,$$

where d_{ijk} is a third-rank tensor that depends only on the material. Show that e_{ij} can only be non-zero in an anisotropic material.

Exercise 1.16: In three dimensions, a rank-five isotropic tensor T_{ijklm} is a linear combination of expressions of the form $\epsilon_{i_1 i_2 i_3}\delta_{i_4 i_5}$ for some assignment of the indices i, j, k, l, m to the $i_1, \ldots, i_5$. Show that, on taking into account the symmetries of the Kronecker and Levi-Civita symbols, we can construct *ten* distinct products $\epsilon_{i_1 i_2 i_3}\delta_{i_4 i_5}$. Only *six* of these are linearly independent, however. Show, for example, that

$$\epsilon_{ijk}\delta_{lm} - \epsilon_{jkl}\delta_{im} + \epsilon_{kli}\delta_{jm} - \epsilon_{lij}\delta_{km} = 0,$$

and find the three other independent relations of this sort.[7]

(*Hint*: Begin by showing that, in three dimensions,

$$\delta^{i_1 i_2 i_3 i_4}_{i_5 i_6 i_7 i_8} \stackrel{\text{def}}{=} \begin{vmatrix} \delta_{i_1 i_5} & \delta_{i_1 i_6} & \delta_{i_1 i_7} & \delta_{i_1 i_8} \\ \delta_{i_2 i_5} & \delta_{i_2 i_6} & \delta_{i_2 i_7} & \delta_{i_2 i_8} \\ \delta_{i_3 i_5} & \delta_{i_3 i_6} & \delta_{i_3 i_7} & \delta_{i_3 i_8} \\ \delta_{i_4 i_5} & \delta_{i_4 i_6} & \delta_{i_4 i_7} & \delta_{i_4 i_8} \end{vmatrix} = 0,$$

and contract with $\epsilon_{i_6 i_7 i_8}$.)

Problem 1.17: *The Plücker Relations.* This problem provides a challenging test of your understanding of linear algebra. It leads you through the task of deriving the necessary and sufficient conditions for

$$\mathbf{A} = A^{i_1 \ldots i_k}\, \mathbf{e}_{i_1} \wedge \ldots \wedge \mathbf{e}_{i_k} \in \bigwedge{}^k V$$

to be decomposable as

$$\mathbf{A} = \mathbf{f}_1 \wedge \mathbf{f}_2 \wedge \ldots \wedge \mathbf{f}_k.$$

The trick is to introduce two subspaces of V,

i) W, the smallest subspace of V such that $\mathbf{A} \in \bigwedge^k W$,
ii) $W' = \{\mathbf{v} \in V : \mathbf{v} \wedge \mathbf{A} = 0\}$,

and explore their relationship.

a) Show that if $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$ constitute a basis for W', then

$$\mathbf{A} = \mathbf{w}_1 \wedge \mathbf{w}_2 \wedge \cdots \wedge \mathbf{w}_n \wedge \boldsymbol{\varphi}$$

for some $\boldsymbol{\varphi} \in \bigwedge^{k-n} V$. Conclude that that $W' \subseteq W$, and that equality holds if and only if $\mathbf{A}$ is decomposable, in which case $W = W' = \text{span}\{\mathbf{f}_1 \ldots \mathbf{f}_k\}$.

[7] Such relations are called *syzygies*. A recipe for constructing linearly independent basis sets of isotropic tensors can be found in: G. F. Smith, *Tensor*, **19** (1968) 79-88.

b) Now show that W is the image space of $\bigwedge^{k-1} V^*$ under the map that takes

$$\mathbf{\Xi} = \Xi_{i_1 \ldots i_{k-1}} \mathbf{e}^{*i_1} \wedge \ldots \wedge \mathbf{e}^{*i_{k-1}} \in \bigwedge{}^{k-1} V^*$$

to

$$i(\mathbf{\Xi})\mathbf{A} \stackrel{\text{def}}{=} \Xi_{i_1 \ldots i_{k-1}} A^{i_1 \ldots i_{k-1} j} \mathbf{e}_j \in V$$

Deduce that the condition $W \subseteq W'$ is that

$$\Big(i(\mathbf{\Xi})\mathbf{A}\Big) \wedge \mathbf{A} = 0, \quad \forall\, \mathbf{\Xi} \in \bigwedge{}^{k-1} V^*.$$

c) By taking

$$\mathbf{\Xi} = \mathbf{e}^{*i_1} \wedge \ldots \wedge \mathbf{e}^{*i_{k-1}},$$

show that the condition in part b) can be written as

$$A^{i_1 \ldots i_{k-1} j_1} A^{j_2 j_3 \ldots j_{k+1}} \mathbf{e}_{j_1} \wedge \ldots \wedge \mathbf{e}_{j_{k+1}} = 0.$$

Deduce that the necessary and sufficient conditions for decomposibility are that

$$A^{i_1 \ldots i_{k-1} [j_1} A^{j_2 j_3 \ldots j_{k+1}]} = 0,$$

for all possible index sets $i_1, \ldots, i_{k-1}, j_1, \ldots j_{k+1}$. Here $[\ldots]$ denotes anti-symmetrization of the enclosed indices.

Chapter 2

Differential Calculus on Manifolds

In this section we will apply what we have learned about vectors and tensors in a linear space to the case of vector and tensor *fields* in a general curvilinear co-ordinate system. Our aim is to introduce the reader to the modern language of advanced calculus, and in particular to the calculus of differential forms on surfaces and manifolds.

2.1 Vector and Covector Fields

Vector fields — electric, magnetic, velocity fields, and so on — appear everywhere in physics. After perhaps struggling with it in introductory courses, we rather take the field concept for granted. There remain subtleties, however. Consider an electric field. It makes sense to add two field vectors at a single point, but there is no physical meaning to the sum of field vectors $\mathbf{E}(x_1)$ and $\mathbf{E}(x_2)$ at two distinct points. We should therefore regard all possible electric fields at a single point as living in a vector space, but each different point in space comes with its own field-vector space. This view seems even more reasonable when we consider velocity vectors describing motion on a curved surface.

A velocity vector lives in the *tangent space* to the surface at each point, and each of these spaces is a differently oriented subspace of the higher-dimensional ambient space.

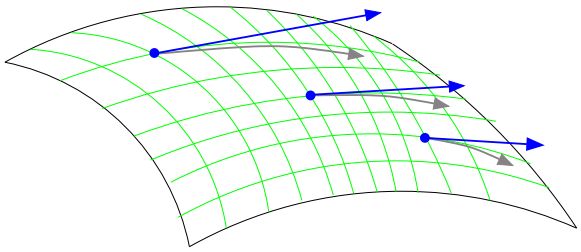

Figure 2.1: *Each point on a surface has its own vector space of tangents.*

Mathematicians call such a collection of vector spaces — one for each of the points in a surface — a *vector bundle* over the surface. Thus the *tangent bundle* over a surface is the totality of all vector spaces tangent to the surface. Why a bundle? This word is used because the individual tangent spaces are not completely independent, but are tied together in a rather non-obvious way. Try to construct a smooth field of unit vectors tangent to the surface of a sphere. However hard you work you will end up in trouble somewhere. You cannot comb a hairy ball. On the surface of torus you will have no problems. You can comb a hairy doughnut. The tangent spaces collectively know something about the surface they are tangent to.

Although we spoke in the previous paragraph of vectors tangent to a curved surface, it is useful to generalize this idea to vectors lying in the tangent space of an n-dimensional *manifold*. An n-manifold M is essentially a space that locally looks like a part of $\mathbb{R}^n$. This means that some open neighbourhood of each point can be parametrized by an n-dimensional co-ordinate system. Such a parametrization is called a *chart*. Unless M is $\mathbb{R}^n$ itself (or part of it), a chart will cover only part of M, and more than one will be required for complete coverage. Where a pair of charts overlap we demand that the transformation formula giving one set of co-ordinates as a function of the other be a smooth (C^∞) function, and to possess a smooth inverse.[1] A collection of such smoothly related co-ordinate charts covering all of M is called an *atlas*. The advantage of thinking in terms of manifolds is that we do not have to understand their properties as arising from some embedding in a higher dimensional space. Whatever structure they have, they possess in, and of, themselves

[1] A formal definition of a manifold contains some further technical restrictions (that the space be *Hausdorff* and *paracompact*) that are designed to eliminate pathologies. We are more interested in doing calculus than in proving theorems, and so we will ignore these niceties.

Classical mechanics provides a familiar illustration of these ideas. The configuration space M of a mechanical system is usually a manifold. When the system has n degrees of freedom we use generalized co-ordinates q^i, $i = 1, \ldots, n$ to parameterize M. The tangent bundle of M then provides the setting for Lagrangian mechanics. This bundle, denoted by TM, is the $2n$-dimensional space whose points consist of a point p in M together with a tangent vector lying in the tangent space TM_p at that point. If we think of the tangent vector as a velocity, the natural co-ordinates on TM become $(q^1, q^2, \ldots, q^n\,; \dot{q}^1, \dot{q}^2, \ldots, \dot{q}^n)$, and these are the variables that appear in the Lagrangian of the system.

If we consider a vector tangent to some curved surface, it will stick out of it. If we have a vector tangent to a manifold, it is a straight arrow lying atop bent co-ordinates. Should we restrict the length of the vector so that it does not stick out too far? Are we restricted to only infinitesimal vectors? It's best to avoid all this by inventing a clever notion of what a vector in a tangent space is. The idea is to focus on a well-defined object such as a derivative. Suppose our space has co-ordinates x^μ (These are *not* the contravariant components of some vector). A *directional derivative* is an object such as $X^\mu \partial_\mu$ where ∂_μ is shorthand for $\partial/\partial x^\mu$. When the numbers X^μ are functions of the co-ordinates x^σ, this object is called a tangent-vector field, and we write[2]

$$X = X^\mu \partial_\mu. \tag{2.1}$$

We regard the ∂_μ at a point x as a basis for TM_x, the tangent-vector space at x, and the $X^\mu(x)$ as the (contravariant) components of the vector X at that point. Although they are not little arrows, what the ∂_μ are is mathematically clear, and so we know perfectly well how to deal with them.

When we change co-ordinate system from x^μ to z^ν by regarding the x^μ's as invertable functions of the z^ν's, *i.e.*

$$\begin{aligned} x^1 &= x^1(z^1, z^2, \ldots, z^n), \\ x^2 &= x^2(z^1, z^2, \ldots, z^n), \\ &\vdots \\ x^n &= x^n(z^1, z^2, \ldots, z^n), \end{aligned} \tag{2.2}$$

[2]We are going to stop using bold symbols to distinguish between intrinsic objects and their components, because from now on almost everything will be something other than a number, and too much black ink would just be confusing.

then the chain rule for partial differentiation gives

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \frac{\partial z^\nu}{\partial x^\mu}\frac{\partial}{\partial z^\nu} = \left(\frac{\partial z^\nu}{\partial x^\mu}\right)\partial'_\nu, \tag{2.3}$$

where ∂'_ν is shorthand for $\partial/\partial z^\nu$. By demanding that

$$X = X^\mu \partial_\mu = X'^\nu \partial'_\nu \tag{2.4}$$

we find the components in the z^ν co-ordinate frame to be

$$X'^\nu = \left(\frac{\partial z^\nu}{\partial x^\mu}\right) X^\mu. \tag{2.5}$$

Conversely, using

$$\frac{\partial x^\sigma}{\partial z^\nu}\frac{\partial z^\nu}{\partial x^\mu} = \frac{\partial x^\sigma}{\partial x^\nu} = \delta^\sigma_\mu, \tag{2.6}$$

we have

$$X^\nu = \left(\frac{\partial x^\nu}{\partial z^\mu}\right) X'^\mu. \tag{2.7}$$

This, then, is the transformation law for a contravariant vector.

It is worth pointing out that the basis vectors ∂_μ are *not* unit vectors. As we have no metric, and therefore no notion of length anyway, we cannot try to normalize them. If you insist on drawing (small?) arrows, think of ∂_1 as starting at a point $(x^1, x^2, \ldots, x^n)$ and with its head at $(x^1 + 1, x^2, \ldots, x^n)$. Of course this is only a good picture if the co-ordinates are not too "curvy."

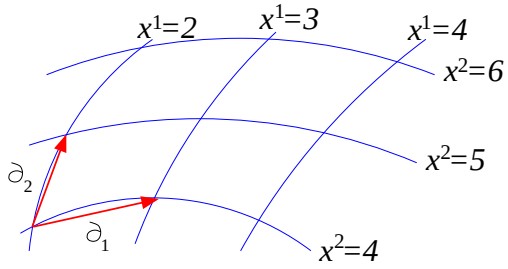

Figure 2.2: *Approximate picture of the vectors* ∂_1 *and* ∂_2 *at the point* $(x^1, x^2) = (2, 4)$.

Example: The surface of the unit sphere is a manifold. It is usually denoted by S^2. We may label its points with spherical polar co-ordinates θ and ϕ,

and these will be useful everywhere except at the north and south poles, where they become singular because at $\theta = 0$ or π all values of ϕ correspond to the same point. In this co-ordinate basis, the tangent vector representing the velocity field due to a rigid rotation of one radian per second about the z axis is

$$V_z = \partial_\phi. \tag{2.8}$$

Similarly

$$\begin{aligned} V_x &= -\sin\phi\,\partial_\theta - \cot\theta\cos\phi\,\partial_\phi, \\ V_y &= \cos\phi\,\partial_\theta - \cot\theta\sin\phi\partial_\phi, \end{aligned} \tag{2.9}$$

represent rigid rotations about the x and y axes.

We now know how to think about vectors. What about their dual-space partners, the covectors? These live in the *cotangent bundle* T^*M, and for them a cute notational game, due to Élie Cartan, is played. We write the basis vectors dual to the ∂_μ as $dx^\mu(\ )$. Thus

$$dx^\mu(\partial_\nu) = \delta^\mu_\nu. \tag{2.10}$$

When evaluated on a vector field $X = X^\mu\partial_\mu$, the basis covectors dx^μ return its components

$$dx^\mu(X) = dx^\mu(X^\nu\partial_\nu) = X^\nu dx^\mu(\partial_\nu) = X^\nu\delta^\mu_\nu = X^\mu. \tag{2.11}$$

Now, any smooth function $f \in C^\infty(M)$ will give rise to a field of covectors in T^*M. This is because a vector field X acts on the scalar function f as

$$Xf = X^\mu\partial_\mu f \tag{2.12}$$

and Xf is another scalar function. This new function gives a number — and thus an element of the field $\mathbb{R}$ — at each point $x \in M$. But this is exactly what a covector does: it takes in a vector at a point and returns a number. We will call this covector field "df." It is essentially the gradient of f. Thus

$$df(X) \stackrel{\text{def}}{=} Xf = X^\mu\frac{\partial f}{\partial x^\mu}. \tag{2.13}$$

If we take f to be the co-ordinate x^ν, we have

$$dx^\nu(X) = X^\mu\frac{\partial x^\nu}{\partial x^\mu} = X^\mu\delta^\nu_\mu = X^\nu, \tag{2.14}$$

so this viewpoint is consistent with our previous definition of dx^ν. Thus

$$df(X) = \frac{\partial f}{\partial x^\mu} X^\mu = \frac{\partial f}{\partial x^\mu} dx^\mu(X) \tag{2.15}$$

for any vector field X. In other words, we can expand df as

$$df = \frac{\partial f}{\partial x^\mu} dx^\mu. \tag{2.16}$$

This is *not* some approximation to a change in f, but is an exact expansion of the covector field df in terms of the basis covectors dx^μ.

We may retain something of the notion that dx^μ represents the (contravariant) components of a small displacement in x provided that we think of dx^μ as a machine into which we insert the small displacement (a vector) and have it spit out the numerical components δx^μ. This is the same distinction that we make between $\sin(\ )$ as a function into which one can plug x, and $\sin x$, the number that results from inserting in this particular value of x. Although seemingly innocent, we know that it is a distinction of great power.

The change of co-ordinates transformation law for a covector field f_μ is found from

$$f_\mu \, dx^\mu = f'_\nu \, dz^\nu, \tag{2.17}$$

by using

$$dx^\mu = \left(\frac{\partial x^\mu}{\partial z^\nu} \right) dz^\nu. \tag{2.18}$$

We find

$$f'_\nu = \left(\frac{\partial x^\mu}{\partial z^\nu} \right) f_\mu. \tag{2.19}$$

A general tensor such as $Q^{\lambda\mu}{}_{\rho\sigma\tau}$ transforms as

$$Q'^{\lambda\mu}{}_{\rho\sigma\tau}(z) = \frac{\partial z^\lambda}{\partial x^\alpha} \frac{\partial z^\mu}{\partial x^\beta} \frac{\partial x^\gamma}{\partial z^\rho} \frac{\partial x^\delta}{\partial z^\sigma} \frac{\partial x^\epsilon}{\partial z^\tau} Q^{\alpha\beta}{}_{\gamma\delta\epsilon}(x). \tag{2.20}$$

Observe how the indices are wired up: Those for the new tensor coefficients in the new co-ordinates, z, are attached to the new z's, and those for the old coefficients are attached to the old x's. Upstairs indices go in the numerator of each partial derivative, and downstairs ones are in the denominator.

The language of bundles and sections

At the beginning of this section, we introduced the notion of a vector bundle. This is a particular example of the more general concept of a *fibre bundle*, where the vector space at each point in the manifold is replaced by a "fibre" *over* that point. The fibre can be any mathematical object, such as a set, tensor space, or another manifold. Mathematicians visualize the bundle as a collection of fibres growing out of the manifold, much as stalks of wheat grow out the soil. When one slices through a patch of wheat with a scythe, the blade exposes a cross-section of the stalks. By analogy, a choice of an element of the the fibre over each point in the manifold is called a *cross-section*, or, more commonly, a *section* of the bundle. In this language a tangent-vector field becomes a section of the tangent bundle, and a field of covectors becomes a section of the cotangent bundle.

We provide a more detailed account of bundles in chapter 7.

2.2 Differentiating Tensors

If f is a function then $\partial_\mu f$ are components of the covariant vector df. Suppose that a^μ is a contravariant vector. Are $\partial_\nu a^\mu$ the components of a type $(1,1)$ tensor? The answer is *no*! In general, differentiating the components of a tensor does not give rise to another tensor. One can see why at two levels:

a) Consider the transformation laws. They contain expressions of the form $\partial x^\mu/\partial z^\nu$. If we differentiate both sides of the transformation law of a tensor, these factors are also differentiated, but tensor transformation laws never contain second derivatives, such as $\partial^2 x^\mu/\partial z^\nu \partial z^\sigma$.

b) Differentiation requires subtracting vectors or tensors at different points — but vectors at different points are in different vector spaces, so their difference is not defined.

These two reasons are really one and the same. We need to be cleverer to get new tensors by differentiating old ones.

2.2.1 Lie Bracket

One way to proceed is to note that the vector field X is an *operator*. It makes sense, therefore, to try to compose two of them to make another. Look at

XY, for example:

$$XY = X^\mu \partial_\mu (Y^\nu \partial_\nu) = X^\mu Y^\nu \partial^2_{\mu\nu} + X^\mu \left(\frac{\partial Y^\nu}{\partial x^\mu} \right) \partial_\nu. \tag{2.21}$$

What are we to make of this? Not much! There is no particular interpretation for the second derivative, and as we saw above, it does not transform nicely. But suppose we take a *commutator*:

$$[X, Y] = XY - YX = (X^\mu (\partial_\mu Y^\nu) - Y^\mu (\partial_\mu X^\nu)) \, \partial_\nu. \tag{2.22}$$

The second derivatives have cancelled, and what remains is a directional derivative and so a *bona-fide* vector field. The components

$$[X, Y]^\nu \equiv X^\mu (\partial_\mu Y^\nu) - Y^\mu (\partial_\mu X^\nu) \tag{2.23}$$

are the components of a new contravariant vector field made from the two old vector fields. It is called the *Lie bracket* of the two fields, and has a geometric interpretation.

To understand the geometry of the Lie bracket, we first define the *flow* associated with a tangent-vector field X. This is the map that takes a point x_0 and maps it to $x(t)$ by solving the family of equations

$$\frac{dx^\mu}{dt} = X^\mu (x^1, x^2, \ldots, x^d), \tag{2.24}$$

with initial condition $x^\mu(0) = x^\mu_0$. In words, we regard X as the velocity field of a flowing fluid, and let x ride along with the fluid.

Now envisage X and Y as two velocity fields. Suppose we flow along X for a brief time t, then along Y for another brief interval s. Next we switch back to X, but with a minus sign, for time t, and then to $-Y$ for a final interval of s. We have tried to retrace our path, but a short exercise with Taylor's theorem shows that we will fail to return to our exact starting point. We will miss by $\delta x^\mu = st[X, Y]^\mu$, plus corrections of cubic order in s and t.

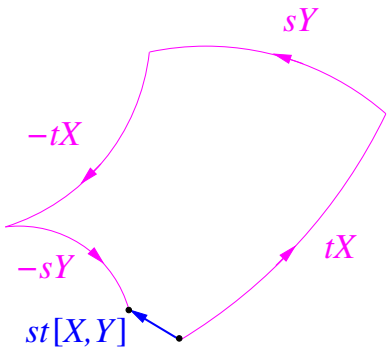

Figure 2.3: *The Lie bracket.*

Example: Let

$$\begin{aligned} V_x &= -\sin\phi\,\partial_\theta - \cot\theta\cos\phi\,\partial_\phi, \\ V_y &= \cos\phi\,\partial_\theta - \cot\theta\sin\phi\,\partial_\phi \end{aligned}$$

be two vector fields in $T(S^2)$. We find that

$$[V_x, V_y] = -V_z,$$

where $V_z = \partial_\phi$.

Frobenius' Theorem

Suppose that in some region of a d-dimensional manifold M we are given $n < d$ linearly independent tangent-vector fields X_i. Such a set is called a *distribution* by differential geometers. (The concept has nothing to do with probability, or with objects like "$\delta(x)$" which are also called "distributions.") At each point x, the span $\langle X_i(x)\rangle$ of the field vectors vectors forms a subspace of the tangent space TM_x, and we can picture this subspace as a fragment of an n-dimensional surface passing through x. It is possible that these surface fragments fit together to make a stack of smooth surfaces — called a *foliation* — that fill out the d-dimensional space, and have the given X_i as their tangent vectors.

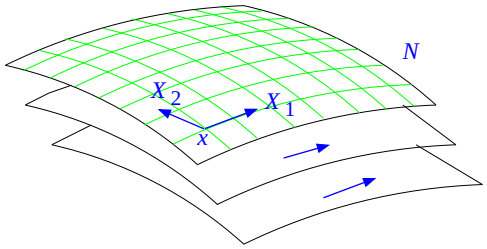

Figure 2.4: *A local foliation.*

If this is the case then starting from x and taking steps only along the X_i we find ourselves restricted to the n-surface, or *n-submanifold*, N passing though the original point x.

Alternatively, the surface fragments may form such an incoherent jumble that starting from x and moving only along the X_i we can find our way to any point in the neighbourhood of x. It is also possible that some intermediate case applies, so that moving along the X_i restricts us to an m-surface, where $d > m > n$. The Lie bracket provides us with the appropriate tool with which to investigate these possibilities.

First a definition: If there are functions $c_{ij}{}^{k}(x)$ such that

$$[X_i, X_j] = c_{ij}{}^{k}(x) X_k, \tag{2.25}$$

i.e. the Lie brackets close within the set $\{X_i\}$ at each point x, then the distribution is said to be *involutive.* When our given distribution is involutive, then the first case holds, and, at least locally, there is a foliation by n-submanifolds N. A formal statement of this is:
Theorem (Frobenius): A smooth (C^∞) involutive distribution is completely integrable: locally, there are co-ordinates $x^\mu, \mu = 1, \ldots, d$ such that $X_i = \sum_{\mu=1}^{n} X_i^\mu \partial_\mu$, and the surfaces N through each point are in the form $x^\mu =$ const. for $\mu = n+1, \ldots, d$. Conversely, if such co-ordinates exist then the distribution is involutive.
Sketch of Proof: If such co-ordinates exist then it is obvious that the Lie bracket of any pair of vectors in the form $X_i = \sum_{\mu=1}^{n} X_i^\mu \partial_\mu$ can also be expanded in terms of the first n basis vectors. A logically equivalent statement exploits the geometric interpretation of the Lie bracket: If the Lie brackets of the fields X_i do *not* close within the n-dimensional span of the X_i, then a sequence of back-and-forth manouvres along the X_i allows us to escape into a new direction, and so the X_i *cannot* be tangent to an n-surface. Establishing

the converse — that closure implies the existence of the foliation — is rather more technical, and we will not attempt it.

The physicist's version of Frobenius' theorem is usually expressed in the language of *holonomic* or *anholonomic* constraints.

For example, consider a particle moving in three dimensions. If we are told that the velocity vector is constrained to be perpendicular to the radius vector, *i.e.* $\mathbf{v} \cdot \mathbf{r} = 0$, we realize that the particle is being forced to move on a the sphere $|\mathbf{r}| = r_0$ passing through the initial point. In spherical co-ordinates the associated distribution is the set $\{\partial_\theta, \partial_\phi\}$, which is clearly involutive. The foliation is the family of nested spheres whose centre is the origin. The foliation is not global because it becomes singular at $r = 0$. Constraints like this, which restrict the motion to a surface, are called *holonomic*.

Suppose, on the other hand, we have a ball rolling on a table. Here, we have a five-dimensional configuration manifold $M = \mathbb{R}^2 \times S^3$ parameterized by the centre of mass $(x, y) \in \mathbb{R}^2$ of the ball and the three Euler angles $(\theta, \phi, \psi) \in S^3$ defining its orientation. Three no-slip rolling conditions

$$\begin{aligned} \dot{x} &= \dot{\psi} \sin\theta \sin\phi + \dot{\theta} \cos\phi, \\ \dot{y} &= -\dot{\psi} \sin\theta \cos\phi + \dot{\theta} \sin\phi, \\ 0 &= \dot{\psi} \cos\theta + \dot{\phi}, \end{aligned} \tag{2.26}$$

(see exercise 2.17) link the rate of change of the Euler angles to the velocity of the centre of mass. At each point in this five-dimensional manifold we are free to roll the ball in two directions, and so might expect that the reachable configurations constitute a two-dimensional surface embedded in the full five-dimensional space. The two vector fields

$$\begin{aligned} \mathbf{roll_x} &= \partial_x - \sin\phi \cot\theta \, \partial_\phi + \cos\phi \, \partial_\theta + \operatorname{cosec}\theta \sin\phi \, \partial_\psi, \\ \mathbf{roll_y} &= \partial_y + \cos\phi \cot\theta \, \partial_\phi + \sin\phi \, \partial_\theta - \operatorname{cosec}\theta \cos\phi \, \partial_\psi, \end{aligned} \tag{2.27}$$

describing the x- and y-direction rolling motion are not in involution, however. By calculating enough Lie brackets we eventually obtain five linearly independent velocity vector fields, and starting from one configuration we can reach any other. The no-slip rolling condition is said to be *non-integrable*, or *anholonomic*. Such systems are tricky to deal with in Lagrangian dynamics.

For a d-dimensional mechanical system, a set of m independent constraints of the form $\omega^i_\mu(q)\dot{q}^\mu = 0$, $i = 1, \ldots, m$ determines an $n = d - m$

dimensional distribution. In terms of the vector $\dot{q} \equiv \dot{q}^\mu \partial_\mu$ and the covectors

$$\omega^i = \sum_{\mu=1}^{d} \omega^i_\mu(q) dq^\mu, \quad i = 1 \leq i \leq m \tag{2.28}$$

we can write the these constraints as $\omega^i(\dot{q}) = 0$. This is known a *Pfaffian* system of equations. The Pfaffian system is said to be *integrable* if the distribution it implicitly defines is in involution, and hence itself integrable. In this case there is a set of m functions $g^i(q)$ and an invertible m-by-m matrix $f^i{}_j(q)$ such that

$$\omega^i = \sum_{j=1}^{m} f^i{}_j(q) dg^j. \tag{2.29}$$

The functions $g^i(q)$ can, for example, be taken to be the co-ordinate functions x^μ, $\mu = n+1, \ldots, d$, that label the foliating surfaces N in the statement of Frobenius' theorem. The system of integrable constraints $\omega^i(\dot{q}) = 0$ thus restricts us to the surfaces $g^i(q) = constant$. Integrable constraints are therefore holonomic.

The following exercise provides a familiar example of the utility of non-holonomic constraints:

Exercise 2.1: *Parallel Parking using Lie Brackets.*

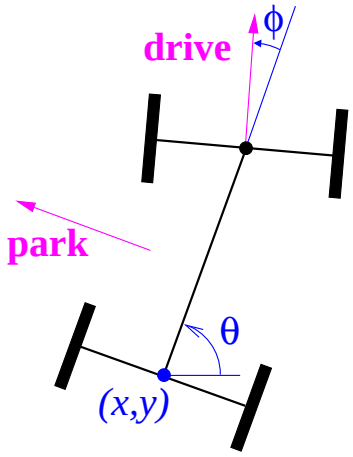

Figure 2.5: Co-ordinates for car parking

The configuration space of a car is four dimensional, and parameterized by co-ordinates (x, y, θ, ϕ), as shown in figure 2.5.

Define the following vector fields:

a) (front wheel) **drive** $= \cos\phi(\cos\theta\,\partial_x + \sin\theta\,\partial_y) + \sin\phi\,\partial_\theta$.
b) **steer** $= \partial_\phi$.
c) (front wheel) **skid** $= -\sin\phi(\cos\theta\,\partial_x + \sin\theta\,\partial_y) + \cos\phi\,\partial_\theta$.
d) **park** $= -\sin\theta\,\partial_x + \cos\theta\,\partial_y$.

Explain why these are apt names for the vector fields, and compute the Lie brackets:

$$[\mathbf{steer}, \mathbf{drive}], \quad [\mathbf{steer}, \mathbf{skid}], \quad [\mathbf{skid}, \mathbf{drive}],$$
$$[\mathbf{park}, \mathbf{drive}], \quad [\mathbf{park}, \mathbf{park}], \quad [\mathbf{park}, \mathbf{skid}].$$

The driver can use only the operations $(\pm)\,\mathbf{drive}$ and $(\pm)\,\mathbf{steer}$ to manouvre the car. Use the geometric interpretation of the Lie bracket to explain how a suitable sequence of motions (forward, reverse, and turning the steering wheel) can be used to manoeuvre a car sideways into a parking space.

2.2.2 Lie Derivative

Another derivative we can define is the *Lie derivative* along a vector field X. It is defined by its action on a scalar function f as

$$\mathcal{L}_X f \stackrel{\text{def}}{=} Xf, \tag{2.30}$$

on a vector field by

$$\mathcal{L}_X Y \stackrel{\text{def}}{=} [X, Y], \tag{2.31}$$

and on anything else by requiring it to be a *derivation*, meaning that it obeys Leibniz' rule. For example, let us compute the Lie derivative of a covector F. We first introduce an arbitrary vector field Y and plug it into F to get the scalar function $F(Y)$. Leibniz' rule is then the statement that

$$\mathcal{L}_X F(Y) = (\mathcal{L}_X F)(Y) + F(\mathcal{L}_X Y). \tag{2.32}$$

Since $F(Y)$ is a function and Y a vector, both of whose derivatives we know how to compute, we know two of the three terms in this equation. From $\mathcal{L}_X F(Y) = XF(Y)$ and $F(\mathcal{L}_X Y) = F([X, Y])$, we have

$$XF(Y) = (\mathcal{L}_X F)(Y) + F([X, Y]), \tag{2.33}$$

and so

$$(\mathcal{L}_X F)(Y) = XF(Y) - F([X,Y]). \tag{2.34}$$

In components, this becomes

$$\begin{aligned}(\mathcal{L}_X F)(Y) &= X^\nu \partial_\nu (F_\mu Y^\mu) - F_\nu (X^\mu \partial_\mu Y^\nu - Y^\mu \partial_\mu X^\nu) \\ &= (X^\nu \partial_\nu F_\mu + F_\nu \partial_\mu X^\nu) Y^\mu. \end{aligned} \tag{2.35}$$

Note how all the derivatives of Y^μ have cancelled, so $\mathcal{L}_X F(\ )$ depends only on the local value of Y. The Lie derivative of F is therefore still a covector field. This is true in general: the Lie derivative does not change the tensor character of the objects on which it acts. Dropping the passive spectator field Y^ν, we have a formula for $\mathcal{L}_X F$ in components:

$$(\mathcal{L}_X F)_\mu = X^\nu \partial_\nu F_\mu + F_\nu \partial_\mu X^\nu. \tag{2.36}$$

Another example is provided by the Lie derivative of a type $(0,2)$ tensor, such as a metric tensor. This is

$$(\mathcal{L}_X g)_{\mu\nu} = X^\alpha \partial_\alpha g_{\mu\nu} + g_{\mu\alpha}\partial_\nu X^\alpha + g_{\alpha\nu}\partial_\mu X^\alpha. \tag{2.37}$$

The Lie derivative of a metric measures the extent to which the displacement $x^\alpha \to x^\alpha + \epsilon X^\alpha(x)$ deforms the geometry. If we write the metric as

$$g(\ ,\ ) = g_{\mu\nu}(x)\, dx^\mu \otimes dx^\nu, \tag{2.38}$$

we can understand both this geometric interpretation and the origin of the three terms appearing in the Lie derivative. We simply make the displacement $x^\alpha \to x^\alpha + \epsilon X^\alpha$ in the coefficients $g_{\mu\nu}(x)$ and in the two dx^α. In the latter we write

$$d(x^\alpha + \epsilon X^\alpha) = dx^\alpha + \epsilon \frac{\partial X^\alpha}{\partial x^\beta} dx^\beta. \tag{2.39}$$

Then we see that

$$\begin{aligned} g_{\mu\nu}(x)\, dx^\mu \otimes dx^\nu &\to [g_{\mu\nu}(x) + \epsilon(X^\alpha \partial_\alpha g_{\mu\nu} + g_{\mu\alpha}\partial_\nu X^\alpha + g_{\alpha\nu}\partial_\mu X^\alpha)]\ dx^\mu \otimes dx^\nu \\ &= [g_{\mu\nu} + \epsilon(\mathcal{L}_X g)_{\mu\nu}]\, dx^\mu \otimes dx^\nu. \end{aligned} \tag{2.40}$$

A displacement field X that does not change distances between points, *i.e.* one that gives rise to an *isometry*, must therefore satisfy $\mathcal{L}_X g = 0$. Such an X is said to be a *Killing field* after Wilhelm Killing who introduced them in his study of non-euclidean geometries.

The geometric interpretation of the Lie derivative of a vector field is as follows: In order to compute the X directional derivative of a vector field Y, we need to be able to subtract the vector $Y(x)$ from the vector $Y(x+\epsilon X)$, divide by ϵ, and take the limit $\epsilon \to 0$. To do this we have somehow to get the vector $Y(x)$ from the point x, where it normally resides, to the new point $x+\epsilon X$, so both vectors are elements of the same vector space. The Lie derivative achieves this by carrying the old vector to the new point along the field X.

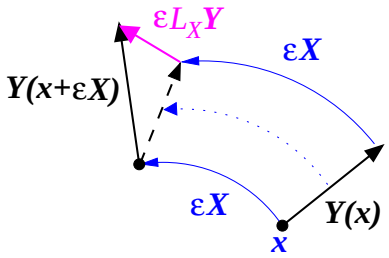

Figure 2.6: Computing the Lie derivative of a vector.

Imagine the vector Y as drawn in ink in a flowing fluid whose velocity field is X. Initially the tail of Y is at x and its head is at $x+Y$. After flowing for a time ϵ, its tail is at $x+\epsilon X$ — *i.e* exactly where the tail of $Y(x+\epsilon X)$ lies. Where the head of transported vector ends up depends how the flow has stretched and rotated the ink, but it is this distorted vector that is subtracted from $Y(x+\epsilon X)$ to get $\epsilon \mathcal{L}_X Y = \epsilon[X,Y]$.

Exercise 2.2: The metric on the unit sphere equipped with polar co-ordinates is

$$g(\ \ ,\ \ ) = d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi.$$

Consider

$$V_x = -\sin\phi\partial_\theta - \cot\theta\cos\phi\partial_\phi,$$

the vector field of a rigid rotation about the x axis. Compute the Lie derivative $\mathcal{L}_{V_x} g$, and show that it is zero.

Exercise 2.3: Suppose we have an unstrained block of material in real space. A co-ordinate system ξ^1, ξ^2, ξ^3, is attached to the atoms of the body. The point with co-ordinate ξ is located at $(x^1(\xi), x^2(\xi), x^3(\xi))$ where x^1, x^2, x^3 are the usual $\mathbf{R^3}$ Cartesian co-ordinates.

a) Show that the induced metric in the ξ co-ordinate system is

$$g_{\mu\nu}(\xi) = \sum_{a=1}^{3} \frac{\partial x^a}{\partial \xi^\mu} \frac{\partial x^a}{\partial \xi^\nu}.$$

b) The body is now deformed by an infinitesimal strain vector field $\eta(\xi)$. The atom with co-ordinate ξ^μ is moved to what was $\xi^\mu + \eta^\mu(\xi)$, or equivalently, the atom initially at Cartesian co-ordinate $x^a(\xi)$ is moved to $x^a + \eta^\mu \partial x^a / \partial \xi^\mu$. Show that the new induced metric is

$$g_{\mu\nu} + \delta g_{\mu\nu} = g_{\mu\nu} + \mathcal{L}_\eta g_{\mu\nu}.$$

c) Define the *strain tensor* to be 1/2 of the Lie derivative of the metric with respect to the deformation. If the original ξ co-ordinate system coincided with the Cartesian one, show that this definition reduces to the familiar form

$$e_{ab} = \frac{1}{2}\left(\frac{\partial \eta_a}{\partial x^b} + \frac{\partial \eta_b}{\partial x^a}\right),$$

all tensors being Cartesian.

d) Part c) gave us the geometric definitition of *infinitesimal strain*. If the body is deformed substantially, the *Cauchy-Green finite strain tensor* is defined as

$$E_{\mu\nu}(\xi) = \frac{1}{2}\left(g_{\mu\nu} - g^{(0)}_{\mu\nu}\right),$$

where $g^{(0)}_{\mu\nu}$ is the metric in the undeformed body and $g_{\mu\nu}$ that of the deformed body. Explain why this is a reasonable definition.

2.3 Exterior Calculus

2.3.1 Differential Forms

The objects we introduced in section 2.1, the dx^μ, are called one-forms, or differential one-forms. They are fields living in the cotangent bundle T^*M of M. More precisely, they are *sections* of the cotangent bundle. Sections of the bundle whose fibre above $x \in M$ is the p-th skew-symmetric tensor power $\bigwedge^p(T^*M_x)$ of the cotangent space are known as p-forms.

For example,

$$A = A_\mu dx^\mu = A_1 dx^1 + A_2 dx^2 + A_3 dx^3, \tag{2.41}$$

is a 1-form,

$$F = \frac{1}{2}F_{\mu\nu}dx^\mu \wedge dx^\nu = F_{12}dx^1 \wedge dx^2 + F_{23}dx^2 \wedge dx^3 + F_{31}dx^3 \wedge dx^1, \quad (2.42)$$

is a 2-form, and

$$\begin{aligned} \Omega &= \frac{1}{3!}\Omega_{\mu\nu\sigma}\, dx^\mu \wedge dx^\nu \wedge dx^\sigma \\ &= \Omega_{123}\, dx^1 \wedge dx^2 \wedge dx^3, \end{aligned} \quad (2.43)$$

is a 3-form. All the coefficients are skew-symmetric tensors, so, for example,

$$\Omega_{\mu\nu\sigma} = \Omega_{\nu\sigma\mu} = \Omega_{\sigma\mu\nu} = -\Omega_{\nu\mu\sigma} = -\Omega_{\mu\sigma\nu} = -\Omega_{\sigma\nu\mu}. \quad (2.44)$$

In each example we have explicitly written out all the independent terms for the case of three dimensions. Note how the $p!$ disappears when we do this and keep only distinct components. In d dimensions the space of p-forms is $d!/p!(d-p)!$ dimensional, and all p-forms with $p > d$ vanish identically.

As with the wedge products in chapter one, we regard a p-form as a p-linear skew-symetric function with p slots into which we can drop vectors to get a number. For example the basis two-forms give

$$dx^\mu \wedge dx^\nu(\partial_\alpha, \partial_\beta) = \delta^\mu_\alpha\delta^\nu_\beta - \delta^\mu_\beta\delta^\nu_\alpha. \quad (2.45)$$

The analogous expression for a p-form would have $p!$ terms. We can define an algebra of differential forms by "wedging" them together in the obvious way, so that the product of a p form with a q form is a $(p+q)$-form. The wedge product is associative and distributive but not, of course, commutative. Instead, if a is a p-form and b a q-form, then

$$a \wedge b = (-1)^{pq}\, b \wedge a. \quad (2.46)$$

Actually it is customary in this game to suppress the "$\wedge$" and simply write $F = \frac{1}{2}F_{\mu\nu}\, dx^\mu dx^\nu$, it being assumed that you know that $dx^\mu dx^\nu = -dx^\nu dx^\mu$ — what else could it be?

2.3.2 The Exterior Derivative

These p-forms may seem rather complicated, so it is perhaps surprising that all the vector calculus (div, grad, curl, the divergence theorem and Stokes'

theorem, *etc.*) that you have learned in the past reduce, in terms of them, to two simple formulæ! Indeed Élie Cartan's calculus of p-forms is slowly supplanting traditional vector calculus, much as Willard Gibbs' and Oliver Heaviside's vector calculus supplanted the tedious component-by-component formulæ you find in Maxwell's *Treatise on Electricity and Magnetism*.

The basic tool is the *exterior derivative* "d", which we now define axiomatically:

i) If f is a function (0-form), then df coincides with the previous definition, *i.e.* $df(X) = Xf$ for any vector field X.

ii) d is an *anti-derivation*: If a is a p-form and b a q-form then

$$d(a \wedge b) = da \wedge b + (-1)^p a \wedge db. \tag{2.47}$$

iii) *Poincaré's lemma*: $d^2 = 0$, meaning that $d(da) = 0$ for any p-form a.

iv) d is linear. That $d(\alpha a) = \alpha da$, for constant α follows already from i) and ii), so the new fact is that $d(a + b) = da + db$.

It is not immediately obvious that axioms i), ii) and iii) are compatible with one another. If we use axiom i), ii) and $d(dx^i) = 0$ to compute the d of $\Omega = \frac{1}{p!}\Omega_{i_1,\ldots,i_p} dx^{i_1} \cdots dx^{i_p}$, we find

$$\begin{aligned} d\Omega &= \frac{1}{p!}(d\Omega_{i_1,\ldots,i_p})\, dx^{i_1} \cdots dx^{i_p} \\ &= \frac{1}{p!}\partial_k \Omega_{i_1,\ldots,i_p}\, dx^k dx^{i_1} \cdots dx^{i_p}. \end{aligned} \tag{2.48}$$

Now compute

$$d(d\Omega) = \frac{1}{p!}\left(\partial_l \partial_k \Omega_{i_1,\ldots,i_p}\right) dx^l dx^k dx^{i_1} \cdots dx^{i_p}. \tag{2.49}$$

Fortunately this is zero because $\partial_l \partial_k \Omega = \partial_k \partial_l \Omega$, while $dx^l dx^k = -dx^k dx^l$.
If $A = A_1 dx^1 + A_2 dx^2 + A_3 dx^3$ then

$$\begin{aligned} dA &= \left(\frac{\partial A_2}{\partial x^1} - \frac{\partial A_1}{\partial x^2}\right) dx^1 dx^2 + \left(\frac{\partial A_1}{\partial x^3} - \frac{\partial A_3}{\partial x^1}\right) dx^3 dx^1 + \left(\frac{\partial A_3}{\partial x^2} - \frac{\partial A_2}{\partial x^3}\right) dx^2 dx^3 \\ &= \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu, \end{aligned} \tag{2.50}$$

where

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.51}$$

You will recognize the components of curl $\mathbf{A}$ hiding in here.

Similarly, if $F = F_{12}dx^1dx^2 + F_{23}dx^2dx^3 + F_{31}dx^3dx^1$ then

$$dF = \left(\frac{\partial F_{23}}{\partial x^1} + \frac{\partial F_{31}}{\partial x^2} + \frac{\partial F_{12}}{\partial x^3}\right) dx^1dx^2dx^3. \tag{2.52}$$

This looks like a divergence.

The axiom $d^2 = 0$ encompasses both "curl grad $= 0$" and "div curl $= 0$", together with an infinite number of higher-dimensional analogues. The familiar "curl $=\nabla\times$", meanwhile, is only defined in three dimensional space.

The exterior derivative takes p-forms to $(p+1)$-forms *i.e.* skew-symmetric type $(0, p)$ tensors to skew-symmetric $(0, p+1)$ tensors. How does "d" get around the fact that the derivative of a tensor is not a tensor? Well, if you apply the transformation law for A_μ, and the chain rule to $\frac{\partial}{\partial x^\mu}$ to find the transformation law for $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, you will see why: all the derivatives of the $\frac{\partial z^\nu}{\partial x^\mu}$ cancel, and $F_{\mu\nu}$ is a *bona-fide* tensor of type $(0, 2)$. This sort of cancellation is why skew-symmetric objects are useful, and symmetric ones less so.

Exercise 2.4: Use axiom ii) to compute $d(d(a \wedge b))$ and confirm that it is zero.

Closed and exact forms

The Poincaré lemma. $d^2 = 0$, leads to some important terminology:

i) A p-form ω is said to be *closed* if $d\omega = 0$.

ii) A p-form ω is said to *exact* if $\omega = d\eta$ for some $(p-1)$-form η.

An exact form is necessarily closed, but a closed form is not necessarily exact. The question of when closed $\Rightarrow$ exact is one involving the global topology of the space in which the forms are defined, and will be subject of chapter 4.

Cartan's formulæ

It is sometimes useful to have expressions for the action of d coupled with the evaluation of the subsequent $(p+1)$ forms.

If f, η, ω, are $0, 1, 2$-forms, respectively, then $df, d\eta, d\omega$, are $1, 2, 3$-forms. When we plug in the appropriate number of vector fields X, Y, Z, then, after some labour, we will find

$$df(X) \;=\; Xf. \tag{2.53}$$

$$d\eta(X,Y) = X\eta(Y) - Y\eta(X) - \eta([X,Y]). \tag{2.54}$$

$$\begin{aligned} d\omega(X,Y,Z) &= X\omega(Y,Z) + Y\omega(Z,X) + Z\omega(X,Y) \\ &\quad -\omega([X,Y],Z) - \omega([Y,Z],X) - \omega([Z,X],Y). \end{aligned} \tag{2.55}$$

These formulæ, and their higher-p analogues, express d in terms of geometric objects, and so make it clear that the exterior derivative is itself a geometric object, independent of any particular co-ordinate choice.

Let us demonstate the correctness of the second formula. With $\eta = \eta_\mu dx^\mu$, the left-hand side, $d\eta(X,Y)$, is equal to

$$\partial_\mu \eta_\nu \, dx^\mu dx^\nu(X,Y) = \partial_\mu \eta_\nu (X^\mu Y^\nu - X^\nu Y^\mu). \tag{2.56}$$

The right hand side is equal to

$$X^\mu \partial_\mu(\eta_\nu Y^\nu) - Y^\mu \partial_\mu(\eta_\nu X^\nu) - \eta_\nu(X^\mu \partial_\mu Y^\nu - Y^\mu \partial_\mu X^\nu). \tag{2.57}$$

On using the product rule for the derivatives in the first two terms, we find that all derivatives of the components of X and Y cancel, and are left with exactly those terms appearing on left.

Exercise 2.5: Let ω^i, $i = 1, \ldots, r$ be a linearly independent set of one-forms defining a Pfaffian system (see sec. 2.2.1) in d dimensions.

i) Use Cartan's formulæ to show that the corresponding $(d-r)$-dimensional distribution is involutive if and only if there is an r-by-r matrix of 1-forms ${\theta^i}_j$ such that

$$d\omega^i = \sum_{j=1}^{r} {\theta^i}_j \wedge \omega^j.$$

ii) Show that the conditions in part i) are satisfied if there are r functions g^i and an invertible r-by-r matrix of functions ${f^i}_j$ such that

$$\omega^i = \sum_{j=1}^{r} {f^i}_j dg^i.$$

In this case foliation surfaces are given by the conditions $g^i(x) = \text{const.}$, $i = 1, \ldots, r$.

It is also possible, but considerably harder, to show that i) $\Rightarrow$ ii). Doing so would constitute a proof of Frobenius' theorem.

Exercise 2.6: Let ω be a closed two-form, and let $\text{Null}(\omega)$ be the space of vector fields X such that $\omega(X, \ ) = 0$. Use the Cartan formulæ to show that if $X, Y \in \text{Null}(\omega)$, then $[X,Y] \in \text{Null}(\omega)$.

Lie Derivative of Forms

Given a p-form ω and a vector field X, we can form a $(p-1)$-form called $i_X\omega$ by writing

$$i_X\omega(\underbrace{\ldots\ldots}_{p-1\,\text{slots}}) = \omega(\overbrace{X, \underbrace{\ldots\ldots}_{p-1\,\text{slots}}}^{p\,\text{slots}}). \tag{2.58}$$

Acting on a 0-form, i_X is defined to be 0. This procedure is called the *interior multiplication* by X. It is simply a contraction

$$\omega_{j_i j_2 \ldots j_p} \rightarrow \omega_{k j_2 \ldots j_p} X^k, \tag{2.59}$$

but it is convenient to have a special symbol for this operation. It is perhaps surprising that i_X turns out to be an anti-derivation, just as is d. If η and ω are p and q forms respectively, then

$$i_X(\eta \wedge \omega) = (i_X\eta) \wedge \omega + (-1)^p \eta \wedge (i_X\omega), \tag{2.60}$$

even though i_X involves no differentiation. For example, if $X = X^\mu \partial_\mu$, then

$$\begin{aligned} i_X(dx^\mu \wedge dx^\nu) &= dx^\mu \wedge dx^\nu(X^\alpha \partial_\alpha, \quad), \\ &= X^\mu dx^\nu - dx^\mu X^\nu, \\ &= (i_X dx^\mu) \wedge (dx^\nu) - dx^\mu \wedge (i_X dx^\nu). \end{aligned} \tag{2.61}$$

One reason for introducing i_X is that there is a nice (and profound) formula for the Lie derivative of a p-form in terms of i_X. The formula is called the *infinitesimal homotopy relation*. It reads

$$\mathcal{L}_X\omega = (d\, i_X + i_X d)\omega. \tag{2.62}$$

This formula is proved by verifying that it is true for functions and one-forms, and then showing that it is a derivation – in other words that it satisfies Leibniz' rule. From the derivation property of the Lie derivative, we immediately deduce that that the formula works for any p-form.

That the formula is true for functions should be obvious: Since $i_X f = 0$ by definition, we have

$$(d\, i_X + i_X d)f = i_X df = df(X) = Xf = \mathcal{L}_X f. \tag{2.63}$$

To show that the formula works for one forms, we evaluate

$$\begin{aligned}(d\,i_X + i_X d)(f_\nu\, dx^\nu) &= d(f_\nu X^\nu) + i_X(\partial_\mu f_\nu\, dx^\mu dx^\nu) \\ &= \partial_\mu(f_\nu X^\nu)dx^\mu + \partial_\mu f_\nu(X^\mu dx^\nu - X^\nu dx^\mu) \\ &= (X^\nu \partial_\nu f_\mu + f_\nu \partial_\mu X^\nu)dx^\mu. \qquad (2.64)\end{aligned}$$

In going from the second to the third line, we have interchanged the dummy labels $\mu \leftrightarrow \nu$ in the term containing dx^ν. We recognize that the 1-form in the last line is indeed $\mathcal{L}_X f$.

To show that $di_X + i_X d$ is a derivation we must apply $d\,i_X + i_X d$ to $a \wedge b$ and use the anti-derivation property of i_x and d. This is straightforward once we recall that d takes a p-form to a $(p+1)$-form while i_X takes a p-form to a $(p-1)$-form.

Exercise 2.7: Let

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

Use the anti-derivation property of i_X to show that

$$i_X\omega = \frac{1}{(p-1)!}\,\omega_{\alpha i_2 \ldots i_p} X^\alpha dx^{i_2} \cdots dx^{i_p},$$

and so verify the equivalence of (2.58) and (2.59).

Exercise 2.8: Use the infinitesimal homotopy relation to show that $\mathcal{L}$ and d commute, *i.e.* for ω a p-form, we have

$$d\,(\mathcal{L}_X\omega) = \mathcal{L}_X(d\omega).$$

2.4 Physical Applications

2.4.1 Maxwell's Equations

In relativistic[3] four-dimensional tensor notation the two source-free Maxwell's equations

$$\begin{aligned}\mathrm{curl}\,\mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\ \mathrm{div}\,\mathbf{B} &= 0, \qquad (2.65)\end{aligned}$$

[3]In this section we will use units in which $c = \epsilon_0 = \mu_0 = 1$. We take the Minkowski metric to be $g_{\mu\nu} = \mathrm{diag}\,(-1, 1, 1, 1)$ where $x^0 = t$, $x^1 = x$, *etc.*

reduce to the single equation

$$\frac{\partial F_{\mu\nu}}{\partial x^\lambda} + \frac{\partial F_{\nu\lambda}}{\partial x^\mu} + \frac{\partial F_{\lambda\mu}}{\partial x^\nu} = 0. \tag{2.66}$$

where

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{2.67}$$

The "F" is traditional, for Michael Faraday. In form language, the relativistic equation becomes the even more compact expression $dF = 0$, where

$$\begin{aligned} F &\equiv \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu \\ &= B_x dydz + B_y dzdx + B_z dxdy + E_x dxdt + E_y dydt + E_z dzdt, \end{aligned} \tag{2.68}$$

is a Minkowski-space 2-form.

Exercise 2.9: Verify that the source-free Maxwell equations are indeed equivalent to $dF = 0$.

The equation $dF = 0$ is automatically satisfied if we introduce a 4-vector 1-form potential $A = -\phi dt + A_x dx + A_y dy + A_z dz$ and set $F = dA$.

The two Maxwell equations with sources

$$\begin{aligned} \operatorname{div} \mathbf{D} &= \rho, \\ \operatorname{curl} \mathbf{H} &= \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t}, \end{aligned} \tag{2.69}$$

reduce in 4-tensor notation to the single equation

$$\partial_\mu F^{\mu\nu} = J^\nu. \tag{2.70}$$

Here $J^\mu = (\rho, \mathbf{j})$ is the current 4-vector.

This source equation takes a little more work to express in form language, but it can be done. We need a new concept: the *Hodge "star" dual* of a form. In d dimensions the "$\star$" map takes a p-form to a $(d-p)$-form. It depends on both the metric and the *orientation*. The latter means a canonical choice of the order in which to write our basis forms, with orderings that differ

by an even permutation being counted as the same. The full d-dimensional definition involves the Levi-Civita duality operation of chapter 1 , combined with the use of the metric tensor to raise indices. Recall that $\sqrt{g} = \sqrt{\det g_{\mu\nu}}$. (In Minkowski-signature metrics we should replace $\sqrt{g}$ by $\sqrt{-g}$.) We define "$\star$" to be a linear map

$$\star : \bigwedge^{p}(T^*M) \to \bigwedge^{(d-p)}(T^*M) \tag{2.71}$$

such that

$$\star\, dx^{i_1} \dots dx^{i_p} \stackrel{\text{def}}{=} \frac{1}{(d-p)!}\sqrt{g} g^{i_1 j_1} \dots g^{i_p j_p} \epsilon_{j_1 \cdots j_p j_{p+1} \cdots j_d} dx^{j_{p+1}} \dots dx^{j_d}. \tag{2.72}$$

Although this definition looks a trifle involved, computations involving it are not so intimidating. The trick is to work, whenever possible, with oriented orthonormal frames. If we are in euclidean space and $\{\mathbf{e}^{*i_1}, \mathbf{e}^{*i_2}, \dots, \mathbf{e}^{*i_d}\}$ is an ordering of the orthonormal basis for $(T^*M)_x$ whose orientation is equivalent to $\{\mathbf{e}^{*1}, \mathbf{e}^{*2}, \dots, \mathbf{e}^{*d}\}$ then

$$\star\,(\mathbf{e}^{*i_1} \wedge \mathbf{e}^{*i_2} \wedge \cdots \wedge \mathbf{e}^{*i_p}) = \mathbf{e}^{*i_{p+1}} \wedge \mathbf{e}^{*i_{p+2}} \wedge \cdots \wedge \mathbf{e}^{*i_d}. \tag{2.73}$$

For example, in three dimensions, and with x, y, z, our usual Cartesian coordinates, we have

$$\begin{aligned} \star\, dx &= dydz, \\ \star\, dy &= dzdx, \\ \star\, dz &= dxdy. \end{aligned} \tag{2.74}$$

An analogous method works for Minkowski-signature $(-,+,+,+)$ metrics, except that now we must include a minus sign for each negatively normed dt factor in the form being "starred." Taking $\{dt, dx, dy, dz\}$ as our oriented basis, we therefore find[4]

$$\begin{aligned} \star\, dxdy &= -dzdt, \\ \star\, dydz &= -dxdt, \\ \star\, dzdx &= -dydt, \\ \star\, dxdt &= dydz, \\ \star\, dydt &= dzdx, \\ \star\, dzdt &= dxdy. \end{aligned} \tag{2.75}$$

[4]See for example: Misner, Thorn and Wheeler, *Gravitation*, (MTW) page 108.

For example, the first of these equations is derived by observing that $(dxdy)(-dzdt) = dtdxdydz$, and that there is no "dt" in the product $dxdy$. The fourth follows from observing that that $(dxdt)(-dydx) = dtdxdydz$, but there is a negative-normed "dt" in the product $dxdt$.

The $\star$ map is constructed so that if

$$\alpha = \frac{1}{p!}\alpha_{i_1 i_2 \ldots i_p} dx^{i_1} dx^{i_2} \cdots dx^{i_p}, \tag{2.76}$$

and

$$\beta = \frac{1}{p!}\beta_{i_1 i_2 \ldots i_p} dx^{i_1} dx^{i_2} \cdots dx^{i_p}, \tag{2.77}$$

then

$$\alpha \wedge (\star\beta) = \beta \wedge (\star\alpha) = \langle \alpha, \beta \rangle \, \sigma, \tag{2.78}$$

where the inner product $\langle \alpha, \beta \rangle$ is defined to be the invariant

$$\langle \alpha, \beta \rangle = \frac{1}{p!} g^{i_1 j_1} g^{i_2 j_2} \cdots g^{i_p j_p} \alpha_{i_1 i_2 \ldots i_p} \beta_{j_1 j_2 \ldots j_p}, \tag{2.79}$$

and σ is the *volume form*

$$\sigma = \sqrt{g}\, dx^1 dx^2 \cdots dx^d. \tag{2.80}$$

In future we will write $\alpha \star \beta$ for $\alpha \wedge (\star\beta)$. Bear in mind that the "$\star$" in this expression is acting β and is not some new kind of binary operation.

We now apply these ideas to Maxwell. From the field-strength 2-form

$$F = B_x dydz + B_y dzdx + B_z dxdy + E_x dxdt + E_y dydt + E_z dzdt, \tag{2.81}$$

we get a dual 2-form

$$\star F = -B_x dxdt - B_y dydt - B_z dzdt + E_x dydz + E_y dzdx + E_z dxdy. \tag{2.82}$$

We can check that we have correctly computed the Hodge star of F by taking the wedge product, for which we find

$$F \star F = \frac{1}{2}(F_{\mu\nu} F^{\mu\nu})\sigma = (B_x^2 + B_y^2 + B_z^2 - E_x^2 - E_y^2 - E_z^2) dtdxdydz. \tag{2.83}$$

Observe that the expression $B^2 - E^2$ is a Lorentz scalar. Similarly, from the current 1-form

$$J \equiv J_\mu dx^\mu = -\rho\, dt + j_x dx + j_y dy + j_z dz, \tag{2.84}$$

we derive the dual current 3-form

$$\star J = \rho\, dxdydz - j_x dtdydz - j_y dtdzdx - j_z dtdxdy, \tag{2.85}$$

and check that

$$J \star J = (J_\mu J^\mu)\sigma = (-\rho^2 + j_x^2 + j_y^2 + j_z^2) dtdxdydz. \tag{2.86}$$

Observe that

$$d \star J = \left(\frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{j} \right) dtdxdydz = 0, \tag{2.87}$$

expresses the charge conservation law.

Writing out the terms explicitly shows that the source-containing Maxwell equations reduce to $d\star F = \star J$. All four Maxwell equations are therefore very compactly expressed as

$$\boxed{dF = 0, \quad d \star F = \star J.}$$

Observe that current conservation $d\star J = 0$ follows from the second Maxwell equation as a consequence of $d^2 = 0$.

Exercise 2.10: Show that for a p-form ω in d euclidean dimensions we have

$$\star \star \omega = (-1)^{p(d-p)} \omega.$$

Show, further, that for a Minkowski metric an additional minus sign has to be inserted. (For example, $\star \star F = -F$, even though $(-1)^{2(4-2)} = +1$.)

2.4.2 Hamilton's Equations

Hamiltonian dynamics takes place in *phase space*, a manifold with co-ordinates $(q^1, \ldots, q^n, p^1, \ldots, p^n)$. Since momentum is a naturally covariant vector[5], phase space is usually the *co-tangent bundle* T^*M of the configuration manifold M. We are writing the indices on the p's upstairs though, because we are considering them as co-ordinates in T^*M.

We expect that you are familiar with Hamilton's equation in their q, p setting. Here, we shall describe them as they appear in a modern book on Mechanics, such as Abrahams and Marsden's *Foundations of Mechanics*, or V. I. Arnold's *Mathematical Methods of Classical Mechanics*.

[5]To convince yourself of this, remember that in quantum mechanics $\hat{p}_\mu = -i\hbar \frac{\partial}{\partial x^\mu}$, and the gradient of a function is a covector.

Phase space is an example of a *symplectic manifold*, a manifold equiped with a *symplectic form* — a non-degenerate 2-form field

$$\omega = \frac{1}{2}\omega_{ij}dx^i dx^j. \tag{2.88}$$

Recall that the word *closed* means that $d\omega = 0$. *Non-degenerate* means that for any point x the statement that $\omega(X, Y) = 0$ for all vectors $Y \in TM_x$ implies that $X = 0$ at that point (or equivalently that for all x the matrix $\omega_{ij}(x)$ has an inverse $\omega^{ij}(x)$).

Given a *Hamiltonian* function H on our symplectic manifold, we define a velocity vector-field v_H by solving

$$dH = -i_{v_H}\omega = -\omega(v_H, \ \ ) \tag{2.89}$$

for v_H. If the symplectic form is $\omega = dp^1dq^1 + dp^2dq^2 + \cdots + dp^ndq^n$, this is nothing but a fancy form of Hamilton's equations. To see this, we write

$$dH = \frac{\partial H}{\partial q^i}dq^i + \frac{\partial H}{\partial p^i}dp^i \tag{2.90}$$

and use the customary notation $(\dot{q}^i, \dot{p}^i)$ for the velocity-in-phase-space components, so that

$$v_H = \dot{q}^i\frac{\partial}{\partial q^i} + \dot{p}^i\frac{\partial}{\partial p^i}. \tag{2.91}$$

Now we work out

$$\begin{aligned} i_{v_H}\omega &= dp^i dq^i(\dot{q}^j\partial_{q^j} + \dot{p}^j\partial_{p^j}, \ \ ) \\ &= \dot{p}^i dq^i - \dot{q}^i dp^i, \end{aligned} \tag{2.92}$$

so, comparing coefficients of dp^i and dq^i on the two sides of $dH = -i_{v_H}\omega$, we read off

$$\dot{q}^i = \frac{\partial H}{\partial p^i}, \quad \dot{p}^i = -\frac{\partial H}{\partial q^i}. \tag{2.93}$$

Darboux' theorem, which we will not try to prove, says that for any point x we can always find co-ordinates p, q, valid in some neigbourhood of x, such that $\omega = dp^1dq^1 + dp^2dq^2 + \cdots dp^ndq^n$. Given this fact, it is not unreasonable to think that there is little to gained by using the abstract differential-form language. In simple cases this is so, and the traditional methods are fine.

It may be, however, that the neigbourhood of x where the Darboux co-ordinates work is not the entire phase space, and we need to cover the space with overlapping p, q co-ordinate charts. Then, what is a p in one chart will usually be a combination of p's and q's in another. In this case, the traditional form of Hamilton's equations loses its appeal in comparison to the co-ordinate-free $dH = -i_{v_H}\omega$.

Given two functions H_1, H_2 we can define their *Poisson bracket* $\{H_1, H_2\}$. Its importance lies in Dirac's observation that the passage from classical mechanics to quantum mechanics is accomplished by replacing the Poisson bracket of two quantities, A and B, with the commutator of the corresponding operators $\hat{A}$, and $\hat{B}$:

$$i[\hat{A}, \hat{B}] \quad \longleftrightarrow \quad \hbar\{A, B\} + O\left(\hbar^2\right). \tag{2.94}$$

We define the Poisson bracket by[6]

$$\{H_1, H_2\} \stackrel{\text{def}}{=} \left.\frac{dH_2}{dt}\right|_{H_1} = v_{H_1} H_2. \tag{2.95}$$

Now, $v_{H_1} H_2 = dH_2(v_{H_1})$, and Hamilton's equations say that $dH_2(v_{H_1}) = \omega(v_{H_1}, v_{H_2})$. Thus,

$$\{H_1, H_2\} = \omega(v_{H_1}, v_{H_2}). \tag{2.96}$$

The skew symmetry of $\omega(v_{H_1}, v_{H_2})$ shows that despite the asymmetrical appearance of the definition we have skew symmetry: $\{H_1, H_2\} = -\{H_2, H_1\}$.

Moreover, since

$$v_{H_1}(H_2 H_3) = (v_{H_1} H_2) H_3 + H_2(v_{H_1} H_3), \tag{2.97}$$

the Poisson bracket is a derivation:

$$\{H_1, H_2 H_3\} = \{H_1, H_2\} H_3 + H_2\{H_1, H_3\}. \tag{2.98}$$

Neither the skew symmetry nor the derivation property require the condition that $d\omega = 0$. What does need ω to be closed is the *Jacobi identity*:

$$\{\{H_1, H_2\}, H_3\} + \{\{H_2, H_3\}, H_1\} + \{\{H_3, H_1\}, H_2\} = 0. \tag{2.99}$$

[6]Our definition differs in sign from the traditional one, but has the advantage of minimizing the number of minus signs in subsequent equations.

We establish Jacobi by using Cartan's formula in the form

$$\begin{aligned} d\omega(v_{H_1}, v_{H_2}, v_{H_3}) &= v_{H_1}\omega(v_{H_2}, v_{H_3}) + v_{H_2}\omega(v_{H_3}, v_{H_1}) + v_{H_3}\omega(v_{H_1}, v_{H_2}) \\ &\quad -\omega([v_{H_1}, v_{H_2}], v_{H_3}) - \omega([v_{H_2}, v_{H_3}], v_{H_1}) - \omega([v_{H_3}, v_{H_1}], v_{H_2}). \end{aligned} \tag{2.100}$$

It is relatively straight-forward to interpret each term in the first of these lines as Poisson brackets. For example,

$$v_{H_1}\omega(v_{H_2}, v_{H_3}) = v_{H_1}\{H_2, H_3\} = \{H_1, \{H_2, H_3\}\}. \tag{2.101}$$

Relating the terms in the second line to Poisson brackets requires a little more effort. We proceed as follows:

$$\begin{aligned} \omega([v_{H_1}, v_{H_2}], v_{H_3}) &= -\omega(v_{H_3}, [v_{H_1}, v_{H_2}]) \\ &= dH_3([v_{H_1}, v_{H_2}]) \\ &= [v_{H_1}, v_{H_2}]H_3 \\ &= v_{H_1}(v_{H_2}H_3) - v_{H_2}(v_{H_1}H_3) \\ &= \{H_1, \{H_2, H_3\}\} - \{H_2, \{H_1, H_3\}\} \\ &= \{H_1, \{H_2, H_3\}\} + \{H_2, \{H_3, H_1\}\}. \end{aligned} \tag{2.102}$$

Adding everything togther now shows that

$$\begin{aligned} 0 &= d\omega(v_{H_1}, v_{H_2}, v_{H_3}) \\ &= -\{\{H_1, H_2\}, H_3\} - \{\{H_2, H_3\}, H_1\} - \{\{H_3, H_1\}, H_2\}. \end{aligned} \tag{2.103}$$

If we rearrange the Jacobi identity as

$$\{H_1, \{H_2, H_3\}\} - \{H_2, \{H_1, H_3\}\} = \{\{H_1, H_2\}, H_3\}, \tag{2.104}$$

we see that it is equivalent to

$$[v_{H_1}, v_{H_2}] = v_{\{H_1, H_2\}}.$$

The algebra of Poisson brackets is therefore *homomorphic* to the algebra of the Lie brackets. The correspondence is not an *isomorphism*, however: the assignment $H \mapsto v_H$ fails to be one-to-one because constant functions map to the zero vector field.

Exercise 2.11: Use the infinitesimal homotopy relation, to show that $\mathcal{L}_{v_H}\omega = 0$, where v_H is the vector field corresponding to H. Suppose now that the phase space is $2n$ dimensional. Show that in local Darboux co-ordinates the $2n$-form $\omega^n/n!$ is, up to a sign, the phase-space volume element $d^n p\, d^n q$. Show that $\mathcal{L}_{v_H}\omega^n/n! = 0$ and that this result is *Liouville's theorem* on the conservation of phase-space volume.

The classical mechanics of spin

It is sometimes said in books on quantum mechanics that the spin of an electron, or other elementary particle, is a purely quantum concept and cannot be described by classical mechanics. This statement is false, but spin *is* the simplest system in which traditional physicist's methods become ugly and it helps to use the modern symplectic language. A "spin" **S** can be regarded as a fixed length vector that can point in any direction in $\mathbb{R}^3$. We will take it to be of unit length so that its components are

$$\begin{aligned} S_x &= \sin\theta\cos\phi, \\ S_y &= \sin\theta\sin\phi, \\ S_z &= \cos\theta, \end{aligned} \tag{2.105}$$

where θ and ϕ are polar co-ordinates on the two-sphere S^2.

The surface of the sphere turns out to be both the configuration space and the phase space. In particular the phase space for a spin is *not* the cotangent bundle of the configuration space. This has to be so: we learned from Niels Bohr that a $2n$-dimensional phase space contains roughly one quantum state for every $\hbar^n$ of phase-space volume. A cotangent bundle always has infinite volume, so its corresponding Hilbert space is necessarily infinite dimensional. A quantum spin, however, has a *finite-dimensional* Hilbert space so its classical phase space must have a finite total volume. This finite-volume phase space seems un-natural in the traditional view of mechanics, but it fits comfortably into the modern symplectic picture.

We want to treat all points on the sphere alike, and so it is natural to take the symplectic 2-form to be proportional to the element of area. Suppose that $\omega = \sin\theta\, d\theta d\phi$. We could write $\omega = d\cos\theta\, d\phi$ and regard ϕ as "q" and $\cos\theta$ as "p' (Darboux' theorem in action!), but this identification is singular at the north and south poles of the sphere, and, besides, it obscures the spherical symmetry of the problem, which is manifest when we think of ω as d(area).

Let us take our hamiltonian to be $H = BS_x$, corresponding to an applied magnetic field in the x direction, and see what Hamilton's equations give for the motion. First we take the exterior derivative

$$d(BS_x) = B(\cos\theta\cos\phi d\theta - \sin\theta\sin\phi d\phi). \tag{2.106}$$

This is to be set equal to

$$-\omega(v_{BS_x}, \ \ ) = v^\theta(-\sin\theta)d\phi + v^\phi\sin\theta d\theta. \tag{2.107}$$

Comparing coefficients of $d\theta$ and $d\phi$, we get

$$v_{(BS_x)} = v^\theta \partial_\theta + v^\phi \partial_\phi = B(\sin\phi\partial_\theta + \cos\phi\cot\theta\partial_\phi), \tag{2.108}$$

i.e. B times the velocity vector for a rotation about the x axis. This velocity field therefore describes a steady Larmor precession of the spin about the applied field. This is exactly the motion predicted by quantum mechanics. Similarly, setting $B = 1$, we find

$$\begin{aligned} v_{S_y} &= -\cos\phi\partial_\theta + \sin\phi\cot\theta\partial_\phi, \\ v_{S_z} &= -\partial_\phi. \end{aligned} \tag{2.109}$$

From these velocity fields we can compute the Poisson brackets:

$$\begin{aligned} \{S_x, S_y\} &= \omega(v_{S_x}, v_{S_y}) \\ &= \sin\theta d\theta d\phi(\sin\phi\partial_\theta + \cos\phi\cot\theta\partial_\phi, -\cos\phi\partial_\theta + \sin\phi\cot\theta\partial_\phi) \\ &= \sin\theta(\sin^2\phi\cot\theta + \cos^2\phi\cot\theta) \\ &= \cos\theta = S_z. \end{aligned}$$

Repeating the exercise leads to

$$\begin{aligned} \{S_x, S_y\} &= S_z, \\ \{S_y, S_z\} &= S_x, \\ \{S_z, S_x\} &= S_y. \end{aligned} \tag{2.110}$$

These Poisson brackets for our classical "spin" are to be compared with the commutation relations $[\hat{S}_x, \hat{S}_y] = i\hbar\hat{S}_z$ *etc.* for the quantum spin operators $\hat{S}_i$.

2.5 Covariant Derivatives

Covariant derivatives are a general class of derivatives that act on sections of a vector or tensor bundle over a manifold. We will begin by considering derivatives on the tangent bundle, and in the exercises indicate how the idea generalizes to other bundles.

2.5.1 Connections

The Lie and exterior derivatives require no structure beyond that which comes for free with our manifold. Another type of derivative that can act on tangent-space vectors and tensors is the *covariant derivative* $\nabla_X \equiv X^\mu \nabla_\mu$. This requires an additional mathematical object called an *affine connection*.

The covariant derivative is defined by:

i) Its action on scalar functions as

$$\nabla_X f = Xf. \tag{2.111}$$

ii) Its action a basis set of tangent-vector fields $\mathbf{e}_a(x) = e_a^\mu(x)\partial_\mu$ (a local frame, or *vielbein*[7]) by introducing a set of functions $\omega^i{}_{jk}(x)$ and setting

$$\nabla_{\mathbf{e}_k}\mathbf{e}_j = \mathbf{e}_i \omega^i{}_{jk}. \tag{2.112}$$

ii) Extending this definition to any other type of tensor by requiring ∇_X to be a derivation.

iii) Requiring that the result of applying ∇_X to a tensor is a tensor of the same type.

The set of functions $\omega^i{}_{jk}(x)$ is the *connection*. In any local co-ordinate chart we can choose them at will, and different choices define different covariant derivatives. (There may be global compatibility constraints, however, which appear when we assemble the charts into an atlas.)

Warning: Despite having the appearance of one, $\omega^i{}_{jk}$ is **not** a tensor. It transforms inhomogeneously under a change of frame or co-ordinates — see equation (2.131).

We can, of course, take as our basis vectors the co-ordinate vectors $\mathbf{e}_\mu \equiv \partial_\mu$. When we do this it is traditional to use the symbol Γ for the co-ordinate frame connection instead of ω. Thus,

$$\nabla_\mu \mathbf{e}_\nu \equiv \nabla_{\mathbf{e}_\mu}\mathbf{e}_\nu = \mathbf{e}_\lambda \Gamma^\lambda{}_{\nu\mu}. \tag{2.113}$$

The numbers $\Gamma^\lambda{}_{\nu\mu}$ are often called *Christoffel symbols*.

As an example consider the covariant derivative of a vector $f^\nu \mathbf{e}_\nu$. Using the derivation property we have

$$\begin{aligned} \nabla_\mu(f^\nu \mathbf{e}_\nu) &= (\partial_\mu f^\nu)\mathbf{e}_\nu + f^\nu \nabla_\mu \mathbf{e}_\nu \\ &= (\partial_\mu f^\nu)\mathbf{e}_\nu + f^\nu \mathbf{e}_\lambda \Gamma^\lambda{}_{\nu\mu} \\ &= \mathbf{e}_\nu \left\{ \partial_\mu f^\nu + f^\lambda \Gamma^\nu{}_{\lambda\mu} \right\}. \end{aligned} \tag{2.114}$$

[7] In practice *viel*, "many", is replaced by the appropriate German numeral: *ein-, zwei-, drei-, vier-, fünf-, ...*, indicating the dimension. The word *bein* means "leg."

In the first line we have used the defining property that $\nabla_{\mathbf{e}_\mu}$ acts on the functions f^ν as ∂_μ, and in the last line we interchanged the dummy indices ν and λ. We often abuse the notation by writing only the components, and set

$$\nabla_\mu f^\nu = \partial_\mu f^\nu + f^\lambda \Gamma^\nu{}_{\lambda\mu}. \tag{2.115}$$

Similarly, acting on the components of a mixed tensor, we would write

$$\nabla_\mu A^\alpha{}_{\beta\gamma} = \partial_\mu A^\alpha{}_{\beta\gamma} + \Gamma^\alpha{}_{\lambda_\mu} A^\lambda{}_{\beta\gamma} - \Gamma^\lambda{}_{\beta\mu} A^\alpha{}_{\lambda\gamma} - \Gamma^\lambda{}_{\gamma\mu} A^\alpha{}_{\beta\lambda}. \tag{2.116}$$

When we use this notation, we are no longer regarding the tensor components as "functions."

Observe that the plus and minus signs in (2.116) are required so that, for example, the covariant derivative of the scalar function $f_\alpha g^\alpha$ is

$$\begin{aligned} \nabla_\mu \left(f_\alpha g^\alpha\right) &= \partial_\mu \left(f_\alpha g^\alpha\right) \\ &= \left(\partial_\mu f_\alpha\right) g^\alpha + f_\alpha \left(\partial_\mu g^\alpha\right) \\ &= \left(\partial_\mu f_\alpha - f_\lambda \Gamma^\lambda{}_{\alpha\mu}\right) g^\alpha + f_\alpha \left(\partial_\mu g^\alpha + g^\lambda \Gamma^\alpha{}_{\lambda\mu}\right) \\ &= \left(\nabla_\mu f_\alpha\right) g^\alpha + f_\alpha \left(\nabla_\mu g^\alpha\right), \end{aligned} \tag{2.117}$$

and so satisfies the derivation property.

Parallel transport

We have defined the covariant derivative *via* its formal calculus properties. It has, however, a geometrical interpretation. As with the Lie derivative, in order to compute the derivative along X of the vector field Y, we have to somehow carry the vector $Y(x)$ from the tangent space TM_x to the tangent space $TM_{x+\epsilon X}$, where we can subtract it from $Y(x+\epsilon X)$. The Lie derivative carries Y along with the X flow. The covariant derivative implicitly carries Y by "parallel transport". If $\gamma : s \mapsto x^\mu(s)$ is a parameterized curve with tangent vector $X^\mu \partial_\mu$, where

$$X^\mu = \frac{dx^\mu}{ds}, \tag{2.118}$$

then we say that the vector field $Y(x^\mu(s))$ is *parallel transported* along the curve γ if

$$\nabla_X Y = 0, \tag{2.119}$$

at each point $x^\mu(s)$. Thus, a vector that in the vielbein frame $\mathbf{e}_i$ at x has components Y^i will, after being parallel transported to $x + \epsilon X$, end up components

$$Y^i - \epsilon \omega^i{}_{jk} Y^j X^k. \tag{2.120}$$

In a co-ordinate frame, after parallel transport through an infinitesimal displacement δx^μ, the vector $Y^\nu \partial_\nu$ will have components

$$Y^\nu \to Y^\nu - \Gamma^\nu{}_{\lambda\mu} Y^\lambda \delta x^\mu, \tag{2.121}$$

and so

$$\begin{aligned} \delta x^\mu \nabla_\mu Y^\nu &= Y^\nu(x^\mu + \delta x^\mu) - \{Y^\nu(x) - \Gamma^\nu{}_{\lambda\mu} Y^\lambda \delta x^\mu\} \\ &= \delta x^\mu \{\partial_\mu Y^\nu + \Gamma^\nu{}_{\lambda\mu} Y^\lambda\}. \end{aligned} \tag{2.122}$$

Curvature and Torsion

As we said earlier, the connection $\omega^i{}_{jk}(x)$ is not itself a tensor. Two important quantities which *are* tensors, are associated with ∇_X:

i) The *torsion*

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \tag{2.123}$$

The quantity $T(X,Y)$ is a vector depending linearly on X, Y, so T at the point x is a map $TM_x \times TM_x \to TM_x$, and so a tensor of type (1,2). In a co-ordinate frame it has components

$$T^\lambda{}_{\mu\nu} = \Gamma^\lambda{}_{\mu\nu} - \Gamma^\lambda{}_{\nu\mu}. \tag{2.124}$$

ii) The *Riemann curvature tensor*

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_Z Z - \nabla_{[X,Y]} Z. \tag{2.125}$$

The quantity $R(X,Y)Z$ is also a vector, so $R(X,Y)$ is a linear map $TM_x \to TM_x$, and thus R itself is a tensor of type (1,3). Written out in a co-ordinate frame, we have

$$R^\alpha{}_{\beta\mu\nu} = \partial_\mu \Gamma^\alpha{}_{\beta\nu} - \partial_\nu \Gamma^\alpha{}_{\beta\mu} + \Gamma^\alpha{}_{\lambda\mu} \Gamma^\lambda{}_{\beta\nu} - \Gamma^\alpha{}_{\lambda\nu} \Gamma^\lambda{}_{\beta\mu}. \tag{2.126}$$

If our manifold comes equipped with a metric tensor $g_{\mu\nu}$ (and is thus a *Riemann manifold*), and if we require both that $T = 0$ and $\nabla_\mu g_{\alpha\beta} = 0$,

then the connection is uniquely determined, and is called the *Riemann*, or *Levi-Civita*, connection. In a co-ordinate frame it is given by

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2} g^{\alpha\lambda} \left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu} \right). \tag{2.127}$$

This is the connection that appears in General Relativity.

The curvature tensor measures the degree of path dependence in parallel transport: if $Y^\nu(x)$ is parallel transported along a path $\gamma : s \mapsto x^\mu(s)$ from a to b, and if we deform γ so that $x^\mu(s) \to x^\mu(s) + \delta x^\mu(s)$ while keeping the endpoints a, b fixed, then

$$\delta Y^\alpha(b) = -\int_a^b R^{\alpha}{}_{\beta\mu\nu}(x) Y^\beta(x) \delta x^\mu \, dx^\nu. \tag{2.128}$$

If $R^{\alpha}{}_{\beta\mu\nu} \equiv 0$ then the effect of parallel transport from a to b will be independent of the route taken.

The geometric interpretation of $T_{\mu\nu}$ is less transparent. On a two-dimensional surface a connection is torsion free when the tangent space "rolls without slipping" along the curve γ.

Exercise 2.12: *Metric compatibility*. Show that the Riemann connection

$$\Gamma^{\alpha}{}_{\mu\nu} = \frac{1}{2} g^{\alpha\lambda} \left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu} \right).$$

follows from the torsion-free condition $\Gamma^{\alpha}{}_{\mu\nu} = \Gamma^{\alpha}{}_{\nu\mu}$ together with the *metric compatibility condition*

$$\nabla_\mu g_{\alpha\beta} \equiv \partial_\mu \, g_{\alpha\beta} - \Gamma^{\nu}{}_{\alpha\mu} \, g_{\nu\beta} - \Gamma^{\nu}{}_{\alpha\mu} \, g_{\alpha\nu} = 0.$$

Show that "metric compatibility" means that that the operation of raising or lowering indices commutes with covariant derivation.

Exercise 2.13: *Geodesic equation.* Let $\gamma : s \mapsto x^\mu(s)$ be a parametrized path from a to b. Show that the Euler-Lagrange equation that follows from minimizing the distance functional

$$S(\gamma) = \int_a^b \sqrt{g_{\mu\nu} \dot{x}^\mu \dot{x}^\nu} \, ds,$$

where the dots denote differentiation with respect to the parameter s, is

$$\frac{d^2 x^\mu}{ds^2} + \Gamma^{\mu}{}_{\alpha\beta} \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = 0.$$

Here $\Gamma^{\mu}{}_{\alpha\beta}$ is the Riemann connection (2.127).

Exercise 2.14: Show that if A^μ is a vector field then, for the Riemann connection,

$$\nabla_\mu A^\mu = \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g} A^\mu}{\partial x^\mu}.$$

In other words, show that

$$\Gamma^\alpha{}_{\alpha\mu} = \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g}}{\partial x^\mu}.$$

Deduce that the Laplacian acting on a scalar field ϕ can be defined by setting either

$$\nabla^2 \phi = g_{\mu\nu} \nabla_\mu \nabla_\nu \phi,$$

or

$$\nabla^2 \phi = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x^\mu} \left(\sqrt{g} g^{\mu\nu} \frac{\partial \phi}{\partial x^\nu} \right),$$

the two definitions being equivalent.

2.5.2 Cartan's Form Viewpoint

Let $\mathbf{e}^{*j}(x) = e^{*j}{}_\mu(x) dx^\mu$ be the basis of one-forms dual to the vielbein frame $\mathbf{e}_i(x) = e_i^\mu(x) \partial_\mu$. Since

$$\delta^i_j = \mathbf{e}^{*i}(\mathbf{e}_j) = e^{*j}{}_\mu e_i^\mu, \tag{2.129}$$

the matrices $e^{*j}{}_\mu$ and e_i^μ are inverses of one-another. We can use them to change from roman vielbein indices to greek co-ordinate frame indices. For example:

$$g_{ij} = g(\mathbf{e}_i, \mathbf{e}_j) = e_i^\mu g_{\mu\nu} e_j^\nu, \tag{2.130}$$

and

$$\omega^i{}_{jk} = e^{*i}{}_\nu (\partial_\mu e_j^\nu) e_k^\mu + e^{*i}{}_\lambda e_j^\nu e_k^\mu \Gamma^\lambda{}_{\nu\mu}. \tag{2.131}$$

Cartan regards the connection as being a matrix $\mathbf{\Omega}$ of one-forms with entries $\omega^i{}_j = \omega^i{}_{j\mu} dx^\mu$. In this language equation (2.112) becomes

$$\nabla_X \mathbf{e}_j = \mathbf{e}_i \omega^i{}_j(X). \tag{2.132}$$

Cartan's viewpoint separates off the index μ, which refers to the direction $\delta x^\mu \propto X^\mu$ in which we are differentiating, from the matrix indices i and j that act on the components of the vector or tensor being differentiated. This separation becomes very natural when the vector space spanned by the

$\mathbf{e}_i(x)$ is no longer the tangent space, but some other "internal" vector space attached to the point x. Such internal spaces are common in physics, an important example being the "colour space" of gauge field theories. Physicists, following Hermann Weyl, call a connection on an internal space a "gauge potential." To mathematicians it is simply a connection on the vector bundle that has the internal spaces as its fibres.

Cartan also regards the torsion $\mathbf{T}$ and curvature $\mathbf{R}$ as forms; in this case vector- and matrix-valued two-forms, respectively, with entries

$$T^i = \frac{1}{2} T^i{}_{\mu\nu} dx^\mu dx^\nu, \tag{2.133}$$

$$R^i{}_k = \frac{1}{2} R^i{}_{k\mu\nu} dx^\mu dx^\nu. \tag{2.134}$$

In his form language the equations defining the torsion and curvature become *Cartan's structure equations*:

$$d\mathbf{e}^{*i} + \omega^i{}_j \wedge \mathbf{e}^{*j} = T^i, \tag{2.135}$$

and

$$d\omega^i{}_k + \omega^i{}_j \wedge \omega^j{}_k = R^i{}_k. \tag{2.136}$$

The last equation can be written more compactly as

$$d\mathbf{\Omega} + \mathbf{\Omega} \wedge \mathbf{\Omega} = \mathbf{R}. \tag{2.137}$$

From this, by taking the exterior derivative, we obtain the *Bianchi identity*

$$d\mathbf{R} - \mathbf{R} \wedge \mathbf{\Omega} + \mathbf{\Omega} \wedge \mathbf{R} = 0. \tag{2.138}$$

On a Riemann manifold, we can take the vielbein frame $\mathbf{e}_i$ to be orthonormal. In this case the roman-index metric $g_{ij} = g(\mathbf{e}_i, \mathbf{e}_j)$ becomes δ_{ij}. There is then no distinction between covariant and contravariant roman indices, and the connection and curvature forms, $\mathbf{\Omega}$, $\mathbf{R}$, being infinitesimal rotations, become skew symmetric matrices:

$$\omega_{ij} = -\omega_{ji}, \quad R_{ij} = -R_{ji}. \tag{2.139}$$

2.6 Further Exercises and Problems

Exercise 2.15: Consider the vector fields $X = y\partial_x$, $Y = \partial_y$ in $\mathbb{R}^2$. Find the flows associated with these fields, and use them to verify the statements made in section 2.2.1 about the geometric interpretation of the Lie bracket.

Exercise 2.16: Show that the pair of vector fields $L_z = x\partial_y - y\partial_x$ and $L_y = z\partial_x - x\partial_z$ in $\mathbb{R}^3$ is in involution wherever they are both non-zero. Show further that the general solution of the system of partial differential equations

$$\begin{aligned}(x\partial_y - y\partial_x)f &= 0,\\ (x\partial_z - z\partial_x)f &= 0,\end{aligned}$$

in $\mathbb{R}^3$ is $f(x, y, z) = F(x^2 + y^2 + z^2)$, where F is an arbitrary function.

Exercise 2.17: In the rolling conditions (2.26) we are using the "Y" convention for Euler angles. In this convention θ and ϕ are the usual spherical polar coordinate angles with respect to the space-fixed xyz axes. They specify the direction of the body-fixed Z axis about which we make the final ψ rotation.

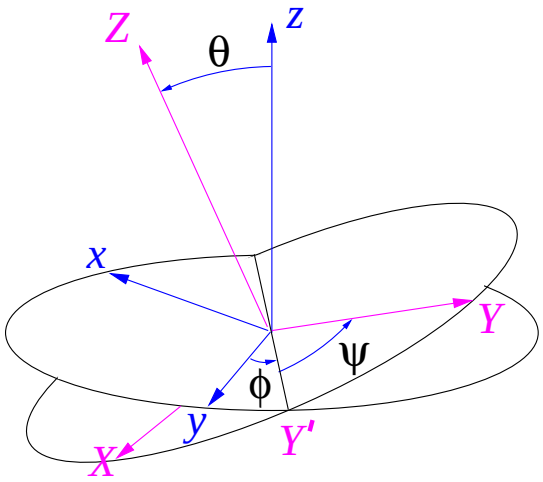

Figure 2.7: *Euler angles: we first rotate the ball through an angle ϕ about the z axis, thus taking $y \to Y'$, then through θ about Y', and finally through ψ about Z, so taking $Y' \to Y$.*

a) Show that (2.26) are indeed the no-slip rolling conditions

$$\begin{aligned}\dot{x} &= \omega_y,\\ \dot{y} &= -\omega_x,\\ 0 &= \omega_z,\end{aligned}$$

where $(\omega_x, \omega_y, \omega_z)$ are the components of the ball's angular velocity in the xyz space-fixed frame.

b) Solve the three constraints in (2.26) so as to obtain the vector fields (2.27).

c) Show that

$$[\mathbf{roll_x}, \mathbf{roll_y}] = -\mathbf{spin_z},$$

where $\mathbf{spin_z} \equiv \partial_\phi$, corresponds to a rotation about a vertical axis through the point of contact. This is a new motion, being forbidden by the $\omega_z = 0$ condition.

d) Show that

$$\begin{aligned} [\mathbf{spin_z}, \mathbf{roll_x}] &= \mathbf{spin_x}, \\ [\mathbf{spin_z}, \mathbf{roll_y}] &= \mathbf{spin_y}, \end{aligned}$$

where the new vector fields

$$\begin{aligned} \mathbf{spin_x} &\equiv -(\mathbf{roll_y} - \partial_y), \\ \mathbf{spin_y} &\equiv \quad (\mathbf{roll_x} - \partial_x), \end{aligned}$$

correspond to rotations of the ball about the space-fixed x and y axes through its centre, and with the centre of mass held fixed.

We have generated five independent vector fields from the original two. Therefore, by sufficient rolling to-and-fro, we can position the ball anywhere on the table, and in any orientation.

Exercise 2.18: The semi-classical dynamics of charge $-e$ electrons in a magnetic solid are governed by the equations[8]

$$\begin{aligned} \dot{\mathbf{r}} &= \frac{\partial \epsilon(\mathbf{k})}{\partial \mathbf{k}} - \dot{\mathbf{k}} \times \mathbf{\Omega}, \\ \dot{\mathbf{k}} &= -\frac{\partial V}{\partial \mathbf{r}} - e\dot{\mathbf{r}} \times \mathbf{B}. \end{aligned}$$

Here $\mathbf{k}$ is the Bloch momentum of the electron, $\mathbf{r}$ is its position, $\epsilon(\mathbf{k})$ its band energy (in the extended-zone scheme), and $\mathbf{B}(\mathbf{r})$ is the external magnetic field. The components Ω_i of the *Berry curvature* $\mathbf{\Omega}(\mathbf{k})$ are given in terms of the periodic part $|u(\mathbf{k})\rangle$ of the Bloch wavefunctions of the band by

$$\Omega_i = i\epsilon_{ijk}\frac{1}{2}\left(\left\langle \frac{\partial u}{\partial k_j}\middle|\frac{\partial u}{\partial k_k}\right\rangle - \left\langle \frac{\partial u}{\partial k_k}\middle|\frac{\partial u}{\partial k_j}\right\rangle\right).$$

[8] M. C. Chang, Q. Niu, *Phys. Rev. Lett.* **75** (1995) 1348.

The only property of $\mathbf{\Omega}(\mathbf{k})$ needed for the present problem, however, is that $\mathrm{div}_{\mathbf{k}}\mathbf{\Omega} = 0$.

a) Show that these equations are Hamiltonian, with

$$H(\mathbf{r}, \mathbf{k}) = \epsilon(\mathbf{k}) + V(\mathbf{r})$$

and with

$$\omega = dk_i dx_i - \frac{e}{2}\epsilon_{ijk}B_i(\mathbf{r})dx_j dx_k + \frac{1}{2}\epsilon_{ijk}\Omega_i(\mathbf{k})dk_j dk_k.$$

as the symplectic form.[9]

b) Confirm that the ω defined in part b) is closed, and that the Poisson brackets are given by

$$\begin{aligned}
\{x_i, x_j\} &= -\frac{\epsilon_{ijk}\Omega_k}{(1 + e\mathbf{B}\cdot\mathbf{\Omega})}, \\
\{x_i, k_j\} &= -\frac{\delta_{ij} + \Omega_i B_j}{(1 + e\mathbf{B}\cdot\mathbf{\Omega})}, \\
\{k_i, k_j\} &= \frac{\epsilon_{ijk}B_k}{(1 + e\mathbf{B}\cdot\mathbf{\Omega})}.
\end{aligned}$$

c) Show that the conserved phase-space volume $\omega^3/3!$ is equal to

$$(1 + e\mathbf{B}\cdot\mathbf{\Omega})d^3k d^3x,$$

instead of the naïvely expected $d^3k d^3x$.

The following pair of exercises show that Cartan's expression for the curvature tensor remains valid for covariant differentiation in "internal" spaces. There is, however, no natural concept analogous to the torsion tensor for internal spaces.

Exercise 2.19: *Non-abelian gauge fields as matrix-valued forms.* In a non-abelian Yang-Mills gauge theory, such as QCD, the vector potential

$$A = A_\mu dx^\mu$$

is matrix-valued, meaning that the components A_μ are matrices which do not necessarily commute with each other. (These matrices are elements of the Lie

[9] C. Duval, Z. Horváth, P. A. Horváthy, L. Martina, P. C. Stichel, *Modern Physics Letters B* **20** (2006) 373.

algebra of the gauge group, but we won't need this fact here.) The matrix-valued curvature, or field-strength, 2-form F is defined by

$$F = dA + A^2 = \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu.$$

Here a combined matrix and wedge product is to be understood:

$$(A^2)^a{}_b \equiv A^a{}_c \wedge A^c{}_b = A^a{}_{c\mu} A^c{}_{b\nu}\, dx^\mu dx^\nu.$$

i) Show that $A^2 = \frac{1}{2}[A_\mu, A_\nu] dx^\mu dx^\nu$, and hence show that

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu].$$

ii) Define the *gauge-covariant derivatives*

$$\nabla_\mu = \partial_\mu + A_\mu,$$

and show that the commutator $[\nabla_\mu, \nabla_\nu]$ of two of these is equal to $F_{\mu\nu}$. Show further that if X, Y are two vector fields with Lie bracket $[X, Y]$ and $\nabla_X \equiv X^\mu \nabla_\mu$, then

$$F(X, Y) = [\nabla_X, \nabla_Y] - \nabla_{[X,Y]}.$$

iii) Show that F obeys the Bianchi identity

$$dF - FA + AF = 0.$$

Again wedge and matrix products are to be understood. This equation is the non-abelian version of the source-free Maxwell equation $dF = 0$.

iv) Show that, in any number of dimensions, the Bianchi identity implies that the 4-form $\mathrm{tr}\,(F^2)$ is closed, *i.e.* that $d\,\mathrm{tr}\,(F^2) = 0$. Similarly show that the $2n$-form $\mathrm{tr}\,(F^n)$ is closed. (Here the "tr" means a trace over the roman matrix indices, and not over the greek space-time indices.)

v) Show that,

$$\mathrm{tr}\,(F^2) = d\left\{ \mathrm{tr}\,\left(AdA + \frac{2}{3} A^3 \right) \right\}.$$

The 3-form $\mathrm{tr}\,(AdA + \frac{2}{3}A^3)$ is called a *Chern-Simons* form.

Exercise 2.20: *Gauge transformations.* Here we consider how the matrix-valued vector potential transforms when we make a change of gauge. In other words, we seek the non-abelian version of $A_\mu \to A_\mu + \partial_\mu \phi$.

i) Let g be an invertable matrix, and δg a matrix describing a small change in g. Show that the corresponding change in the inverse matrix is given by $\delta(g^{-1}) = -g^{-1}(\delta g)g^{-1}$.

ii) Show that under the *gauge transformation*

$$A \to A^g \equiv g^{-1}Ag + g^{-1}dg,$$

we have $F \to g^{-1}Fg$. (Hint: The labour is minimized by exploiting the covariant derivative identity in part ii) of the previous exercise).

iii) Deduce that $\mathrm{tr}\,(F^n)$ is *gauge invariant*.

iv) Show that a necessary condition for the matrix-valued gauge field A to be "pure gauge", *i.e.* for there to be a position dependent matrix g such that $A = g^{-1}dg$, is that $F = 0$, where F is the curvature two-form of the previous exercise.

In a gauge theory based on a Lie group G, the matrices g will be elements of the group, or, more generally, they will form a matrix representation of the group.

Chapter 3

Integration on Manifolds

One usually thinks of integration as requiring *measure* – a notion of volume, and hence of size and length, and so a *metric*. A metric however is not required for integrating differential forms. They come pre-equipped with whatever notion of length, area, or volume is required.

3.1 Basic Notions

3.1.1 Line Integrals

Consider, for example, the form df. We want to try to give a meaning to the symbol

$$I_1 = \int_\Gamma df. \tag{3.1}$$

Here Γ is a path in our space starting at some point P_0 and ending at the point P_1. Any reasonable definition of I_1 should end up with the answer we would immediately write down if we saw an expression like I_1 in an elementary calculus class. This answer is

$$I_1 = \int_\Gamma df = f(P_1) - f(P_0). \tag{3.2}$$

No notion of a metric is needed here. There is however a geometric picture of what we have done. We draw in our space the surfaces $\ldots, f(x) = -1, f(x) = 0, f(x) = 1, \ldots$, and perhaps fill in intermediate values if necessary. We then start at P_0 and travel from there to P_1, keeping track of how many of

these surfaces we pass through (with sign -1, if we pass back through them). The integral of df is this number. Figure 3.1 illustrates a case in which $\int_\Gamma df = 5.5 - 1.5 = 4$.

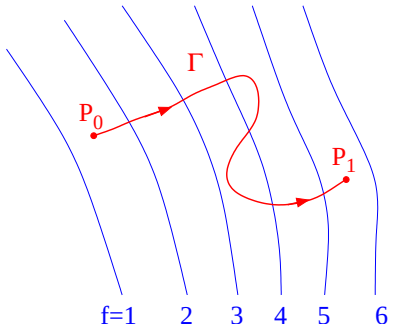

Figure 3.1: *The integral of a one-form*

What we have defined is a *signed integral*. If we parameterise the path as $x(s)$, $0 \le s \le 1$, and with $x(0) = P_0$, $x(1) = P_1$ we have

$$I_1 = \int_0^1 \left(\frac{df}{ds}\right) ds \tag{3.3}$$

where the right hand side is an ordinary one-variable integral. It is important that we did not write $\left|\frac{df}{ds}\right|$ in this integral. The absence of the modulus sign ensures that if we partially retrace our route, so that we pass over some part of Γ three times—twice forward and once back—we obtain the same answer as if we went only forward.

3.1.2 Skew-symmetry and Orientations

What about integrating 2 and 3-forms? Why the skew-symmetry? To answer these questions, think about assigning some sort of "area" in $\mathbb{R}^2$ to the parallelogram defined by the two vectors $\mathbf{x}, \mathbf{y}$. This is going to be some function of the two vectors. Let us call it $\omega(\mathbf{x}, \mathbf{y})$. What properties do we demand of this function? There are at least three:

i) Scaling: If we double the length of one of the vectors, we expect the area to double. Generalizing this, we demand $\omega(\lambda\mathbf{x}, \mu\mathbf{y}) = (\lambda\mu)\omega(\mathbf{x}, \mathbf{y})$. (Note that we are not putting modulus signs on the lengths, so we are allowing negative "areas", and for the sign to change when we reverse the direction of a vector.)

ii) Additivity: The drawing in figure 3.2 shows that we ought to have

$$\omega(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = \omega(\mathbf{x}_1, \mathbf{y}) + \omega(\mathbf{x}_2, \mathbf{y}), \tag{3.4}$$

similarly for the second slots.

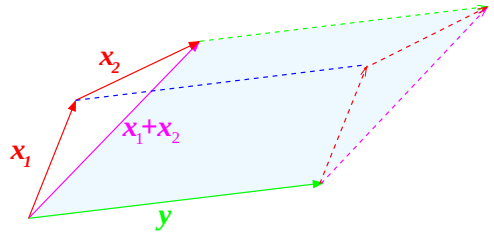

Figure 3.2: *Additivity of* $\omega(\mathbf{x}, \mathbf{y})$.

iii) Degeneration: If the two sides coincide, the area should be zero. Thus $\omega(\mathbf{x}, \mathbf{x}) = 0$.

The first two properties, show that ω should be a multilinear form. The third shows that it must be skew-symmetric!

$$\begin{aligned} 0 = \omega(\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}) &= \omega(\mathbf{x}, \mathbf{x}) + \omega(\mathbf{x}, \mathbf{y}) + \omega(\mathbf{y}, \mathbf{x}) + \omega(\mathbf{y}, \mathbf{y}) \\ &= \omega(\mathbf{x}, \mathbf{y}) + \omega(\mathbf{y}, \mathbf{x}). \end{aligned} \tag{3.5}$$

So

$$\omega(\mathbf{x}, \mathbf{y}) = -\omega(\mathbf{y}, \mathbf{x}). \tag{3.6}$$

These are exactly the properties possessed by a 2-form. Similarly, a 3-form outputs a volume element.

These volume elements are *oriented*. Remember that an orientation of a set of vectors is a choice of order in which to write them. If we interchange two vectors, the orientation changes sign. We do not distinguish orientations related by an even number of interchanges. A p-form assigns a signed ($\pm$) p-dimensional volume element to an orientated set of vectors. If we change the orientation, we change the sign of the volume element.

Orientable and Non-orientable Manifolds

In the classic video game *Asteroids* you could select periodic boundary conditions so that your spaceship would leave the right-hand side of the screen

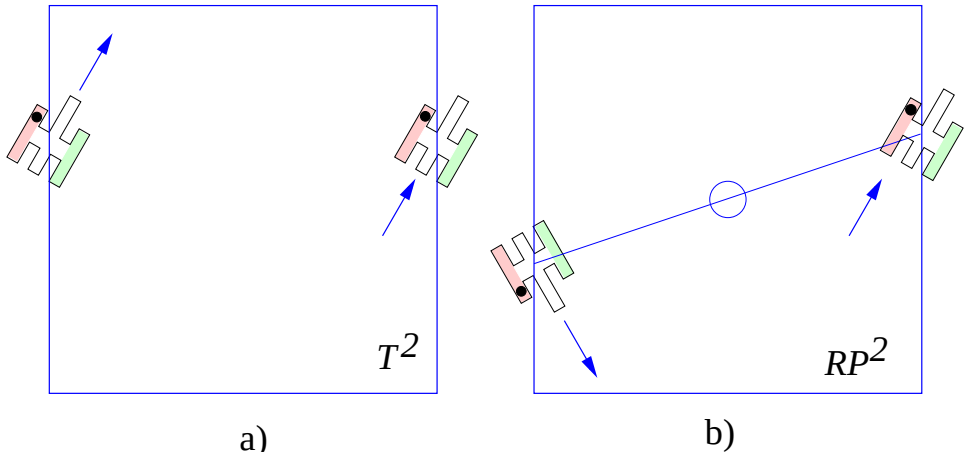

Figure 3.3: *A spaceship leaves one side of the screen and returns on the other with a) torus boundary conditions, b) projective-plane boundary conditions. Observe how, in case b), the spaceship has changed from being left handed to being right-handed.*

and re-appear on the left. The game universe was topologically a torus T^2. Suppose that we modify the game code so that each bit of the spaceship re-appears at the point *diametrically opposite* the point it left. This does not seem like a drastic change until you play a game with a left-hand-drive (US) spaceship. If you send the ship off the screen and watch as it re-appears on the opposite side, you will observe the ship transmogrify into a right-hand-drive (British) craft. If we ourselves made such an excursion, we would end up starving to death because all our left-handed digestive enzymes would have been converted to right-handed ones. The manifold we have constructed is topologically equivalent to the *real projective plane* $\mathbb{R}P^2$. The lack of a global notion of being left or right-handed makes it an example of a *non-orientable* manifold.

A manifold or surface is *orientable* if we can choose a global orientation for the tangent bundle. The simplest way to do this would be to find a smoothly varying set of basis-vector fields, $\mathbf{e}_\mu(x)$, on the surface and define the orientation by chosing an order, $\mathbf{e}_1(x), \mathbf{e}_2(x), \ldots, \mathbf{e}_d(x)$, in which to write them. In general, however, a globally-defined smooth basis will not exist (try to construct one for the two-sphere, S^2!). We will, however, be able to find a continously varying orientated basis $\mathbf{e}_1^{(i)}(x), \mathbf{e}_2^{(i)}(x), \ldots, \mathbf{e}_d^{(i)}(x)$ for each member, labelled by (i), of an atlas of coordinate charts. We should chose

the charts so the intersection of any pair forms a connected set. Assuming that this has been done, the orientation of pair of overlapping charts is said to coincide if the determinant, $\det A$, of the map $\mathbf{e}_\mu^{(i)} = A_\mu^\nu \mathbf{e}_\nu^{(j)}$ relating the bases in the region of overlap, is positive.[1] If bases can be chosen so that all overlap determinants are positive, the manifold is *orientable* and the selected bases define the orientation. If bases cannot be so chosen, the manifold or surface is *non-orientable*.

Exercise 3.1: Consider a *three-dimensional* ball B^3 with diametrically opposite points of its surface identified. What would happen to an aircraft flying through the surface of the ball? Would it change handedness, turn inside out, or simply turn upside down? Is this ball an orientable 3-manifold?

3.2 Integrating p-Forms

A p-form is naturally integrated over an oriented p-dimensional surface or manifold. Rather than start with an abstract definition, We will first explain this pictorially, and then translate the pictures into mathematics.

3.2.1 Counting Boxes

To visualize integrating 2-forms let us try to make sense of

$$\int_\Omega df dg, \tag{3.7}$$

where Ω is an oriented region embedded in three dimensions. The surfaces $f = const.$ and $g = const.$ break the space up into a series of tubes. The oriented surface Ω cuts these tubes in a two-dimensional mesh of (oriented) parallelograms.

[1] The determinant will have the same sign in the entire overlap region. If it did not, continuity and connectedness would force it to be zero somewhere, implying that one of the putative bases was not linearly independent there

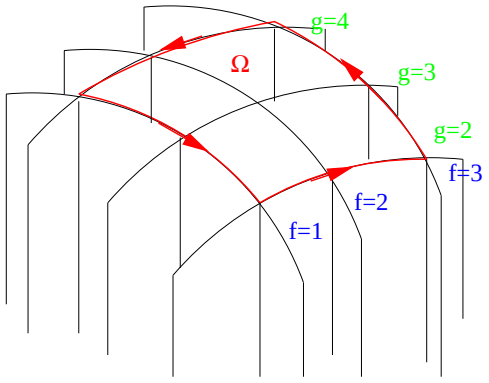

Figure 3.4: *The integration region cuts the tubes into parallelograms.*

We define an integral by counting how many parallelograms (including fractions of a parallelogram) there are, taking the number to be positive if the parallelogram given by the mesh is oriented in the same way as the surface, and negative otherwise. To compute

$$\int_\Omega h df\, dg \tag{3.8}$$

we do the same, but weight each parallelogram, by the value of h at that point. The integral $\int_\Omega f dx dy$, over a region in $\mathbb{R}^2$ thus ends up being the number we would compute in a multivariate calculus class, but the integral $\int_\Omega f dy dx$, would be minus this. Similarly we compute

$$\int_\Xi df\, dg\, dh \tag{3.9}$$

of the 3-form $df\, dg\, dh$ over the oriented volume Ξ, by counting how many boxes defined by the surfaces $f, g, h =$ constant, are included in Ξ.

An equivalent way of thinking of the integral of a p-form uses its definition as a skew-symmetric p-linear function. Accordingly we evaluate

$$I_2 = \int_\Omega \omega, \tag{3.10}$$

where ω is a 2-form, and Ω is an oriented 2-surface, by plugging vectors into ω. We tile the surface Ω with collection of (small) parallelograms, each defined by an oriented pair of basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

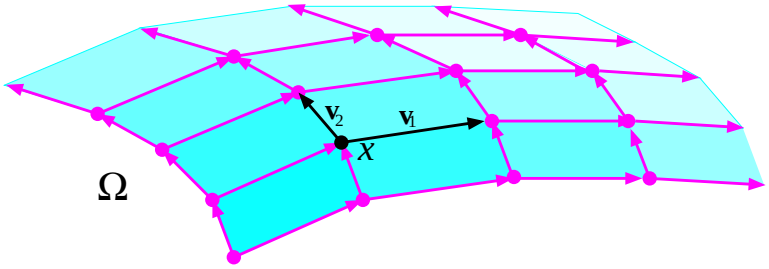

Figure 3.5: *We tile Ω with small oriented parallelograms and compute $\sum_{x\in\Omega}\omega(\mathbf{v}_1(x),\mathbf{v}_2(x))$.*

At each base point x we insert these vectors into the 2-form (in the order specified by their orientation) to get $\omega(\mathbf{v}_1,\mathbf{v}_2)$, and then sum the resulting numbers to get I_2. Similarly, we integrate p-form over an oriented p-dimensional region by decomposing the region into infinitesimal p-dimensional oriented parallelepipeds, inserting their defining vectors into the form, and summing their contributions.

3.2.2 Relation to conventional integrals

The previous section explained how to think pictorially about the integral. Here we interpret the pictures as multi-variable calculus.

We begin by motivating our recipe by considering a change of variables in an integral in $\mathbb{R}^2$. Suppose we set $x_1 = x(y_1,y_2)$, $x_2 = x_2(y_1,y_2)$ in

$$I_4 = \int_\Omega f(x)dx^1dx^2 \tag{3.11}$$

and use

$$\begin{aligned} dx^1 &= \frac{\partial x^1}{\partial y^1}dy^1 + \frac{\partial x^1}{\partial y^2}dy^2, \\ dx^2 &= \frac{\partial x^2}{\partial y^1}dy^1 + \frac{\partial x^2}{\partial y^2}dy^2. \end{aligned} \tag{3.12}$$

Since $dy^1dy^2 = -dy^2dy^1$, we have

$$dx^1dx^2 = \left(\frac{\partial x^1}{\partial y^1}\frac{\partial x^2}{\partial y^2} - \frac{\partial x^2}{\partial y^1}\frac{\partial x^1}{\partial y^2}\right)dy^1dy^2. \tag{3.13}$$

Thus

$$\int_\Omega f(x)dx^1dx^2 = \int_{\Omega'} f(x(y))\frac{\partial(x^1,x^2)}{\partial(y^1,y^2)}dy^1dy^2 \tag{3.14}$$

where $\frac{\partial(x^1,y^1)}{\partial(y^1,y^2)}$ is the Jacobean determinant

$$\frac{\partial(x^1,y^1)}{\partial(y^1,y^2)} \equiv \left(\frac{\partial x^1}{\partial y^1}\frac{\partial x^2}{\partial y^2} - \frac{\partial x^2}{\partial y^1}\frac{\partial x^1}{\partial y^2}\right), \tag{3.15}$$

and Ω' the integration region in the new variables. There is therefore no need to include an explicit Jacobean factor when changing variables in an integral of a p-form over a p-dimensional space—it comes for free with the form.

This observation leads us to the general prescription: To evaluate $\int_\Omega \omega$, the integral of a p-form

$$\omega = \frac{1}{p!}\omega_{\mu_1\mu_2\ldots\mu_p}dx^{\mu_1}\cdots dx^{\mu_p} \tag{3.16}$$

over the region Ω of a p dimensional surface in a $d \geq p$ dimensional space, substitute a paramaterization

$$\begin{aligned} x^1 &= x^1(\xi^1,\xi^2,\ldots,\xi^p), \\ &\vdots \\ x^d &= x^d(\xi^1,\xi^2,\ldots,\xi^p), \end{aligned} \tag{3.17}$$

of the surface into ω. Next, use

$$dx^\mu = \frac{\partial x^\mu}{\partial \xi^i}d\xi^i, \tag{3.18}$$

so that

$$\omega \to \omega(x(\xi))_{i_1i_2\ldots i_p}\frac{\partial x^{i_1}}{\partial \xi^1}\cdots\frac{\partial x^{i_p}}{\partial \xi^p}d\xi^1\cdots d\xi^p, \tag{3.19}$$

which we regard as a p-form on Ω. (Our customary $1/p!$ is absent here because we have chosen a particular order for the $d\xi$'s.) Then

$$\int_\Omega \omega \stackrel{\text{def}}{=} \int_\Omega \omega(x(\xi))_{i_1i_2\ldots i_p}\frac{\partial x^{i_1}}{\partial \xi^1}\cdots\frac{\partial x^{i_p}}{\partial \xi^p}d\xi^1\cdots d\xi^p, \tag{3.20}$$

where the right hand side is an ordinary multiple integral. This recipe is a generalization of the formula (3.3) which reduced the integral of a one-form

to an ordinary single-variable integral. Because the appropriate Jacobean factor appears automatically, the numerical value of the integral does not depend on the choice of parameterization of the surface.
Example: To integrate the 2-form $x\,dydz$ over the surface of a two dimensional sphere of radius R, we parameterize the surface with polar angles as

$$\begin{aligned} x &= R\sin\phi\sin\theta, \\ y &= R\cos\phi\sin\theta, \\ z &= R\cos\theta. \end{aligned} \tag{3.21}$$

Then

$$\begin{aligned} dy &= -R\sin\phi\sin\theta d\phi + R\cos\phi\cos\theta d\theta, \\ dz &= -R\sin\theta d\theta, \end{aligned} \tag{3.22}$$

and so

$$x\,dydz = R^3\sin^2\phi\sin^3\theta\,d\phi d\theta. \tag{3.23}$$

We therefore evaluate

$$\begin{aligned} \int_{\text{sphere}} x\,dydz &= R^3\int_0^{2\pi}\int_0^{\pi}\sin^2\phi\sin^3\theta\,d\phi d\theta \\ &= R^3\int_0^{2\pi}\sin^2\phi\,d\phi\int_0^{\pi}\sin^3\theta\,d\theta \\ &= R^3\pi\int_{-1}^{1}(1-\cos^2\theta)\,d\cos\theta \\ &= \frac{4}{3}\pi R^3. \end{aligned} \tag{3.24}$$

The volume form

Although we do not need any notion of length to integrate a differential form, a p-dimensional surface embedded or immersed in $\mathbb{R}^d$ does inherit a distance scale from the $\mathbb{R}^d$ Euclidean metric, and this is used to define the area or volume of the surface. When the Cartesian co-ordinates $x^1,\ldots,x^d$ of a point in the surface are given as $x^a(\xi^1,\ldots,\xi^p)$, where the $\xi^1,\ldots,\xi^p$, are co-ordinates on the surface, then the inherited, or *induced*, metric is

$$\text{“}ds^2\text{”} \equiv g(\ ,\ ) \equiv g_{\mu\nu}\,d\xi^\mu\otimes d\xi^\nu \tag{3.25}$$

where

$$g_{\mu\nu} = \sum_{a=1}^{d} \frac{\partial x^a}{\partial \xi^\mu} \frac{\partial x^a}{\partial \xi^\nu}. \tag{3.26}$$

The *volume form* associated with the induced metric is

$$d(\text{Volume}) = \sqrt{g}\, d\xi^1 \cdots d\xi^p, \tag{3.27}$$

where $g = \det(g_{\mu\nu})$. The integral of this p-form over the surface gives the area, or p-dimensional volume, of the surface.

If we change the parameterization of the surface from ξ^μ to ζ^μ, neither the $d\xi^1 \cdots d\xi^p$ nor the $\sqrt{g}$ are separately invariant, but the Jacobean arising from the change of the p-form, $d\xi^1 \cdots d\xi^p \to d\zeta^1 \cdots d\zeta^p$ cancels against the factor coming from the transformation law of the metric tensor $g_{\mu\nu} \to g'_{\mu\nu}$, leading to

$$\sqrt{g}\, d\xi^1 \cdots d\xi^p = \sqrt{g'} d\zeta^1 \cdots d\zeta^p. \tag{3.28}$$

The volume of the surface is therefore independent of the co-ordinate system used to evaluate it.

Example: The induced metric on the surface of a unit-radius two-sphere embedded in $\mathbb{R}^3$, is, expressed in polar angles,

$$\text{“}ds^2\text{”} = \mathbf{g}(\ ,\ ) = d\theta \otimes d\theta + \sin^2\theta\, d\phi \otimes d\phi.$$

Thus

$$g = \begin{vmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{vmatrix} = \sin^2\theta,$$

and

$$d(\text{Area}) = \sin\theta\, d\theta d\phi.$$

3.3 Stokes' Theorem

All the integral theorems of classical vector calculus are special cases of **Stokes' Theorem**: If $\partial\Omega$ denotes the (oriented) boundary of the (oriented) region Ω, then

$$\boxed{\int_\Omega d\omega = \int_{\partial\Omega} \omega.}$$

We will not provide a detailed proof. Apart from notation, it would parallel the proof of Stokes' or Green's theorems in ordinary vector calculus: The exterior derivative d has been defined so that the theorem holds for an infinitesimal square, cube, or hypercube. We therefore divide Ω into many such small regions. We then observe that the contributions of the interior boundary faces cancel because all interior faces are shared between two adjacent regions, and so occur twice with opposite orientations. Only the contribution of the outer boundary remains.
Example: If Ω is a region of $\mathbb{R}^2$, then from

$$d\left[\frac{1}{2}(x\,dy - y\,dx)\right] = dxdy,$$

we have

$$\text{Area}\,(\Omega) = \int_\Omega dxdy = \frac{1}{2}\int_{\partial\Omega}(x\,dy - y\,dx).$$

Example: Again, if Ω is a region of $\mathbb{R}^2$, then from $d[r^2 d\theta/2] = r\,dr d\theta$ we have

$$\text{Area}\,(\Omega) = \int_\Omega r\,dr d\theta = \frac{1}{2}\int_{\partial\Omega} r^2 d\theta.$$

Example: If Ω is the interior of a sphere of radius R, then

$$\int_\Omega dxdydz = \int_{\partial\Omega} x\,dydx = \frac{4}{3}\pi R^3.$$

Here we have referred back to (3.24) to evaluate the surface integral.
Example: Archimedes' tombstone.
Archimedes of Syracuse gave instructions that his tombstone should have displayed on it a diagram consisting of a sphere and circumscribed cylinder. Cicero, while serving as quæstor in Sicily, had the stone restored.[2] This has been said to be the only significant contribution by a Roman to pure mathematics. The carving on the stone was to commemorate Archimedes' results about the areas and volumes of spheres, including the one illustrated in figure 3.6, that the area of the spherical cap cut off by slicing through the cylinder is equal to the area cut off on the cylinder.

We can understand this result via Stokes' theorem: If the two-sphere S^2 is parameterized by spherical polar co-ordinates θ, ϕ, and Ω is a region on

[2]Marcus Tullius Cicero, *Tusculan Disputations*, Book V, Sections 64 – 66

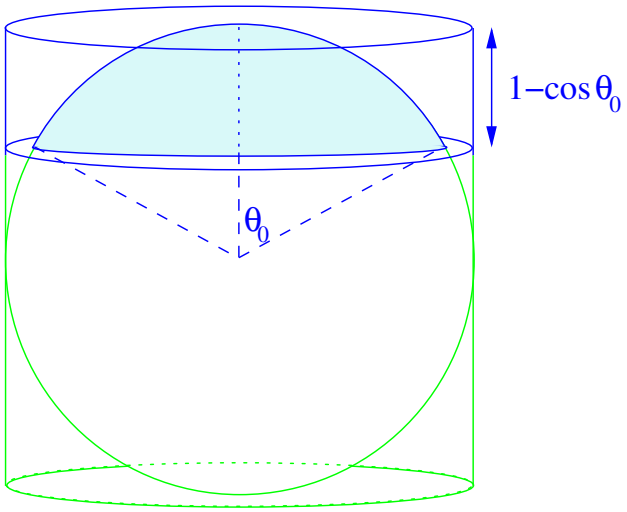

Figure 3.6: *Sphere and circumscribed cylinder.*

the sphere, then

$$\text{Area}\,(\Omega) = \int_\Omega \sin\theta d\theta d\phi = \int_{\partial\Omega} (1-\cos\theta) d\phi,$$

and applying this to the figure, where the cap is defined by $\theta < \theta_0$ gives

$$\text{Area}\,(\text{cap}) = 2\pi(1-\cos\theta_0)$$

which is indeed the area of the blue cylinder.

Exercise 3.2: The sphere S^n can be thought of as the locus of points in $\mathbb{R}^{n+1}$ obeying $\sum_{i=1}^{n+1}(x^i)^2 = 1$. Use its invariance under orthogonal transformations to show that the element of surface "volume" of the n-sphere can be written as

$$d(\text{Volume on } S^n) = \frac{1}{n!}\epsilon_{\alpha_1\alpha_2\ldots\alpha_{n+1}} x^{\alpha_1}\, dx^{\alpha_2}\ldots dx^{\alpha_{n+1}}.$$

Use Stokes' theorem to relate the integral of this form over the *surface* of the sphere to the volume of the *solid* unit sphere. Confirm that we get the correct proportionality between the volume of the solid unit sphere and the volume or area of its surface.

3.4 Applications

We now know how to integrate forms. What sort of forms should we seek to integrate? For a physicist working with a classical or quantum field, a plentiful supply of intesting forms is obtained by using the field to *pull back* geometric objects.

3.4.1 Pull-backs and Push-forwards

If we have a map ϕ from a manifold M to another manifold N, and we choose a point $x \in M$, we can *push forward* a vector from TM_x to $TN_{\phi(x)}$, in the obvious way (map head-to-head and tail-to-tail). This map is denoted by $\phi_* : TM_x \to TN_{\phi(x)}$.

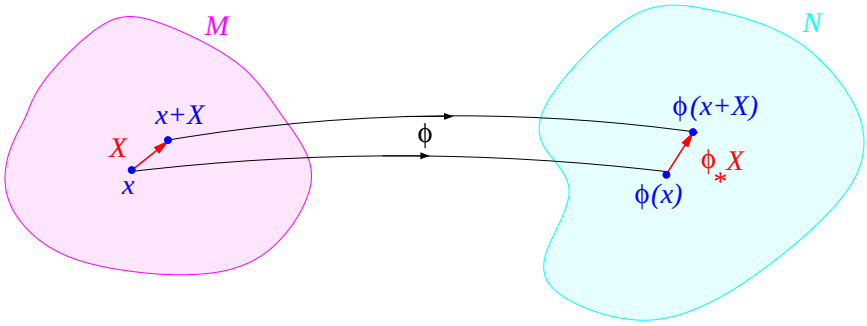

Figure 3.7: *Pushing forward a vector X from TM_x to $TN_{\phi(x)}$.*

If the vector X has components X^μ and the map takes the point with coordinates x^μ to one with coordinates $\xi^\mu(x)$, the vector $\phi_* X$ has components

$$(\phi_* X)^\mu = \frac{\partial \xi^\mu}{\partial x^\nu} X^\nu. \tag{3.29}$$

This looks very like the transformation formula for contravariant vector components under a change of coordinate system. What we are doing here is conceptually different, however. A change of co-ordinates produces a *passive* transformation — *i.e.* a new description for an unchanging vector. A push forward is an *active* transformation — we are changing a vector into different one. Furthermore, the map from $M \to N$ is not being assumed to be

one-to-one, so, contrary to the requirement imposed on a co-ordinate transformation, it may not be possible to invert the functions $\xi^\mu(x)$ and write the x^ν's as functions of the ξ^μ's.

While we can push forward individual vectors, we cannot always push forward a vector *field* X from TM to TN. If two distinct points x_1 and x_2, chanced to map to the same point $\xi \in N$, and $X(x_1) \neq X(x_2)$, we would not know whether to chose $\phi_*[X(x_1)]$ or $\phi_*[X(x_2)]$ as $[\phi_* X](\xi)$. This problem does not occur for differential forms. A map $\phi : M \to N$ induces a natural, and always well defined, *pull-back* map $\phi^* : \bigwedge^p (T^*N) \to \bigwedge^p (T^*M)$ which works as follows: Given a form $\omega \in \bigwedge^p (T^*N)$, we define $\phi^*\omega$ as a form on M by specifying what we get when we plug the vectors $X_1, X_2, \ldots, X_p \in TM$ into it. We evaluate the form at $x \in M$ by pushing the vectors $X_i(x)$ forward from TM_x to $TN_{\phi(x)}$, plugging them into ω at $\phi(x)$ and declaring the result to be the evaluation of $\phi^*\omega$ on the X_i at x. Symbolically

$$[\phi^*\omega](X_1, X_2, \ldots, X_p) = \omega(\phi_* X_1, \phi_* X_2, \ldots, \phi_* X_p). \tag{3.30}$$

This may seem rather abstract, but the idea is in practice quite simple: If the map takes $x \in M \to \xi(x) \in N$, and

$$\omega = \frac{1}{p!}\omega_{i_1 \ldots i_p}(\xi) d\xi^{i_1} \ldots d\xi^{i_p}, \tag{3.31}$$

then

$$\begin{aligned} \phi^*\omega &= \frac{1}{p!}\omega_{i_1 i_2 \ldots i_p}[\xi(x)] d\xi^{i_1}(x) d\xi_2^i(x) \cdots d\xi^{i_p}(x) \\ &= \frac{1}{p!}\omega_{i_1 i_2 \ldots i_p}[\xi(x)] \frac{\partial \xi^{i_1}}{\partial x^{\mu_1}} \frac{\partial \xi^{i_2}}{\partial x^{\mu_2}} \cdots \frac{\partial \xi^{i_p}}{\partial x^{\mu_1}} dx^{\mu_1} \cdots dx^{\mu_p}. \end{aligned} \tag{3.32}$$

Computationally, the process of pulling back a form is so transparent that it easy to confuse it with a simple change of variable. That it is not the same operation will become clear in the next few sections where we consider maps that are many-to-one.

Exercise 3.3: Show that the operation of taking an exterior derivative commutes with a pull back:

$$d\left[\phi^*\omega\right] = \phi^*(d\omega).$$

Exercise 3.4: If the map $\phi : M \to N$ is invertible then we may push forward a vector field X on M to get a vector field $\phi_* X$ on N. Show that

$$\mathcal{L}_X[\phi^*\omega] = \phi^*\left[\mathcal{L}_{\phi_* X}\omega\right].$$

Exercise 3.5: Again assume that $\phi : M \to N$ is invertible. By using the coordinate expressions for the Lie bracket and the effect of a push-forward, show that if X, Y are vector fields on TM then

$$\phi_*([X, Y]) = [\phi_* X, \phi_* Y],$$

as vector fields on TN.

3.4.2 Spin textures

As an application of pull-backs we will consider some of the topological aspects of *spin textures* which are fields of unit vectors $\mathbf{n}$, or "spins", in two or three dimensions.

Consider a smooth map $n : \mathbb{R}^2 \to S^2$ that assigns $x \mapsto \mathbf{n}(x)$, where $\mathbf{n}$ is a three-dimensional unit vector whose tip defines a point on the 2-sphere S^2. A physical example of such an $\mathbf{n}(x)$ would be the local direction of the spin polarization in a ferromagnetically-coupled two-dimensional electron gas.

In terms of $\mathbf{n}$, the area 2-form on the sphere becomes

$$\Omega = \frac{1}{2}\mathbf{n} \cdot (d\mathbf{n} \times d\mathbf{n}) \equiv \frac{1}{2}\epsilon_{ijk} n^i dn^j dn^k. \tag{3.33}$$

The n map pulls this area-form back to

$$F \equiv n^*\Omega = \frac{1}{2}(\epsilon_{ijk} n^i \partial_\mu n^j \partial_\nu n^k) dx^\mu dx^\nu = (\epsilon_{ijk} n^i \partial_1 n^j \partial_2 n^k)\, dx^1 dx^2 \tag{3.34}$$

which is a differential form in $\mathbb{R}^2$. We will call it the *topological charge density*. It measures the area on the two-sphere swept out by the $\mathbf{n}$ vectors as we explore a square in $\mathbb{R}^2$ of side dx^1 by dx^2.

Suppose now that the vector $\mathbf{n}$ tends some fixed direction at large distance. This allows us to think of "infinity" as a single point, and the assignment $x \mapsto \mathbf{n}(x)$ as a map from S^2 to S^2. Such maps are characterized topologically by their *"topological charge,"* or *winding number* N which counts the number of times the image of the originating x sphere wraps round the target $\mathbf{n}$-sphere. A mathematician would call this number the *Brouwer degree* of the map $\mathbf{n}$. It is intuitively plausible that a continuous map from a sphere to itself will wrap a whole number of times, and so we expect

$$N = \frac{1}{4\pi}\int_{\mathbb{R}^2} \left\{\epsilon_{ijk} n^i \partial_1 n^j \partial_2 n^k\right\} dx^1 dx^2, \tag{3.35}$$

to be an integer. We will soon show that this is indeed so, but first we will demonstrate that N is a *topological invariant*.

In two dimensions the form $F = n^*\Omega$ is automatically closed because the exterior derivative of any two-form is zero — there being no three-forms in two dimensions. Even if we consider an $\mathbf{n}(x^1, \ldots, x^m)$ field in $m > 2$ dimensions, however, we still have $dF = 0$. This is because

$$dF = \frac{1}{2}\epsilon^{ijk}\partial_\sigma n^i \partial_\mu n^j \partial_\nu n^k dx^\sigma dx^\mu dx^\nu. \tag{3.36}$$

If we insert infinitesimal vectors into the dx^μ to get their components δx^μ, we have to evaluate the triple-product of three vectors $\delta n^i = \partial_\mu n^i \delta x^\mu$, each of which is tangent to the two-sphere. But the tangent space of S^2 is two-dimensional, so any three tangent vectors $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, are linearly dependent and their triple-product $\mathbf{t}_1 \cdot (\mathbf{t}_2 \times \mathbf{t}_3)$ is zero.

Although it is closed, $F = n^*\Omega$ will not generally be the d of a globally defined one-form. Suppose, however, that we vary the map, $\mathbf{n} \to \mathbf{n} + \delta\mathbf{n}$. The change in the topological charge density is

$$\delta F = n^*[\mathbf{n} \cdot (d(\delta\mathbf{n}) \times d\mathbf{n})], \tag{3.37}$$

and this variation *can* be written as a total derivative

$$\delta F = d\{n^*[\mathbf{n} \cdot (\delta\mathbf{n} \times d\mathbf{n})]\} \equiv d\{\epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu\}. \tag{3.38}$$

In these manipulations we have used $\delta\mathbf{n} \cdot (d\mathbf{n} \times d\mathbf{n}) = d\mathbf{n} \cdot (\delta\mathbf{n} \times d\mathbf{n}) = 0$, the triple-products being zero for the same reason adduced earlier. From Stokes' theorem, we have

$$\delta N = \int_{S^2} \delta F = \int_{\partial S^2} \epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu. \tag{3.39}$$

Since $\partial S^2 = \emptyset$, we conclude that $\delta N = 0$ under any smooth deformation of the map $\mathbf{n}(x)$. This is what we mean when we say that N is a topological invariant. Equivalently, on $\mathbb{R}^2$, with $\mathbf{n}$ constant at infinity, we have

$$\delta N = \int_{\mathbb{R}^2} \delta F = \int_\Gamma \epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu, \tag{3.40}$$

where Γ is a curve surrounding the origin at large distance. Again $\delta N = 0$, this time because $\partial_\mu n^k = 0$ everywhere on Γ.

In some physical applications, the field $\mathbf{n}$ winds in localized regions called *Skyrmions*. These knots in the spin field behave very much as elementary particles, retaining their identity as they move through the material. The winding number counts how many Skyrmions (minus the number of anti-Skyrmions, which wind with opposite orientation) there are. To construct a smooth multi-Skyrmion map $\mathbb{R}^2 \to S^2$ with positive winding number N, take a set of $N+1$ complex numbers $\lambda, a_1, \ldots, a_N$ and another set of N numbers $b_1, \ldots, b_N$ such that no b coincides with any a. Then set

$$e^{i\phi} \tan \frac{\theta}{2} = \lambda \frac{(z-a_1)\ldots(z-a_N)}{(z-b_1)\ldots(z-b_N)} \tag{3.41}$$

where $z = x^1 + ix^2$, and θ and ϕ are spherical polar co-ordinates specifying the direction $\mathbf{n}$. At the points a_i the vector $\mathbf{n}$ points straight up, and at the points b_i it points straight down. You will show in exercise 3.12 that this particular $\mathbf{n}$-field configuration minimizes the energy functional

$$\begin{aligned} E[\mathbf{n}] &= \frac{1}{2}\int (\partial_1 \mathbf{n}\cdot\partial_1\mathbf{n} + \partial_2\mathbf{n}\cdot\partial_2\mathbf{n})\; dx^1 dx^2 \\ &= \frac{1}{2}\int \left(|\nabla n^1|^2 + |\nabla n^2|^2 + |\nabla n^3|^2\right)\; dx^1 dx^2 \end{aligned} \tag{3.42}$$

for the given winding number N. The next section will explain the geometric origin of the mysterious combination $e^{i\phi}\tan\theta/2$.

3.4.3 The Hopf Map

You may recall that in section 1.2.3 we defined *complex projective space* $\mathbb{C}P^n$ to be the set of *rays* in a complex $n+1$ dimensional vector space. A ray is an equivalence classes of vectors $[\zeta_1, \zeta_2, \ldots, \zeta_{n+1}]$, where the ζ_i are not all zero, and where we do not distinguish between $[\zeta_1, \zeta_2, \ldots, \zeta_{n+1}]$ and $[\lambda\zeta_1, \lambda\zeta_2, \ldots, \lambda\zeta_{n+1}]$ for non-zero λ. The space of rays is a $2n$-dimensional real manifold: in a region where ζ_{n+1} does not vanish, we can take as co-ordinates the real numbers $\xi_1, \ldots, \xi_n$, $\eta_1, \ldots, \eta_n$ where

$$\xi_1 + i\eta_1 = \frac{\zeta_1}{\zeta_{n+1}}, \quad \xi_2 + i\eta_2 = \frac{\zeta_2}{\zeta_{n+1}}, \ldots, \xi_n + i\eta_n = \frac{\zeta_n}{\zeta_{n+1}}. \tag{3.43}$$

Similar co-ordinate charts can be constructed in the regions where other ζ_i are non-zero. Every point in $\mathbb{C}P^n$ lies in at least one of these co-ordinate charts,

and the co-ordinate transformation rules for going from chart to another are smooth.

The simplest complex projective space, $\mathbb{C}P^1$, is the real two-sphere S^2 in disguise. This rather non-obvious fact is revealed by the use of a *stereographic map* to make the equivalence class $[\zeta_1, \zeta_2] \in \mathbb{C}P^1$ correspond to a point $\mathbf{n}$ on the sphere. When ζ_1 is non zero, the class $[\zeta_1, \zeta_2]$ is uniquely determined by the ratio $\zeta_2/\zeta_1 = |\zeta_2/\zeta_1|e^{i\phi}$, which we plot on the complex plane. We think of this copy of $\mathbb{C}$ as being the x, y plane in $\mathbb{R}^3$. We then draw a straight line connecting the plotted point to the south pole of a unit sphere circumscribed about the origin in $\mathbb{R}^3$. The point where this line (continued if necessary) intersects the sphere is the tip of the unit vector $\mathbf{n}$.

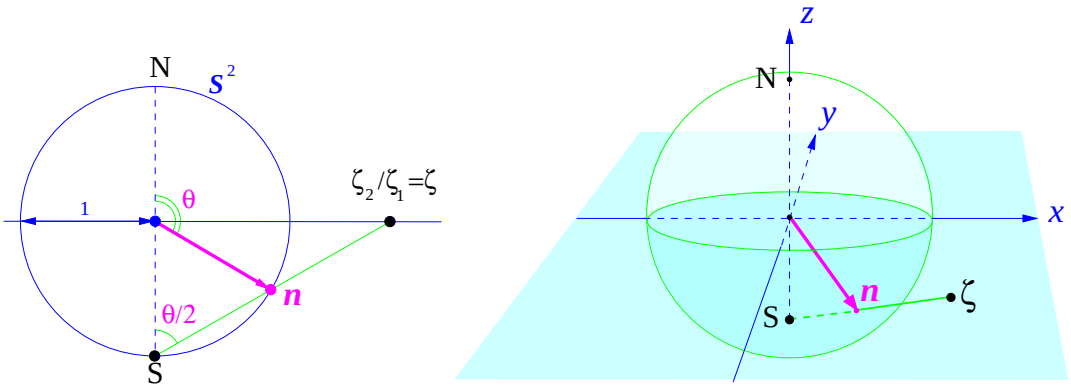

Figure 3.8: *Two views of the sterographic map between the two-sphere and the complex plane. The point* $\zeta = \zeta_2/\zeta_1 \in \mathbb{C}$ *corresponds to the unit vector* $\mathbf{n} \in S^2$.

If ζ_2, were zero, we would end up at the north pole where the $\mathbb{R}^3$ co-ordinate z takes the value $z = 1$. If ζ_1 goes to zero with ζ_2 fixed, we move smoothly to the south pole $z = -1$. We therefore extend the definition of our map to the case $\zeta_1 = 0$ by making the equivalence class $[0, \zeta_2]$ correspond to the south pole. We can find an explicit formula for this map. Figure 3.8 shows that $\zeta_2/\zeta_1 = e^{i\phi} \tan\theta/2$, and this relation suggests the use of the "t"-substitution formulae

$$\sin\theta = \frac{2t}{1+t^2}, \quad \cos\theta = \frac{1-t^2}{1+t^2}, \tag{3.44}$$

where $t = \tan\theta/2$. Since the x, y, z components of $\mathbf{n}$ are given by

$$n^1 = \sin\theta\cos\phi,$$

$$\begin{aligned} n^2 &= \sin\theta\sin\phi, \\ n^3 &= \cos\theta, \end{aligned}$$

we find that

$$n^1 + in^2 = \frac{2(\zeta_2/\zeta_1)}{1+|\zeta_2/\zeta_1|^2}, \quad n^3 = \frac{1-|\zeta_2/\zeta_1|^2}{1+|\zeta_2/\zeta_1|^2}. \tag{3.45}$$

We can multiply through by $|\zeta_1|^2 = \overline{\zeta}_1\zeta_1$, and so write this correspondence in a more symmetrical manner:

$$\begin{aligned} n^1 &= \frac{\overline{\zeta}_1\zeta_2 + \overline{\zeta}_2\zeta_1}{|\zeta_1|^2+|\zeta_2|^2} \\ n^2 &= \frac{1}{i}\left(\frac{\overline{\zeta}_1\zeta_2 - \overline{\zeta}_2\zeta_1}{|\zeta_1|^2+|\zeta_2|^2}\right), \\ n^3 &= \frac{|\zeta_1|^2-|\zeta_2|^2}{|\zeta_1|^2+|\zeta_2|^2}. \end{aligned} \tag{3.46}$$

This last form can be conveniently expressed in terms of the Pauli sigma matrices

$$\hat{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \hat{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \hat{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.47}$$

as

$$\begin{aligned} n^1 &= (\overline{z}_1, \overline{z}_2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \\ n^2 &= (\overline{z}_1, \overline{z}_2)\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \\ n^3 &= (\overline{z}_1, \overline{z}_2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \end{aligned} \tag{3.48}$$

where

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \frac{1}{\sqrt{|\zeta_1|^2+|\zeta_2|^2}}\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{3.49}$$

is a normalized 2-vector, which we think of as a *spinor*.

The $\mathbb{C}P^1 \simeq S^2$ correspondence now has a quantum mechanical interpretation: Any unit three-vector $\mathbf{n}$ can be obtained as the expectation value

of the $\hat{\sigma}$ matrices in a normalized spinor state. Conversly, any normalized spinor $\psi = (z_1, z_2)^T$ gives rise to a unit vector *via*

$$n^i = \psi^\dagger \hat{\sigma}^i \psi. \tag{3.50}$$

Now, since

$$1 = |z_1|^2 + |z_2|^2, \tag{3.51}$$

the normalized spinor can be thought of as defining a point in S^3. This means that the one-to-one correspondence $[z_1, z_2] \leftrightarrow \mathbf{n}$ also gives rise to a map from $S^3 \to S^2$. This is called the *Hopf map*:

$$\text{Hopf} : S^3 \to S^2. \tag{3.52}$$

The dimension reduces from three to two, so the Hopf map cannot be one-to-one. Even after we have normalized $[\zeta_1, \zeta_2]$, we are still left with a choice of overall phase. Both (z_1, z_2) and $(z_1 e^{i\theta}, z_2 e^{i\theta})$, although distinct points in S^3, correspond to the same point in $\mathbb{C}P^1$, and hence in S^2. The inverse image of a point in S^2 is a geodesic circle in S^3. Later we will show that any two such geodesic circles are linked, and this makes the Hopf map topologically non-trivial in that it cannot be continuously deformed to a constant map, *i.e.* to a map that takes all of S^3 to a single point in S^2.

Exercise 3.6: We have seen that the stereographic map relates the point with spherical polar co-ordinates θ, ϕ to the complex number

$$\zeta = e^{i\phi} \tan \theta/2.$$

We can therefore set $\zeta = \xi + i\eta$ and take ξ, η as *stereographic co-ordinates* on the sphere. Show that in these co-ordinates the sphere metric is given by

$$\begin{aligned} g(\ ,\ ) &\equiv d\theta \otimes d\theta + \sin^2\theta \, d\phi \otimes d\phi \\ &= \frac{2}{(1+|\zeta|^2)^2}(d\overline{\zeta} \otimes d\zeta + d\zeta \otimes d\overline{\zeta}) \\ &= \frac{4}{(1+\xi^2+|\eta|^2)^2}(d\xi \otimes d\xi + d\eta \otimes d\eta), \end{aligned}$$

and the area 2-form becomes

$$\begin{aligned} \Omega &\equiv \sin\theta \, d\theta \wedge d\phi \\ &= \frac{2i}{(1+|\zeta|^2)^2} \, d\zeta \wedge d\overline{\zeta} \\ &= \frac{4}{(1+\xi^2+\eta^2)^2} \, d\xi \wedge d\eta. \end{aligned} \tag{3.53}$$

3.4.4 Homotopy and the Hopf map

We can use the Hopf map to factor the map $n : x \mapsto \mathbf{n}(x)$ through the three-sphere by specifying the spinor ψ at each point, instead of the vector $\mathbf{n}$, and so mapping indirectly

$$\mathbb{R}^2 \stackrel{\psi}{\to} S^3 \stackrel{\text{Hopf}}{\to} S^2.$$

It might seem that for a given spin-field $\mathbf{n}(x)$ we can choose the overall phase of $\psi(x) \equiv (z_1(x), z_2(x))^T$ as we like, but if we demand that the z_i's be *continuous* functions of x there is a rather non-obvious topological restriction which has important physical consequences. To see how this comes about we first express the winding number in terms of the z_i. We find (after a page or two of algebra)

$$F = (\epsilon_{ijk} n^i \partial_1 n^j \partial_2 n^k)\, dx^1 dx^2 = \frac{2}{i} \sum_{i=1}^{2} (\partial_1 \overline{z}_i \partial_2 z_i - \partial_2 \overline{z}_i \partial_1 z_i)\, dx^1 dx^2, \quad (3.54)$$

and so the topological charge N is given by

$$N = \frac{1}{2\pi i} \int \sum_{i=1}^{2} (\partial_1 \overline{z}_i \partial_2 z_i - \partial_2 \overline{z}_i \partial_1 z_i)\, dx^1 dx^2. \quad (3.55)$$

Now, when written in terms of the z_i variables, the form F becomes a total derivative:

$$\begin{aligned} F &= \frac{2}{i} \sum_{i=1}^{2} (\partial_1 \overline{z}_i \partial_2 z_i - \partial_2 \overline{z}_i \partial_1 z_i)\, dx^1 dx^2 \\ &= d \left\{ \frac{1}{i} \sum_{i=1}^{2} (\overline{z}_i \partial_\mu z_i - (\partial_\mu \overline{z}_i) z_i)\, dx^\mu \right\}. \end{aligned} \quad (3.56)$$

Further, because $\mathbf{n}$ is fixed at large distance, we have $(z_1, z_2) = e^{i\theta}(c_1, c_2)$ near infinity, where c_1, c_2 are constants with $|c_1|^2 + |c_2|^2 = 1$. Thus, near infinity,

$$\frac{1}{2i} \sum_{i=1}^{2} (\overline{z}_i \partial_\mu z_i - (\partial_\mu \overline{z}_i) z_i) \to (|c_1|^2 + |c_2|^2) d\theta = d\theta. \quad (3.57)$$

We combine this observation with Stokes' theorem to obtain

$$N = \frac{1}{2\pi i} \int_\Gamma \frac{1}{2} \sum_{i=1}^{2} (\overline{z}_i \partial_\mu z_i - (\partial_\mu \overline{z}_i) z_i)\, dx^\mu = \frac{1}{2\pi} \int_\Gamma d\theta. \quad (3.58)$$

Here, as in the previous section, Γ is a curve surrounding the origin at large distance. Now $\int d\theta$ is the total change in θ as we circle the boundary. While the phase $e^{i\theta}$ has to return to its original value after a round trip, the angle θ can increase by an integer multiple of 2π. The *winding number* $\oint d\theta/2\pi$ can therefore be non-zero, but must be an integer.

We have uncovered the rather surpring fact that the topological charge of the map $n : S^2 \to S^2$ is equal to the winding number of the phase angle θ at infinity. This is the topological constraint refered to earlier. As a byproduct, we have confirmed our conjecture that the topological charge N is an integer. The existence of this integer invariant shows that the smooth maps $n : S^2 \to S^2$ fall into distinct *homotopy classes* labeled by N. Maps with different values of N cannot be continuously deformed into one another, and, while we have not shown that it is so, two maps with the same value of N can be deformed into each other.

Maps that can be continuously deformed one into the other are said to be *homotopic*. The set of homotopy classes of the maps of the n-sphere into a manifold M is denoted by $\pi_n(M)$. In the present case $M = S^2$. We are therefore claiming that

$$\pi_2(S^2) = \mathbb{Z}, \tag{3.59}$$

where we are identifying the homotopy class with its winding number $N \in \mathbb{Z}$.

3.4.5 The Hopf index

We have so far discussed maps from S^2 to S^2. It is perhaps not too surprising that such maps are classified by a winding number. What is rather more surprising is that maps $n : S^3 \to S^2$ also have an associated topological number. If we continue to assume that $\mathbf{n}$ tends to a constant direction at infinity so that we can think of $\mathbb{R}^3 \cup \{\infty\}$ as being S^3, this number will label the homotopy classes $\pi_3(S^2)$ of fields of unit vectors $\mathbf{n}$ in *three* dimensions. We will think of the third dimension as being time. In this situation an interesting set of $\mathbf{n}$ fields to consider are the $\mathbf{n}(x,t)$ corresponding moving Skyrmions. The world lines of these Skyrmions will be tubes outside of which $\mathbf{n}$ is constant, and such that on any slice through the tube, $\mathbf{n}$ will cover the target $\mathbf{n}$-sphere once.

To motivate the formula we will find for the topological number, we begin with a problem from magnetostatics. Suppose we are given a cable originally made up of a bundle of many parallel wires. The cable is then twisted N

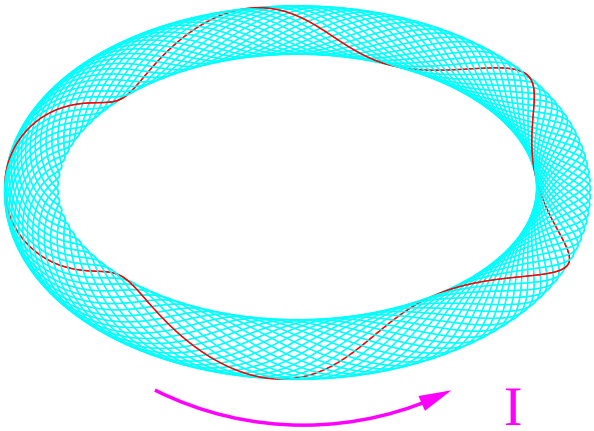

Figure 3.9: *A twisted cable with* $N = 5$.

times about its axis and bent into a closed loop, the end of each individual wire being attached to its begining to make a continuous circuit. A current I flows in the cable in such a manner that each individual wire carries only a small part δI_i of the total. The sense of the current is such that as we flow with it around the cable each wire wraps N times anticlockwise about all the others. The current produces a magnetic field $\mathbf{B}$. Can we determine the integer twisting number N knowing only this $\mathbf{B}$ field?
The answer is *yes*. We use Ampere's law in integral form,

$$\oint_\Gamma \mathbf{B} \cdot d\mathbf{r} = (\text{current encircled by } \Gamma). \tag{3.60}$$

We also observe that the current density $\nabla \times \mathbf{B} = \mathbf{J}$ at a point is directed along the tangent to the wire passing through that point. We therefore integrate along each individual wire as it encircles the others, and sum over the wires to find

$$\sum_{\text{wires } i} \delta I_i \oint \mathbf{B} \cdot d\mathbf{r}_i = \int \mathbf{B} \cdot \mathbf{J}\, d^3x = \int \mathbf{B} \cdot (\nabla \times \mathbf{B})\, d^3x = NI^2. \tag{3.61}$$

We now apply this insight to our three-dimensional field of unit vectors $\mathbf{n}(x)$.

The quantity playing the role of the current density $\mathbf{J}$ is the *topological current*

$$J^\sigma = \frac{1}{2}\epsilon^{\sigma\mu\nu}\epsilon_{ijk}n^i\partial_\mu n^j\partial_\nu n^k. \tag{3.62}$$

We note that $\nabla\cdot\mathbf{J} = 0$. This is simply another way of saying that the 2-form $F = n^*\Omega$ is closed.

The flux of $\mathbf{J}$ through a surface S is

$$\int_S \mathbf{J}\cdot d\mathbf{S} = \int_S F \tag{3.63}$$

and this is the area of the spherical surface covered by the $\mathbf{n}$'s. A Skyrmion, for example, has total topological current $I = 4\pi$, the total surface area of the 2-sphere. The Skyrmion world-line will play the role of the cable, and the inverse images in $\mathbb{R}^3$ of points on S^2 correspond to the individual wires.

If form language, the field corresponding to $\mathbf{B}$ can be any one-form A such that $dA = F$. Thus

$$N_{\text{Hopf}} = \frac{1}{I^2}\int_{\mathbb{R}^3}\mathbf{B}\cdot\mathbf{J}\,d^3x = \frac{1}{16\pi^2}\int_{\mathbb{R}^3}AF \tag{3.64}$$

will be an integer. This integer is the *Hopf linking number*, or *Hopf index*, and counts the number of times the Skyrmion twists before it bites its tail to form a closed-loop world-line.

There is another way of obtaining this formula, and of understanding the number $16\pi^2$. We observe that the two-form F and the one-form A are the pull-back from S^3 to $\mathbb{R}^3$ along ψ of the forms

$$\begin{aligned}\mathcal{F} &= \frac{1}{i}\sum_{i=1}^{2}\left(d\overline{z}_i dz_i - dz_i d\overline{z}_i\right),\\ \mathcal{A} &= \frac{1}{i}\sum_{i=1}^{2}\left(\overline{z}_i dz_i - z_i d\overline{z}_i\right),\end{aligned} \tag{3.65}$$

respectively. If we substitute $z_{1,2} = \xi_{1,2} + i\eta_{1,2}$, we find that

$$\mathcal{A}F = 8(\xi_1 d\eta_1 d\xi_2 d\eta_2 - \eta_1 d\xi_1 d\xi_2 d\eta_2 + \xi_2 d\eta_2 d\xi_1 d\eta_1 - \eta_2 d\xi_2 d\xi_1 d\eta_1). \tag{3.66}$$

We know from exercise 3.2 that this expression is eight times the volume 3-form on the three-sphere. Now the total volume of the unit three-sphere is $2\pi^2$, and so, from our factored map $x \mapsto \psi \mapsto \mathbf{n}$ we have that

$$N_{\text{Hopf}} = \frac{1}{16\pi^2}\int_{\mathbb{R}^3}\psi^*(\mathcal{A}F) = \frac{1}{2\pi^2}\int_{\mathbb{R}^3}\psi^* d(\text{Volume on } S^3) \tag{3.67}$$

is the number of times the normalized spinor $\psi(x)$ covers S^3 as x covers $\mathbb{R}^3$. For the Hopf map itself, this number is unity, and so the loop in S^3 which is the inverse image of a point in S^2 will twist once around any other such inverse image loop.

We have now established that

$$\pi_3(S^2) = \mathbb{Z}. \tag{3.68}$$

This result, implying that there are many maps from the three-sphere to the two-sphere that are not smoothly deformable to a constant map, was an great surprise when Hopf discovered it.

One of the principal physics consequences of the existence of the Hopf index is that "quantum lump" quasi-particles like the Skyrmion can be fermions, even though they are described by commuting (and therefore boson) fields. To understand how this can be, we first explain that the collection of homotopy classes $\pi_n(M)$ is not just a *set*. It has the additional structure of being a *group*: we can compose two homotopy classes to get a third, the composition is associative, and each homotopy class has an inverse. To define the group composition law, we think of S^n as the interior of an n-dimensional cube with the map $f : S^n \to M$ taking a fixed value $m_0 \in M$ at all points on the boundary of the cube. The boundary can then be considered to be a single point on S^n. We then take one of the n dimensions as being "time" and place two cubes and their maps f_1, f_2 into contact, with f_1 being "earlier" and f_2 being "later." We thus get a continuous map from a bigger box into M. The homotopy class of this map, after we relax the condition that the map takes the value m_0 on the common boundary, defines the composition $[f_2] \circ [f_1]$ of the two homotopy classes corresponding to f_1 and f_2. The composition may be shown to be independent of the choice of representative functions in the two classes. The inverse of a homotopy class $[f]$ is obtained by reversing the direction of "time" for each of the maps in the class. This group structure appears to depend on the fixed point m_0. As long as M is arcwise connected, however, the groups obtained from different m_0's are *isomorphic*, or equivalent. In the case of $\pi_2(S^2) = \mathbb{Z}$ and $\pi_3(S^2) = \mathbb{Z}$, the composition law is simply the addition of the integers $N \in \mathbb{Z}$ that label the classes. A full account of homotopy theory for working physicists is to be found in a readable review article by David Mermin.[3]

[3]N. D. Mermin, "The topological theory of defects in ordered media." *Rev. Mod. Phys.* **51** (1979) 591.

When we quantize using Feynman's "sum over histories" path integral, we may multiply the contributions of histories f that are not deformable into one another by different phase factors $\exp\{i\phi([f])\}$. The choice of phases must, however, be compatible with the composition of histories by concatenating one after the other – essentially the same operation as composing homotopy classes. This means that the product $\exp\{i\phi([f_1])\}\exp\{i\phi([f_2])\}$ of the phase factors for two possible histories must be the phase factor $\exp\{i\phi([f_2]\circ[f_1])\}$ assigned to the composition of their homotopy classes. If our quantum system consists of spins $\mathbf{n}$ in two space and one time dimension we can consistently assign a phase factor $\exp(i\pi N_{\rm Hopf})$ to a history. The rotation of a single Skyrmion through 2π makes $N_{\rm Hopf}=1$ and so the wavefunction changes sign. We will show in the next section, that a history where two particles change places can be continuously deformed into a history where they do not interchange, but instead one of them is twisted through 2π. The wavefunction of a pair of Skyrmions therefore changes sign when they are interchanged. This means that the quantized Skyrmion is a fermion.

3.4.6 Twist and Writhe

Consider two oriented non-intersecting closed curves γ_1 and γ_2. We can use Ampère's law to count the number of times γ_1 encircles γ_2 by imagining that γ_2 carries a unit current in the direction of its orientation, and evaluating

$$\begin{aligned}\mathrm{Lk}(\gamma_1,\gamma_2) &= \oint_{\gamma_1}\mathbf{B}(\mathbf{r}_1)\cdot d\mathbf{r}_1\\ &= \frac{1}{4\pi}\oint_{\gamma_1}\oint_{\gamma_2}\frac{(\mathbf{r}_1-\mathbf{r}_2)\cdot(d\mathbf{r}_1\times d\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|^3}.\end{aligned}\tag{3.69}$$

Here the second line follows from the first by an application of the Biot-Savart law to compute the $\mathbf{B}$ field due the current. The second line shows that the *Gauss linking number* $\mathrm{Lk}(\gamma_1,\gamma_2)$ is symmetric under the interchange $\gamma_1\leftrightarrow\gamma_2$ of the two curves. It changes sign, however, if one of the curves changes orientation, or if the pair of curves is reflected in a mirror.

Introduce parameters t_1, t_2 with $0<t_1,t_2\le 1$ to label points on the two curves. The curves are closed, so $\mathbf{r}_1(0)=\mathbf{r}_1(1)$, and similarly for $\mathbf{r}_2$. Let us also define a unit vector

$$\mathbf{n}(t_1,t_2)=\frac{\mathbf{r}_1(t_1)-\mathbf{r}_2(t_2)}{|\mathbf{r}_1(t_1)-\mathbf{r}_2(t_2)|}.\tag{3.70}$$

Then

$$\begin{aligned} \mathrm{Lk}(\gamma_1,\gamma_2) &= \frac{1}{4\pi}\oint_{\gamma_1}\oint_{\gamma_2}\frac{\mathbf{r}_1(t_1)-\mathbf{r}_2(t_2)}{|\mathbf{r}_1(t_1)-\mathbf{r}_2(t_2)|^3}\cdot\left(\frac{\partial\mathbf{r}_1}{\partial t_1}\times\frac{\partial\mathbf{r}_2}{\partial t_2}\right)dt_1dt_2 \\ &= -\frac{1}{4\pi}\int_{T^2}\mathbf{n}\cdot\left(\frac{\partial\mathbf{n}}{\partial t_1}\times\frac{\partial\mathbf{n}}{\partial t_2}\right)dt_1dt_2. \end{aligned} \tag{3.71}$$

is seen to be (minus) the winding number of the map

$$n : [0,1]\times[0,1]\to S^2. \tag{3.72}$$

of the 2-torus into the sphere. Our previous results on maps into the 2-sphere therefore confirm our Ampère-law intuition that $\mathrm{Lk}(\gamma_1,\gamma_2)$ is an integer. The linking number is also topological invariant, being unchanged under any deformation of the curves that does not cause one to pass through the other.

An important application of these ideas occurs in biology, where the curves are the two complementary strands of a closed loop of DNA. We can think of two such parallel curves as forming the edges of a *ribbon* $\{\gamma_1,\gamma_2\}$ of width ϵ. Let use denote by γ the curve $\mathbf{r}(t)$ running along the axis of the ribbon midway between γ_1 and γ_2. The unit tangent to γ at the point $\mathbf{r}(t)$ is

$$\mathbf{t}(t) = \frac{\dot{\mathbf{r}}(t)}{|\dot{\mathbf{r}}(t)|}, \tag{3.73}$$

where the dots denote differentiation with respect to t. We also introduce a unit vector $\mathbf{u}(t)$ that is perpendicular to $\mathbf{t}(t)$ and lies in the ribbon, pointing from $\mathbf{r}_1(t)$ to $\mathbf{r}_2(t)$.

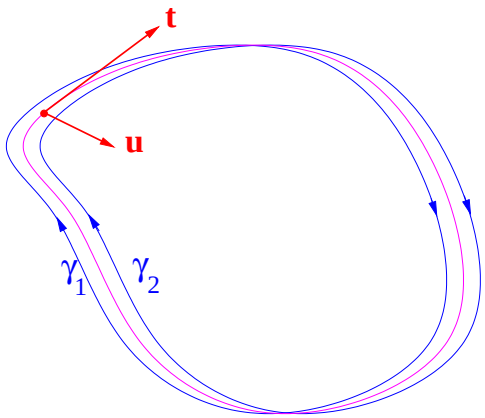

Figure 3.10: *An oriented ribbon* $\{\gamma_1,\gamma_2\}$ *showing the vectors* $\mathbf{t}$ *and* $\mathbf{u}$.

We will assign a common value of the parameter t to a point on γ and the points nearest to $\mathbf{r}(t)$ on γ_1 and γ_2. Consequently

$$\begin{aligned} \mathbf{r}_1(t) &= \mathbf{r}(t) - \frac{1}{2}\epsilon\,\mathbf{u}(t) \\ \mathbf{r}_2(t) &= \mathbf{r}(t) + \frac{1}{2}\epsilon\,\mathbf{u}(t) \end{aligned} \tag{3.74}$$

We can express $\dot{\mathbf{u}}$ as

$$\dot{\mathbf{u}} = \boldsymbol{\omega} \times \mathbf{u} \tag{3.75}$$

for some angular-velocity vector $\boldsymbol{\omega}(t)$. The quantity

$$\mathrm{Tw} = \frac{1}{2\pi}\oint_\gamma (\boldsymbol{\omega}\cdot\mathbf{t})\,dt \tag{3.76}$$

is called the *Twist* of the ribbon. It is not usually an integer, and is a property of the ribbon $\{\gamma_1, \gamma_2\}$ itself, being independent of the choice of parameterization t.

If we set $\mathbf{r}_1(t)$ and $\mathbf{r}_2(t)$ equal to the single axis curve $\mathbf{r}(t)$ in the integrand of (3.69), the resulting "self-linking" integral, or *Writhe*,

$$\mathrm{Wr} \stackrel{\mathrm{def}}{=} \frac{1}{4\pi}\oint_\gamma\oint_\gamma \frac{(\mathbf{r}(t_1)-\mathbf{r}(t_2))\cdot(\dot{\mathbf{r}}(t_1)\times\dot{\mathbf{r}}(t_2))}{|\mathbf{r}(t_1)-\mathbf{r}(t_2)|^3}\,dt_1 dt_2. \tag{3.77}$$

remains convergent despite the factor of $|\mathbf{r}(t_1)-\mathbf{r}(t_2)|^3$ in the denominator. However, if we try to achieve this substitution by making the width of the ribbon ϵ tend to zero, we find that the vector $\mathbf{n}(t_1, t_2)$ abruptly reverses its direction as t_1 passes t_2. In the limit of infinitesimal width this violent motion provides a delta-function contribution

$$-(\boldsymbol{\omega}\cdot\mathbf{t})\delta(t_1-t_2)\,dt_1\wedge dt_2 \tag{3.78}$$

to the 2-sphere area swept out by $\mathbf{n}$, and this contribution is invisible to the Writhe integral. The Writhe is a property only of the overall shape of the axis curve γ, and is independent both of the ribbon that contains it, and of the choice of parameterization. The linking number, on the other hand, is independent of ϵ, so the $\epsilon \to 0$ limit of the linking-number integral is not the integral of the $\epsilon \to 0$ limit of its integrand. Instead we have

$$\mathrm{Lk}(\gamma_1,\gamma_2) = \frac{1}{2\pi}\oint_\gamma (\boldsymbol{\omega}\cdot\mathbf{t})\,dt + \frac{1}{4\pi}\oint_\gamma\oint_\gamma \frac{(\mathbf{r}(t_1)-\mathbf{r}(t_2))\cdot(\dot{\mathbf{r}}(t_1)\times\dot{\mathbf{r}}(t_2))}{|\mathbf{r}(t_1)-\mathbf{r}(t_2)|^3}\,dt_1 dt_2 \tag{3.79}$$

This formula

$$\mathrm{Lk} = \mathrm{Tw} + \mathrm{Wr} \tag{3.80}$$

is known as the *Calugareanu-White-Fuller* relation, and is the basis for the claim, made in the previous section, that the worldline of an extended particle with an exchange ($\mathrm{Wr} = \pm 1$) can be deformed into a worldline with a 2π rotation ($\mathrm{Tw} = \pm 1$) without changing the topologically invariant linking number.

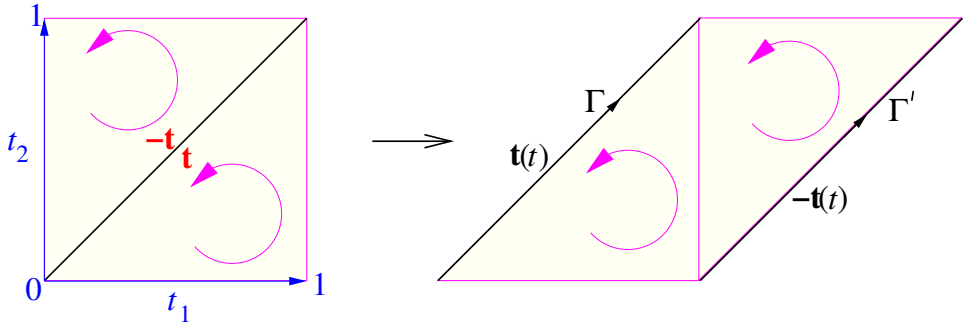

Figure 3.11: *Cutting and reassembling the domain of integration in* (3.82).

By setting

$$\mathbf{n}(t_1, t_2) = \frac{\mathbf{r}(t_1) - \mathbf{r}(t_2)}{|\mathbf{r}(t_1) - \mathbf{r}(t_2)|}. \tag{3.81}$$

we can express the Writhe as

$$\mathrm{Wr} = -\frac{1}{4\pi} \int_{T^2} \mathbf{n} \cdot \left(\frac{\partial \mathbf{n}}{\partial t_1} \times \frac{\partial \mathbf{n}}{\partial t_2} \right) dt_1 dt_2, \tag{3.82}$$

but we must take care to recognize that this new $\mathbf{n}(t_1, t_2)$ is discontinuous across the line $t = t_1 = t_2$. It is equal to $\mathbf{t}(t)$ for t_1 infinitesimally larger than t_2, and equal to $-\mathbf{t}(t)$ when t_1 is infinitesimally smaller than t_2. By cutting the square domain of integration and reassembling it into a rhomboid, as shown in figure 3.11, we obtain a continuous integrand and see that the Writhe is (minus) the 2-sphere area (counted with multiplicies and divided by 4π) of a region whose boundary is composed of two curves Γ, the *tangent indicatrix*, or *tantrix*, on which $\mathbf{n} = \mathbf{t}(t)$, and its oppositely oriented antipodal counterpart Γ' on which $\mathbf{n} = -\mathbf{t}(t)$.

The 2-sphere area $\Omega(\Gamma)$ bounded by Γ is only determined by Γ up to the addition of integer multiples of 4π. Taking note that the "wrong" orientation

of the boundary Γ (see figure 3.11 again) compensates for the minus sign before the integral in (3.82), we have

$$4\pi \mathrm{Wr} = 2\Omega(\Gamma) + 4\pi n. \tag{3.83}$$

Thus,

$$\mathrm{Wr} = \frac{1}{2\pi}\Omega(\Gamma), \ \mathrm{mod}\, 1. \tag{3.84}$$

We can do better than (3.84) once we realize that by allowing crossings we can continuously deform any closed curve into a perfect circle. Each self-crossing causes Lk and Wr (but not Tw which, being a local functional, does not care about crossings) to jump by ± 2. For a perfect circle $\mathrm{Wr} = 0$ whilst $\Omega = 2\pi$. We therefore have an improved estimate of the additive integer that is left undetermined by Γ, and from it we obtain

$$\mathrm{Wr} = 1 + \frac{1}{2\pi}\Omega(\Gamma), \ \mathrm{mod}\, 2. \tag{3.85}$$

This result is due to Brock Fuller.[4]

We can use our ribbon language to describe conformational transitions in long molecules. The elastic energy of a closed rod (or DNA molecule) can be approximated by

$$E = \int_\gamma \left\{ \frac{1}{2}\alpha(\boldsymbol{\omega}\cdot\mathbf{t})^2 + \frac{1}{2}\beta\kappa^2 \right\} ds \tag{3.86}$$

Here we are parameterizing the curve by its arc-length s. The constant α is the torsional stiffness coefficient, β is the flexural stiffness, and

$$\kappa(s) = \left| \frac{d^2\mathbf{r}(s)}{ds^2} \right| = \left| \frac{d\mathbf{t}(s)}{ds} \right|, \tag{3.87}$$

is the local curvature. Suppose that our molecule has linking number n, *i.e* it was twisted n times before the ends were joined together to make a loop.

[4]F. Brock Fuller, *Proc. Natl. Acad. Sci. USA,* **75** (1978) 3557 - 61.

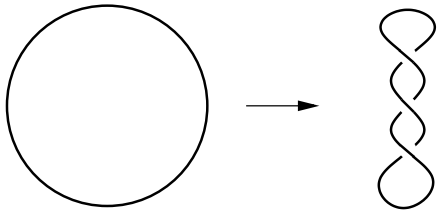

Figure 3.12: *A molecule initially with* Lk = 3, Tw = 3, Wr = 0 *writhes to a new configuration with* Lk = 3, Tw = 0, Wr = 3.

When $\beta \gg \alpha$ the molecule will minimize its bending energy by forming a planar circle with Wr ≈ 0 and Tw $\approx n$. If we increase α, or decrease β, there will come a point at which the molecule will seek to save torsional energy at the expense of bending, and will suddenly writhe into a new configuration with Wr $\approx n$ and Tw ≈ 0. Such twist-to-writhe transformations will be familiar to anyone who has struggled to coil a garden hose or electric cable.

3.5 Exercises and Problems

Exercise 3.7: *Old exam problem.* A two-form is expressed in Cartesian coordinates as,

$$\omega = \frac{1}{r^3}(zdxdy + xdydz + ydzdx)$$

where $r = \sqrt{x^2 + y^2 + z^2}$.

a) Evaluate $d\omega$ for $r \neq 0$.

b) Evaluate the integral

$$\Phi = \int_P \omega$$

over the infinite plane $P = \{-\infty < x < \infty, -\infty < y < \infty, z = 1\}$.

c) A sphere is embedded into $\mathbb{R}^3$ by the map φ, which takes the point $(\theta, \phi) \in S^2$ to the point $(x, y, z) \in \mathbb{R}^3$, where

$$\begin{aligned} x &= R\cos\phi\sin\theta \\ y &= R\sin\phi\sin\theta \\ z &= R\cos\theta. \end{aligned}$$

Pull back ω and find the 2-form $\varphi^*\omega$ on the sphere. (**Hint**: The form $\varphi^*\omega$ is both familiar and simple. If you end up with an intractable mess of trigonometric functions, you have made an algebraic error.)

d) By exploiting the result of part c), or otherwise, evaluate the integral

$$\Phi = \int_{S^2(R)} \omega$$

where $S^2(R)$ is the surface of a two-sphere of radius R centered at the origin.

The following four exercises all explore the same geometric facts relating to Stokes' theorem and the area 2-form of a sphere, but in different physical settings.

Exercise 3.8: A flywheel of moment of inertia I can rotate without friction about an axle whose direction is specified by a unit vector $\mathbf{n}$. The flywheel and axle are initially stationary. The direction $\mathbf{n}$ of the axle is made to describe a simple closed curve $\gamma = \partial\Omega$ on the unit sphere, and is then left stationary.

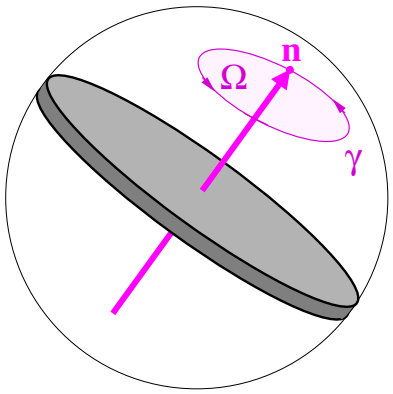

Figure 3.13: *Flywheel*

Show that once the axle has returned to rest in its initial direction, the flywheel has also returned to rest, but has rotated through an angle $\theta = \text{Area}(\Omega)$ when compared with its initial orientation. The area of Ω is to be counted as positive if the path γ surrounds it in a clockwise sense, and negative otherwise. Observe that the path γ bounds two regions with opposite orientations. Taking into account that we cannot define the rotation angle at intermediate steps, show that the area of either region can be used to compute θ, the results being physically indistinguishable. (Hint: Show that the component $L_Z = I(\dot{\psi} + \dot{\phi}\cos\theta)$ of the flywheel's angular momentum along the axle is a constant of the motion.)

Exercise 3.9: A ball of unit radius rolls without slipping on a table. The ball moves in such a way that the point in contact with table describes a closed path $\gamma = \partial\Omega$ on the *ball*. (The corresponding path on the *table* will not necessarily be closed.) Show that the final orientation of the ball will be such that it has rotated, when compared with its initial orientation, through an angle $\phi = \text{Area}(\Omega)$ about a vertical axis through its center, As in the previous problem, the area is counted positive if γ encircles Ω in an anti-clockwise sense. (Hint: recall the no-slip rolling condition $\dot{\phi} + \dot{\psi}\cos\theta = 0$ from (2.26).)

Exercise 3.10: Let a curve in $\mathbb{R}^3$ be parameterized by its arc length s as $\mathbf{r}(s)$. Then the unit tangent to the curve is given by

$$\mathbf{t}(s) = \dot{\mathbf{r}} \equiv \frac{d\mathbf{r}}{ds}.$$

The *principal normal* $\mathbf{n}(s)$ and the *binormal* $\mathbf{b}(s)$ are defined by the requirement that $\dot{\mathbf{t}} = \kappa\mathbf{n}$ with the *curvature* $\kappa(s)$ positive, and that $\mathbf{t}$, $\mathbf{n}$ and $\mathbf{b} = \mathbf{t}\times\mathbf{n}$ form a right-handed orthonormal frame.

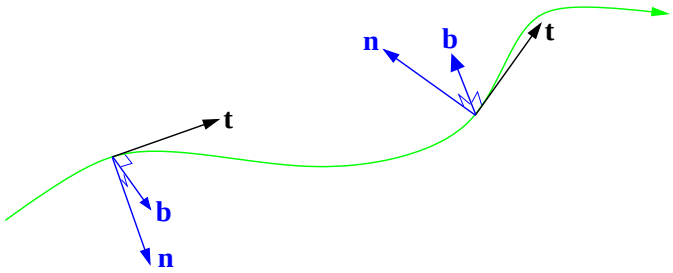

Figure 3.14: *Serret-Frenet frames.*

a) Show that there exists a scalar $\tau(s)$, the *torsion* of the curve, such that $\mathbf{t}$, $\mathbf{n}$ and $\mathbf{b}$ obey the *Serret-Frenet* relations

$$\begin{pmatrix} \dot{\mathbf{t}} \\ \dot{\mathbf{n}} \\ \dot{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}.$$

b) Any pair of mutually orthogonal unit vectors $\mathbf{e}_1(s)$, $\mathbf{e}_2(s)$ perpendicular to $\mathbf{t}$ and such that $\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{t}$ can serve as an orthonormal frame for vectors in the normal plane. A basis pair $\mathbf{e}_1$, $\mathbf{e}_2$ with the property

$$\dot{\mathbf{e}}_1 \cdot \mathbf{e}_2 - \dot{\mathbf{e}}_2 \cdot \mathbf{e}_1 = 0$$

is said to be *parallel*, or *Fermi-Walker*, transported along the curve. In other words, a parallel-transported 3-frame $\mathbf{t}$, $\mathbf{e}_1$, $\mathbf{e}_2$ slides along the curve $\mathbf{r}(s)$ in such a way that the component of its angular velocity in the $\mathbf{t}$ direction is always zero. Show that the Serret-Frenet frame $\mathbf{e}_1 = \mathbf{n}$, $\mathbf{e}_2 = \mathbf{b}$ is *not* parallel transported, but instead rotates at angular velocity $\dot{\theta} = \tau$ with respect to a parallel-transported frame.

c) Consider a finite segment of curve such that the initial and final Serret-Frenet frames are parallel, and so $\mathbf{t}(s)$ defines a closed path $\gamma = \partial\Omega$ on the unit sphere. Fill in the line-by-line justications for the following sequence of manipulations:

$$\begin{aligned}
\int_\gamma \tau\, ds &= \frac{1}{2}\int_\gamma (\mathbf{b}\cdot\dot{\mathbf{n}} - \mathbf{n}\cdot\dot{\mathbf{b}})\, ds \\
&= \frac{1}{2}\int_\gamma (\mathbf{b}\cdot d\mathbf{n} - \mathbf{n}\cdot d\mathbf{b}) \\
&= \frac{1}{2}\int_\Omega (d\mathbf{b}\cdot d\mathbf{n} - d\mathbf{n}\cdot d\mathbf{b}) \qquad (*) \\
&= \frac{1}{2}\int_\Omega \{(d\mathbf{b}\cdot\mathbf{t})(\mathbf{t}\cdot d\mathbf{n}) - (d\mathbf{n}\cdot\mathbf{t})(\mathbf{t}\cdot d\mathbf{b})\} \\
&= \frac{1}{2}\int_\Omega \{(\mathbf{b}\cdot d\mathbf{t})(d\mathbf{t}\cdot\mathbf{n}) - (\mathbf{n}\cdot d\mathbf{t})(d\mathbf{t}\cdot\mathbf{b})\} \\
&= -\frac{1}{2}\int_\Omega \mathbf{t}\cdot(d\mathbf{t}\times d\mathbf{t}) \\
&= -\mathrm{Area}(\Omega).
\end{aligned}$$

(The line marked '$*$' is the one that requires most thought. How can we define "$\mathbf{b}$" and "$\mathbf{n}$" in the interior of Ω?)

d) Conclude that a Fermi-Walker transported frame will have rotated through an angle $\theta = \mathrm{Area}(\Omega)$, compared to its initial orientation, by the time it reaches the end of the curve.

The plane of transversely polarized light propagating in a monomode optical fibre is Fermi-Walker transported, and this rotation can be studied experimentally.[5]

Exercise 3.11: *Foucault's pendulum (in disguise).* A particle of mass m is constrained by a pair of frictionless plates to move in a plane Π that passes through the origin O. The particle is attracted to O by a force $-\kappa\mathbf{r}$, and it therefore executes simple harmonic motion within Π. The orientation of the

[5] A. Tomita, R. Y. Chao, *Phys. Rev. Lett.* **57** (1986) 937-940.

plane, specified by a normal vector $\mathbf{n}$, can be altered in such a way that Π continues to pass through the centre of attraction O.

a) Show that the constrained motion is described by the equation

$$m\ddot{\mathbf{r}} + \kappa\mathbf{r} = \lambda(t)\mathbf{n},$$

and determine $\lambda(t)$ in terms of m, $\mathbf{n}$ and $\ddot{\mathbf{r}}$.

b) Seek a solution in the form

$$\mathbf{r}(t) = \mathbf{A}(t)\cos(\omega t + \phi),$$

and, by assuming that $\mathbf{n}$ changes direction slowly compared to the frequency $\omega = \sqrt{\kappa/m}$, show that $\dot{\mathbf{A}} = -\mathbf{n}(\dot{\mathbf{n}} \cdot \mathbf{A})$. Deduce that $|\mathbf{A}|$ remains constant, and so $\dot{\mathbf{A}} = \boldsymbol{\omega} \times \mathbf{A}$ for some angular velocity vector $\boldsymbol{\omega}$. Show that $\boldsymbol{\omega}$ is perpendicular to $\mathbf{n}$.

c) Show that the results of part b) imply that the direction of oscillation $\mathbf{A}$ is "parallel transported" in the sense of the previous problem. Conclude that if $\mathbf{n}$ slowly describes a closed loop $\gamma = \partial\Omega$ on the unit sphere, then the direction of oscillation $\mathbf{A}$ ends up rotated through an angle $\theta = \text{Area}(\Omega)$.

The next exercise introduces an clever trick for solving some of the non-linear partial differential equations of field theory. The class of equations to which it and its generalizations are applicable is rather restricted, but when they work they provide a complete multi-soliton solution.

Problem 3.12: In this problem you will find the spin field $\mathbf{n}(x)$ that minimizes the energy functional

$$E[\mathbf{n}] = \frac{1}{2}\int_{\mathbb{R}^2} \left(|\nabla n^1|^2 + |\nabla n^2|^2 + |\nabla n^3|^2\right)\, dx^1 dx^2$$

for a given positive winding number N.

a) Use the results of exercise 3.6 to write the winding number N, defined in (3.35), and the energy functional $E[\mathbf{n}]$ as

$$\begin{aligned} 4\pi N &= \int \frac{4}{(1+\xi^2+\eta^2)^2}\left(\partial_1\xi\partial_2\eta - \partial_1\eta\partial_2\xi\right)\, dx^1 dx^2, \\ E[\mathbf{n}] &= \frac{1}{2}\int \frac{4}{(1+\xi^2+\eta^2)^2}\left((\partial_1\xi)^2 + (\partial_2\xi)^2 + (\partial_1\eta)^2 + (\partial_2\eta)^2\right)\, dx^1 dx^2, \end{aligned}$$

where ξ and η are stereographic co-ordinates on S^2 specifying the direction of the unit vector $\mathbf{n}$.

b) Deduce the inequality

$$E - 4\pi N \equiv \frac{1}{2}\int \frac{4}{(1+\xi^2+\eta^2)^2}|(\partial_1 + i\partial_2)(\xi + i\eta)|^2\, dx^1 dx^2 > 0.$$

c) Deduce that for winding number $N > 0$ the minimum energy solutions have energy $E = 4\pi N$ and are obtained by solving the first-order linear equation

$$\left(\frac{\partial}{\partial x^1} + i\frac{\partial}{\partial x^2}\right)(\xi + i\eta) = 0.$$

d) Solve the equation in part c) and show that the minimal energy solutions with winding number $N > 0$ are given by

$$\xi + i\eta = \lambda\frac{(z-a_1)\dots(z-a_N)}{(z-b_1)\dots(z-b_N)}$$

where $z = x^1 + ix^2$, and $\lambda, a_1, \dots, a_N$, and $b_1, \dots, b_N$, are arbitrary complex numbers—except that no a may coincide with any b. This is the solution we displayed at the end of section 3.4.2.

e) Repeat the analysis for $N < 0$. Show that the solutions are given in terms of rational functions of $\bar{z} = x^1 - ix^2$.

The idea of combining the energy functional and the topological charge into a single, manifestly positive, functional is due to Evgueny Bogomol'nyi. The the resulting first order linear equation is therefore called a *Bogomolnyi equation*. If we had tried to find a solution directly in terms of $\mathbf{n}$, we would have ended up with a horribly non-linear second-order partial differential equation..

Exercise 3.13: *Lobachevski space*. The hyperbolic plane of Lobachevski geometry can be realized by embedding the $Z \geq R$ branch of the two-sheeted hyperboloid $Z^2 - X^2 - Y^2 = R^2$ into a Minkowski space with metric $ds^2 = -dZ^2 + dX^2 + dY^2$.

We can parametrize the emebedded surface by making an "imaginary radius" version of the stereographic map, in which the point P on the hyperboloid is labelled by the co-ordinates of the point Q on the X-Y plane (see figure 3.15).

i) Show that the embedding induces the metric

$$g(\ ,\ ) = \frac{4R^4}{(R^2 - X^2 - Y^2)^2}(dX \otimes dX + dY \otimes dY), \quad X^2 + Y^2 < R^2$$

of the Poincaré disc model (see problem **??.??**) on the hyperboloid.

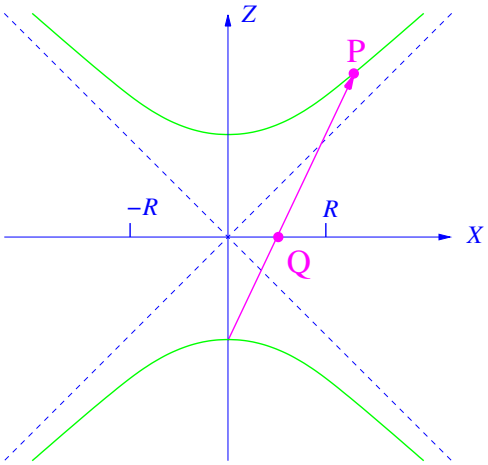

Figure 3.15: *A slice through the embedding of two-dimensional Lobachevski space into three-dimensional Minkowski space, showing the sterographic parameterization of the embedded space by the Poincaré disc* $X^2 + Y^2 < R^2$.

ii) Use the induced metric to show that the area of a disc of hyperbolic radius ρ is given by

$$Area = 4\pi R^2 \sinh^2\left(\frac{\rho}{2R}\right) = 2\pi R^2(\cosh(\rho/R) - 1),$$

and so is only given by $\pi\rho^2$ when ρ is small compared to the scale R of the hyperbolic space. It suffices to consider circles with their centres at the origin. You will first need to show that the hyperbolic distance ρ from the center of the disc to a point at Euclidean distance r is

$$\rho = R \ln\left(\frac{R+r}{R-r}\right).$$

Exercise 3.14: Faraday's "flux rule" for computing the electromotive force $\mathcal{E}$ in a circuit containing a thin moving wire is usually derived by the following manipulations:

$$\begin{aligned}
\mathcal{E} &\equiv \oint_{\partial\Omega} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r} \\
&= \int_\Omega \operatorname{curl} \mathbf{E} \cdot d\mathbf{S} - \oint_{\partial\Omega} \mathbf{B} \cdot (\mathbf{v} \times d\mathbf{r}) \\
&= -\int_\Omega \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} - \oint_{\partial\Omega} \mathbf{B} \cdot (\mathbf{v} \times d\mathbf{r}) \\
&= -\frac{d}{dt} \int_\Omega \mathbf{B} \cdot d\mathbf{S}.
\end{aligned}$$

a) Show that if we parameterize the surface Ω as $x^\mu(u, v, \tau)$, with u, v labelling points on Ω and τ parametrizing the evolution of Ω, then the corresponding manipulations in the covariant differential-form version of Maxwell's equations lead to

$$\frac{d}{d\tau}\int_\Omega F = \int_\Omega \mathcal{L}_V F = \int_{\partial\Omega} i_V F = -\int_{\partial\Omega} f$$

where $V^\mu = \partial x^\mu/\partial\tau$ and $f = -i_V F$.

b) Show that if we take τ to be the proper time along the world-line of each element of Ω, then V is the 4-velocity

$$V^\mu = \frac{1}{\sqrt{1-\mathbf{v}^2}}(1, \mathbf{v}),$$

and $f = -i_V F$ becomes the one-form corresponding to the Lorentz-force 4-vector.

It is not clear that the terms in this covariant form of Farday's law can be given any physical interpretation outside the low-velocity limit. When parts of $\partial\Omega$ have different velocities, the relation of the integrals to measurements made at fixed co-ordinate time requires thought.[6]

The next pair of exercises explores some physics appearances of the continuum Hopf linking number (3.64).

Exercise 3.15: The equations governing the motion of an incompressible inviscid fluid are $\nabla \cdot \mathbf{v} = 0$ and Euler's equation

$$\frac{D\mathbf{v}}{Dt} \equiv \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla P.$$

Recall that the operator $\partial/\partial t + \mathbf{v}\cdot\nabla$, here written as D/Dt, is called the *convective derivative*.

a) Take the curl of Euler's equation to show that if $\boldsymbol{\omega} = \nabla\times\mathbf{v}$ is the *vorticity* then

$$\frac{D\boldsymbol{\omega}}{Dt} \equiv \frac{\partial \boldsymbol{\omega}}{\partial t} + (\mathbf{v}\cdot\nabla)\boldsymbol{\omega} = (\boldsymbol{\omega}\cdot\nabla)\mathbf{v}.$$

b) Combine Euler's equation with part a) to show that

$$\frac{D}{Dt}(\mathbf{v}\cdot\boldsymbol{\omega}) = \nabla\cdot\left\{\boldsymbol{\omega}\left(\frac{1}{2}\mathbf{v}^2 - P\right)\right\}.$$

[6] See E. Marx, *Journal of the Franklin Institute,* **300** (1975) 353-364.

c) Show that if Ω is a volume moving with the fluid, then

$$\frac{d}{dt}\int_\Omega f(\mathbf{r},t)\,dV = \int_\Omega \frac{Df}{Dt}\,dV.$$

e) Conclude that when $\boldsymbol{\omega}$ is zero at infinity the *helicity*

$$I = \int \mathbf{v}\cdot(\nabla\times\mathbf{v})\,dV = \int \mathbf{v}\cdot\boldsymbol{\omega}\,dV$$

is a constant of the motion.

The helicity measures the Hopf linking number of the vortex lines. The discovery[7] of its conservation founded the field of *topological fluid dynamics*.

Exercise 3.16: Let $\mathbf{B} = \nabla\times\mathbf{A}$ and $\mathbf{E} = -\partial\mathbf{A}/\partial t - \nabla\phi$ be the electric and magnetic field in an incompressible and perfectly conducting fluid. In such a fluid the co-moving electromotive force $\mathbf{E}+\mathbf{v}\times\mathbf{B}$ must vanish everywhere.

a) Use Maxwell's equations to show that

$$\begin{aligned}\frac{\partial\mathbf{A}}{\partial t} &= \mathbf{v}\times(\nabla\times\mathbf{A}) - \nabla\phi,\\ \frac{\partial\mathbf{B}}{\partial t} &= \nabla\times(\mathbf{v}\times\mathbf{B}).\end{aligned}$$

b) From part a) show that the convective derivative of $\mathbf{A}\cdot\mathbf{B}$ is given by

$$\frac{D}{Dt}(\mathbf{A}\cdot\mathbf{B}) = \nabla\cdot\{\mathbf{B}(\mathbf{A}\cdot\mathbf{v}-\phi)\}.$$

c) By using the same reasoning as the previous problem, and assuming that $\mathbf{B}$ is zero at infinity, conclude that *Woltjer's invariant*

$$I = \int(\mathbf{A}\cdot\mathbf{B})\,dV = \int \epsilon_{ijk}A_i\partial_j A_k d^3x = \int AF$$

is a constant of the motion.

This result shows that the Hopf linking number of the magnetic field lines is independent of time. It is an essential ingredient in the geodynamo theory of the Earth's magnetic field.

[7] H. K. Moffatt, *J. Fluid Mech.* **35** (1969) 117.

Chapter 4

An Introduction to Topology

Topology is the study of the consequences of continuity. We all know that a continuous real function defined on a connected interval and positive at one point and negative at another must take the value zero at some point between. This fact seems obvious—although a course of real analysis will convince you of the need for a proof. A less obvious fact, but one that follows from the previous one, is that a continuous function defined on the unit circle must posses two diametrically opposite points at which it takes the same value. To see that this is so, consider $f(\theta + \pi) - f(\theta)$. This difference (if not initially zero, in which case there is nothing further to prove) changes sign as θ is advanced through π, because the two terms exchange roles. It was therefore zero somewhere. This observation has practical application in daily life: Our local coffee shop contains four-legged tables that wobble because the floor is not level. They are round tables, however, and because they possess no misguided levelling screws all four legs have the same length. We are therefore guaranteed that by rotating the table about its center through an angle of less than $\pi/2$ we will find a stable location. A ninety-degree rotation interchanges the pair of legs that are both on the ground with the pair that are rocking, and at the change-over point all four legs must be simultaneously on the ground.

Similar effects with a practical significance for physics appear when we try to extend our vector and tensor calculus from a local region to an entire manifold. A smooth field of vectors tangent to the sphere S^2 will always possess a zero — *i.e.* a point at which the the vector field vanishes. On the torus T^2, however, we can construct a nowhere-zero vector field. This shows that the global topology of the manifold influences the way in which

the tangent spaces are glued together to form the tangent bundle. To study this influence in a systematic manner we need first to understand how to characterize the global structure of a manifold, and then to see how this structure affects the mathematical and physical objects that live on it.

4.1 Homeomorphism and Diffeomorphism

In the previous chapter we met with a number of *topological invariants*, quantities that are unaffected by continuous deformations. Some invariants help to distinguish topologically distinct manifolds. An important example is the set of *Betti numbers* of the manifold. If two manifolds have different Betti numbers they are certainly distinct. If, however, they have the same Betti numbers, we cannot be sure that they are topologically identical. It is a holy grail of topology to find a complete set of invariants such that having them all coincide would be enough to say that two manifolds were topologically the same.

In the previous paragraph we were deliberately vague in our use of the terms "distinct" and the "same". Two topological spaces (spaces equipped with a definition of what is to be considered an open set) are regarded as being the "same", or *homeomorphic*, if there is a one-to-one, onto, continuous map between them whose inverse is also continuous. Manifolds come with the additional structure of differentiability: we may therefore talk of "smooth" maps, meaning that their expression in coordinates is infinitely (C^∞) differentiable. We regard two manifolds as being the "same", or *diffeomorphic*, if there is a one-to-one onto C^∞ map between them whose inverse is also C^∞. The distinction between homeomorphism and diffeomorphism sounds like a mere technical nicety, but it has consequences for physics. Edward Witten discovered[1] that there are 992 distinct 11-spheres. These are manifolds that are all homeomorphic to the 11-sphere, but diffeomorphically inequivalent. This fact is crucial for the cancellation of global graviational anomalies in the $\mathrm{E}_8 \times \mathrm{E}_8$ or SO(32) symmetric superstring theories.

Since we are interested in the consequences of topology for calculus, we will restrict ourselves to the interpretation "same" = diffeomorphic.

[1] E. Witten, *Comm. Math. Phys.* **117** (1986), 197.

4.2 Cohomology

Betti numbers arise in answer to what seems like a simple calculus problem: when can a vector field whose divergence vanishes be written as the curl of something? We will see that the answer depends on the global structure of the space the field inhabits.

4.2.1 Retractable Spaces: Converse of Poincaré Lemma

Poincaré's lemma asserts that $d^2 = 0$. In traditional vector calculus language this reduces to the statements $\operatorname{curl}(\operatorname{grad}\phi) = 0$ and $\operatorname{div}(\operatorname{curl}\mathbf{w}) = 0$. We often assume that the converse is true: If $\operatorname{curl}\mathbf{v} = 0$, we expect that we can find a ϕ such that $\mathbf{v} = \operatorname{grad}\phi$, and, if $\operatorname{div}\mathbf{v} = 0$, that we can find a $\mathbf{w}$ such that $\mathbf{v} = \operatorname{curl}\mathbf{w}$. You know a formula for the first case:

$$\phi(x) = \int_{x_0}^{x} \mathbf{v} \cdot d\mathbf{x}, \tag{4.1}$$

but probably do not know the corresponding formula for $\mathbf{w}$. Using differential forms, and provided the space in which these forms live has suitable *topological* properties, it is straightforward to find a solution for the general problem: If ω is closed, meaning that $d\omega = 0$, find χ such that $\omega = d\chi$.

The "suitable topological properties" referred to in the previous paragraph is that the space be *retractable*. Suppose that the closed form ω is defined in a domain Ω. We say that Ω is retractable to the point O if there exists a smooth map $\varphi_t : \Omega \to \Omega$ which depends continuously on a parameter $t \in [0, 1]$ and for which $\varphi_1(x) = x$ and $\varphi_0(x) = \mathrm{O}$. Applying this retraction map to the form, we will then have $\varphi_1^*\omega = \omega$ and $\varphi_0^*\omega = 0$. Let us set $\varphi_t(x^\mu) = x^\mu(t)$. Define $\eta(x, t)$ to be the velocity-vector field that corresponds to the co-ordinate flow:

$$\frac{dx^\mu}{dt} = \eta^\mu(x, t). \tag{4.2}$$

An easy exercise, using the interpretation of the Lie derivative in (2.40), shows that

$$\frac{d}{dt}(\varphi_t^*\omega) = \mathcal{L}_\eta(\varphi_t^*\omega). \tag{4.3}$$

We now use the infinitesimal homotopy relation and our assumption that $d\omega = 0$, and hence (from exercise 3.3) that $d(\varphi_t^*\omega) = 0$, to write

$$\mathcal{L}_\eta(\varphi_t^*\omega) = (i_\eta d + d i_\eta)(\varphi_t^*\omega) = d[i_\eta(\varphi_t^*\omega)]. \tag{4.4}$$

Using this we can integrate up with respect to t to find

$$\omega = \varphi_1^*\omega - \varphi_0^*\omega = d\left(\int_0^1 i_\eta(\varphi_t^*\omega)dt\right). \tag{4.5}$$

Thus

$$\chi = \int_0^1 i_\eta(\varphi_t^*\omega)dt, \tag{4.6}$$

solves our problem.

This magic formula for χ makes use of the nearly all the "calculus on manifolds" concepts that we have introduced so far. The notation is so powerful that it has suppressed nearly everything that a traditionally-educated physicist would find familiar. We will therefore unpack the symbols by means of a concrete example. Let us take Ω to be the whole of $\mathbb{R}^3$. This can be retracted to the origin via the map $\varphi_t(x^\mu) = x^\mu(t) = tx^\mu$. The velocity field whose flow gives

$$x^\mu(t) = t\, x^\mu(0)$$

is $\eta^\mu(x,t) = x^\mu/t$. To verify this, compute

$$\frac{dx^\mu(t)}{dt} = x^\mu(0) = \frac{1}{t}x^\mu(t),$$

so $x^\mu(t)$ is indeed the solution to

$$\frac{dx^\mu}{dt} = \eta^\mu(x(t),t).$$

Now let us apply this retraction to $\omega = Adydz + Bdzdx + Cdxdy$ with

$$d\omega = \left(\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z}\right)dxdydz = 0. \tag{4.7}$$

The pull-back φ_t^* gives

$$\varphi_t^*\omega = A(tx,ty,tz)d(ty)d(tz) + (\text{two similar terms}). \tag{4.8}$$

The interior product with

$$\eta = \frac{1}{t}\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + z\frac{\partial}{\partial z}\right) \tag{4.9}$$

then gives

$$i_\eta \varphi_t^* \omega = tA(tx, ty, tz)(y\,dz - z\,dy) + (\text{two similar terms}). \tag{4.10}$$

Finally we form the ordinary integral over t to get

$$\begin{aligned}
\chi &= \int_0^1 i_\eta(\varphi_t^*\omega)dt \\
&= \left[\int_0^1 A(tx, ty, tz)t\,dt\right](ydz - zdy) \\
&\quad + \left[\int_0^1 B(tx, ty, tz)t\,dt\right](zdx - xdz) \\
&\quad + \left[\int_0^1 C(tx, ty, tz)t\,dt\right](xdy - ydx).
\end{aligned} \tag{4.11}$$

In this expression the integrals in the square brackets are just numerical coefficients, *i.e.*, the "dt" is not part of the one-form. It is instructive, because not entirely trivial, to let "d" act on χ and verify that the construction works. If we focus first on the term involving A, we find that $d[\int_0^1 A(tx, ty, tz)t\,dt](ydz - zdy)$ can be grouped as

$$\begin{aligned}
&\left[\int_0^1 \left\{2tA + t^2\left(x\frac{\partial A}{\partial x} + y\frac{\partial A}{\partial y} + z\frac{\partial A}{\partial z}\right)\right\} dt\right] dydz \\
&\qquad - \int_0^1 t^2 \frac{\partial A}{\partial x}\,dt\,(xdydz + ydzdx + zdxdy).
\end{aligned} \tag{4.12}$$

The first of these terms is equal to

$$\left[\int_0^1 \frac{d}{dt}\left\{t^2 A(tx, ty, tz)\right\}\,dt\right] dydz = A(x, y, x)\,dydz, \tag{4.13}$$

which is part of ω. The second term will combine with the terms involving B, C, to become

$$-\int_0^1 t^2\left(\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z}\right) dt\,(xdydz + ydzdx + zdxdy), \tag{4.14}$$

which is zero by our hypothesis. Putting togther the A, B, C, terms does therefore reconstitute ω.

4.2.2 Obstructions to Exactness

The condition that Ω be retractable plays an essential role in the converse to Poincaré's lemma. In its absence $d\omega = 0$ does not guarantee that there is an χ such that $\omega = d\chi$. Consider, for example, a vector field $\mathbf{v}$ with $\operatorname{curl}\mathbf{v} \equiv 0$ in an annulus $\Omega = \{R_0 < |\mathbf{r}| < R_1\}$. In the annulus (a non-retractable space) the condition that $\operatorname{curl}\mathbf{v} \equiv 0$ does not prohibit $\oint_\Gamma \mathbf{v}\cdot d\mathbf{r}$ being non zero for some closed path Γ encircling the central hole. When this line integral is non-zero then there can be no single-valued χ such that $\mathbf{v} = \nabla\chi$. If there were such a χ, then

$$\oint_\Gamma \mathbf{v}\cdot d\mathbf{r} = \chi(0) - \chi(0) = 0. \tag{4.15}$$

A non-zero value for $\oint_\Gamma \mathbf{v}\cdot d\mathbf{r}$ therefore consititutes an *obstruction* to the existence of an ϕ such that $\mathbf{v} = \nabla\chi$.
Example: The sphere S^2 is not retractable. The area 2-form $\sin\theta d\theta d\phi$ is closed, but, although we can write

$$\sin\theta d\theta d\phi = d[(1-\cos\theta)d\phi], \tag{4.16}$$

the 1-form $(1-\cos\theta)d\phi$ is singular at the south pole, $\theta = \pi$. We could try

$$\sin\theta d\theta d\phi = d[(-1-\cos\theta)d\phi], \tag{4.17}$$

but this is singular at the north pole, $\theta = 0$. There is no escape: we know that

$$\int_{S^2} \sin\theta d\theta d\phi = 4\pi, \tag{4.18}$$

but if $\sin\theta d\theta d\phi = d\chi$ then Stokes says that

$$\int_{S^2} \sin\theta d\theta d\phi \stackrel{?}{=} \int_{\partial S^2} \chi = 0 \tag{4.19}$$

because $\partial S^2 = 0$. Again, a non-zero value for $\int\omega$ over some boundary-less region has provided an obstruction to finding an χ such that $\omega = d\chi$.

4.2.3 De Rham Cohomology

We have seen that sometimes the condition $d\omega = 0$ allows us to find an χ such that $\omega = d\chi$, and sometimes it does not. If the region in which we seek χ is

retractable, we can always construct it. If the region is not retractable there may be an obstruction to the existence of χ. In order to describe the various possibilities we introduce the language of *cohomology*, or more precisely *de Rham cohomology*, named for the Swiss mathematician Georges de Rham who did the most to create it.

The significance of cohomology for physics is that many important quantities can be expressed as integrals of differential forms that lie in some cohomology space.

For simplicity suppose that we are working in a compact manifold M without boundary. Let $\Omega^p(M) = \bigwedge^p(T^*M)$ be the space of all smooth p-form fields. It is a vector space over $\mathbb{R}$: we can add p-form fields and multiply them by real constants, but, as is the vector space $C^\infty(M)$ of smooth functions on M, it is infinite dimensional. The subspace $Z^p(M)$ of *closed* forms—those with $d\omega = 0$—is also an infinite dimensional vector space, and the same is true of the space $B^p(M)$ of *exact* forms — those that can be written as $\omega = d\chi$ for some globally defined $(p-1)$-form χ. Now consider the space $H^p = Z^p/B^p$, which is the space of closed forms *modulo* exact forms. In this space we do not distinguish between two forms, ω_1 and ω_2 when there an χ, such that $\omega_1 = \omega_2 + d\chi$. We say that ω_1 and ω_2 are *cohomologous*, and write $\omega_1 \sim \omega_2 \in H^p(M)$. We will use the symbol $[\omega]$ to denote the equivalence class of forms cohomologous to ω. Now a miracle happens! For a compact manifold M the space $H^p(M)$ is *finite* dimensional! It is called the p-th (de Rham) cohomology space of the manifold, and depends only on the global topology of M. In particular, it does not depend on any metric we may have chosen for M.

Sometimes we write $H^p_{\rm DR}(M, \mathbb{R})$ to make clear that we are dealing with de Rham cohomolgy, and that we are working with vector spaces over the real numbers. This is because there is also a space $H^p_{\rm DR}(M, \mathbb{Z})$, where we only allow multiplication by integers.

The cohomology space $H^p_{\rm DR}(M, \mathbb{R})$ codifies all potential obstructions to solving the problem of finding a $(p-1)$-form χ such that $d\chi = \omega$: we can find such a χ if and only if ω is cohomologous to zero in $H^p_{\rm DR}(M, \mathbb{R})$. If $H^p_{\rm DR}(M, \mathbb{R}) = \{0\}$, which is the case if M is retractable, then all closed p-forms are cohomologous to zero. If $H^p_{\rm DR}(M, \mathbb{R}) \neq \{0\}$, then some closed p-forms ω will not be cohomologous to zero. We can test whether $\omega \sim 0 \in H^p_{\rm DR}(M, \mathbb{R})$ by forming suitable integrals.

4.3 Homology

The language of cohomology seems rather abstract. To understand its origin it may be more intuitive to think about the spaces that are the cohomology spaces' vector-space duals. These *homology* spaces are simple to understand pictorially.

The basic idea is that, given a region Ω, we can find its boundary $\partial\Omega$. Inspection of a few simple cases will soon lead to the conclusion that the "boundary of a boundary" consists of nothing. In symbols, $\partial^2 = 0$. The statement "$\partial^2 = 0$" is clearly analgous to "$d^2 = 0$," and, pursuing the analogy, we can construct a vector space of "regions" and define two "regions" as being *homologous* if they differ by the boundary of another "region."

4.3.1 Chains, Cycles and Boundaries

We begin by making precise the vague notions of region and boundary.

Simplicial Complexes

The set of all curves and surfaces in a manifold M is infinite dimensional, but the homology spaces are finite dimensional. Life would be much easier if we could use finite dimensional spaces throughout. Mathematicians therefore do what any computationally-minded physicist would do: they approximate the smooth manifold by a discrete polygonal grid . Were they interested in distances, they would necessarily use many small polygons so as to obtain a good approximation to the detailed shape of the manifold. The global topology, though, can often be captured by a rather coarse discretization. The result of this process is to reduce a complicated problem in differential geometry to one of simple algebra. The resulting theory is therefore known as *algebraic topology.*

It turns out to be convenient to approximate the manifold by generalized triangles. We therefore dissect M into line segments (if one dimensional), triangles, (if two dimensional), tetrahedra (if three dimensional) or higher dimensional *p-simplices* (singular: *simplex*). The rules for the dissection are:

a) Every point must belong to at least one simplex.
b) A point can belong to only a finite number of simplices.
c) Two different simplices either have no points in common, or
    i) one is a face (or edge, or vertex) of the other,

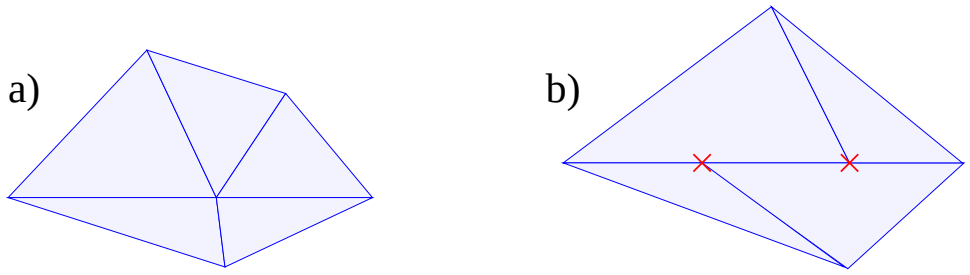

Figure 4.1: *Triangles, or 2-simplices, that are a) allowed, b) not allowed in a dissection. In b) only parts of edges are in common.*

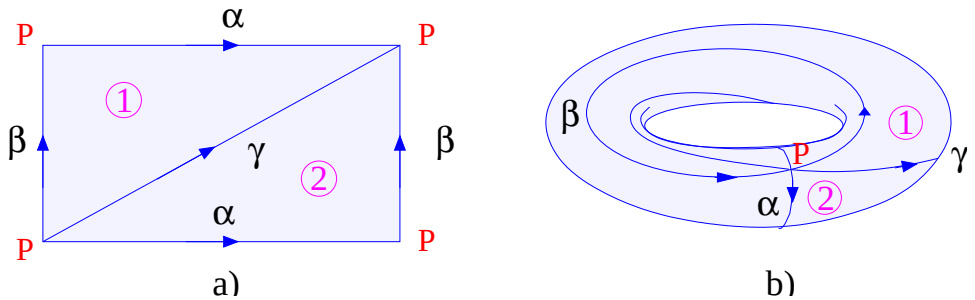

Figure 4.2: *A triangulation of the 2-torus. a) The torus as a rectangle with periodic boundary conditions: The two edges labled α will be glued togther point-by-point along the arrows when we reassemble the torus, and so are to be regarded as a single edge. The two sides labeled β will be glued similarly. b) The assembled torus: All four P's are now in the same place, and correspond to a single point.*

ii) the set of points in common is the whole of a shared face (or edge, or vertex).

The collection of simplices composing the dissected space is called a *simplicial complex*. We will denote it by S.

We may not need many triangles to capture the global topology. For example, figure 4.2 shows how a two-dimensional torus can be decomposed into two 2-simplices (triangles) bounded by three 1-simplices (edges) α, β, γ, and with only a single 0-simplex (vertex) P. Computations are easier to describe, however, if each simplex in the decomposition is uniquely specified by its vertices. For this we usually need a slightly finer dissection. Figure 4.3 shows a decomposition of the torus into 18 triangles each of which is

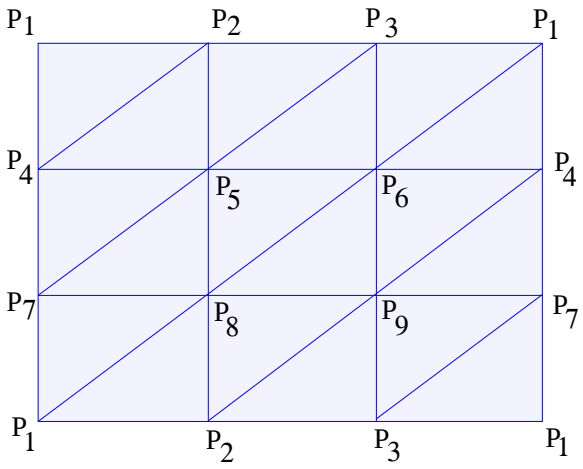

Figure 4.3: *A second triangulation of the 2-torus.*

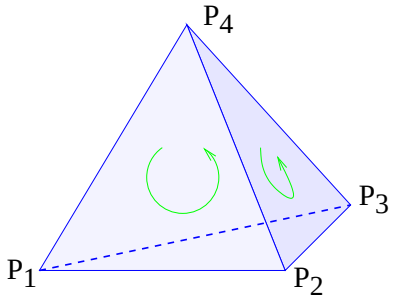

Figure 4.4: *A tetrahedral triangulation of the 2-sphere. The circulating arrows on the faces indicate the choice of orientation* $P_1P_2P_4$ *and* $P_2P_3P_4$.

uniquely labeled by three points drawn from a set of nine vertices. In this figure vertices with identical labels are to be regarded as the same vertex, as are the corresponding sides of triangles. Thus, each of the edges P_1P_2, P_2P_3, P_3P_1, at the top of the figure are to be glued point-by-point to the corresponding edges on bottom of the figure. Similarly along the sides. The resulting simplicial complex then has 27 edges.

We may triangulate the sphere S^2 as a tetrahedron with vertices P_1, P_2, P_3, P_4. This dissection has six edges: P_1P_2, P_1P_3, P_1P_4, P_2P_3, P_2P_4, P_3P_4, and four faces: $P_2P_3P_4$, $P_1P_3P_4$, $P_1P_2P_4$ and $P_1P_2P_3$.

p-Chains

We assign to simplices an orientation defined by the order in which we write their defining vertices. The interchange of of any pair of vertices reverses the orientation, and we consider there to be a relative minus sign between oppositely oriented but otherwise identical simplices: $P_2P_1P_3P_4 = -P_1P_2P_3P_4$.

We now construct abstract vector spaces $C_p(S, \mathbb{R})$ of *p-chains* which have the oriented p-simplices as their basis vectors. The most general elements of $C_2(S, \mathbb{R})$, with S being the tetrahedral triangulation of the sphere S^2, would be

$$a_1P_2P_3P_4 + a_2P_1P_3P_4 + a_3P_1P_2P_4 + a_4P_1P_2P_3, \tag{4.20}$$

where $a_1, \ldots, a_4$, are real numbers. We regard the distinct faces as being linearly independent basis elements for $C_2(S, \mathbb{R})$. The space is therefore four dimensional. If we had triangulated the sphere so that it had 16 triangular faces, the space C_2 would be 16 dimensional.

Similarly, the general element of $C_1(S, \mathbb{R})$ would be

$$b_1P_1P_2 + b_2P_1P_3 + b_3P_1P_4 + b_4P_2P_3 + b_5P_2P_4 + b_6P_3P_4, \tag{4.21}$$

and so $C_1(S, \mathbb{R})$ is a six-dimensional space spanned by the *edges* of the tetrahedron. For $C_0(S, \mathbb{R})$ we have

$$c_1P_1 + c_2P_2 + c_3P_3 + c_4P_4, \tag{4.22}$$

and so $C_0(S, \mathbb{R})$ is four dimensional, and spanned by the *vertices*.

Our manifold comprises only the *surface* of the two-sphere, so there is no such thing as $C_3(S, \mathbb{R})$.

The reason for making the field $\mathbb{R}$ explicit in these definitions is that we sometimes gain more information about the topology if we allow only integer coefficients. The space of such p-chains is then denoted by $C_p(S, \mathbb{Z})$. Because a vector space requires that coefficients be drawn from a field, these objects are no longer vector spaces. They can be thought of as either *modules*—"vector spaces" whose coefficient are drawn from a ring—or as additive abelian groups.

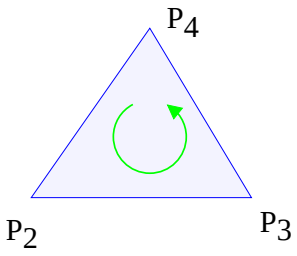

Figure 4.5: *The oriented triangle $P_2P_3P_4$ has boundary $P_3P_4 + P_4P_2 + P_2P_3$.*

The Boundary Operator

We now introduce a linear map $\partial_p : C_p \to C_{p-1}$, called the *boundary operator*. Its action on a p-simplex is

$$\partial_p P_{i_1} P_{i_2} \cdots P_{i_{p+1}} = \sum_{j=1}^{p+1} (-1)^{j+1} P_{i_1} \ldots \widehat{P}_{i_j} \ldots P_{i_{p+1}}, \tag{4.23}$$

where the "hat" indicates that P_{i_j} is to be omitted. The resulting $(p-1)$-chain is called the *boundary* of the simplex. For example

$$\begin{aligned} \partial_2(P_2P_3P_4) &= P_3P_4 - P_2P_4 + P_2P_3, \\ &= P_3P_4 + P_4P_2 + P_2P_3. \end{aligned} \tag{4.24}$$

The boundary of a line segment is the difference of its endpoints

$$\partial_1(P_1P_2) = P_2 - P_1. \tag{4.25}$$

Finally, for any point,

$$\partial P_i = 0. \tag{4.26}$$

Because ∂ is defined to be a linear map, when it is applied to a p-chain $c = a_1s_1 + a_2s_2 + \cdots + a_ns_n$, where the s_i are p-simplices, we have $\partial_p c = a_1\partial_p s_1 + a_2\partial_p s_2 + \cdots + a_n\partial_p s_n$.
When we take the "∂" of a chain of compatibly oriented simplices that together make up some region, the internal boundaries cancel in pairs, and the "boundary" of the chain really is the oriented geometric boundary of the region. For example in figure 4.6 we find that

$$\partial(P_1P_5P_2 + P_2P_5P_4 + P_3P_4P_5 + P_1P_3P_5) = P_1P_3 + P_3P_4 + P_4P_2 + P_2P_1, \tag{4.27}$$

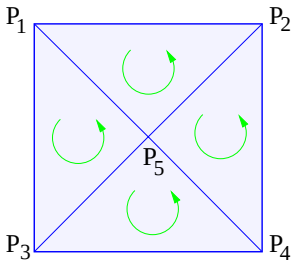

Figure 4.6: *Compatibly oriented simplices.*

which is the counter-clockwise directed boundary of the square.

For each of the examples we find that $\partial_{p-1}\partial_p\, s = 0$. From the definition (4.23) we can easily establish that this identity holds for any p-simplex s. As chains are sums of simplices and ∂_p is linear, it remains true for any $c \in C_p$. Thus $\partial_{p-1}\partial_p = 0$. We will usually abbreviate this statement as $\partial^2 = 0$.

Cycles, Boundaries and Homology

A *chain complex* is a doubly infinite sequence of spaces (these can be vector spaces, modules, abelian groups, or many other mathematical objects) such as $\ldots, C_{-2}, C_{-1}, C_0, C_1, C_2 \ldots$, together with structure-preserving maps

$$\ldots \stackrel{\partial_{p+1}}{\to} C_p \stackrel{\partial_p}{\to} C_{p-1} \stackrel{\partial_{p-1}}{\to} C_{p-2} \stackrel{\partial_{p-1}}{\to} \ldots, \tag{4.28}$$

with the property that $\partial_{p-1}\partial_p = 0$. The finite sequence of C_p's we constructed from our simplicial complex is an example of a chain complex where C_p is zero-dimensional for $p < 0$ or $p > d$. Chain complexes are a useful tool in mathematics, and the ideas we explain in this section have many applications.

Given any chain complex we can define two important linear subspaces of each of the C_p's. The first is the space Z_p of *p-cycles*. This consists of those $z \in C_p$ such that $\partial_p z = 0$. The second is the space B_p of *p-boundaries*, and consists of those $b \in C_p$ such that $b = \partial_{p+1} c$ for some $c \in C_{p+1}$. Because $\partial^2 = 0$, the boundaries B_p constitute a subspace of Z_p. From these spaces we form the quotient space $H_p = Z_p/B_p$, consisting of *equivalence classes* of p-cycles, where we deem z_1 and z_2 to be equivalent, or *homologous*, if they differ by a boundary: $z_2 = z_1 + \partial c$. We will write the equivalence class of cycles homologous z_i to as $[z_i]$. The space H_p, or more accurately, $H_p(\mathbb{R})$, is called the p-th (simplicial) *homology space* of the chain complex. It becomes the p-th homology *group* if $\mathbb{R}$ is replaced by the integers.

We can construct these homology spaces for any chain complex. When the chain complex is derived from a simplicial complex decomposition of a manifold M a remarkable thing happens. The spaces C_p, Z_p, and B_p, all depend on the details of how the manifold M has been dissected to form the simplicial complex S. The homology space H_p, however, is independent the dissection. This is neither obvious nor easy to prove. We will rely on examples to make it plausible. Granted this independence, we will write $H_p(M)$, or $H_p(M, \mathbb{R})$, so as to make it clear that H_p is a property of M. The dimension b_p of $H_p(M)$ is called the p-th *Betti number* of the manifold:

$$b_p \stackrel{\text{def}}{=} \dim H_p(M). \tag{4.29}$$

Example: The Two-Sphere. For the tetrahedral dissection of the two-sphere, any vertex is P_i homologous to any other, as $P_i - P_j = \partial(P_j P_i)$ and all $P_j P_i$ belong to C_2. Furthermore, $\partial P_i = 0$, so $H_0(S^2)$ is one dimensional. In general, the dimension of $H_0(M)$ is the number of disconnected pieces making up M. We will write $H_0(S^2) = \mathbb{R}$, regarding $\mathbb{R}$ as the archetype of a one-dimensional vector space.

Now let us consider $H_1(S^2)$. We first find the space of 1-cycles Z_1. An element of C_1 will be in Z_1 only if each vertex that is the begining of an edge is also the end of an edge, and that these edges have the same coefficient. Thus

$$z_1 = P_2P_3 + P_3P_4 + P_4P_2$$

is a cycle, as is

$$z_2 = P_1P_4 + P_4P_2 + P_2P_1.$$

These are both boundaries of faces of the tetrahedron. It should be fairly easy to convince yourself that Z_1 is the space of linear combinations of these together with boundaries of the other faces

$$\begin{aligned} z_3 &= P_1P_4 + P_4P_3 + P_3P_1, \\ z_4 &= P_1P_3 + P_3P_2 + P_2P_1. \end{aligned}$$

Any three of these are linearly independent, and so Z_1 is three dimensional. Because all of the cycles are boundaries, every element of Z_1 is homologous to $\mathbf{0}$, and so $H_1(S^2) = \{\mathbf{0}\}$.

We also see that $H_2(S^2) = \mathbb{R}$. Here the basis element is

$$P_2P_3P_4 - P_1P_3P_4 + P_1P_2P_4 - P_1P_2P_3 \tag{4.30}$$

which is the 2-chain corresponding to the entire surface of the sphere. It would be the boundary of the solid tedrahedron, but does not count as a boundary as the interior of the tetrahedron is not part of the simplicial complex.

Example: The Torus. Consider the 2-torus T^2.We will see that $H_0(T^2) = \mathbb{R}$, $H_1(T^2) = \mathbb{R}^2 \equiv \mathbb{R} \oplus \mathbb{R}$, and $H_2(T^2) = \mathbb{R}$. A natural basis for the two-dimensional $H_1(T^2)$ consists of the 1-cycles α, β portrayed in figure 4.7.

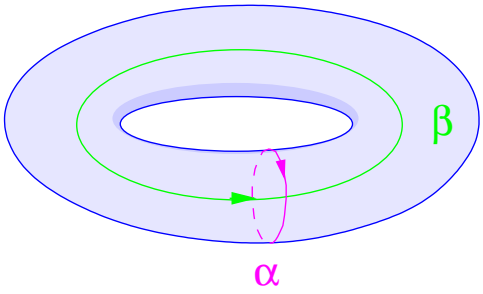

Figure 4.7: *A basis of 1-cycles on the 2-torus.*

The cycle γ that, in figure 4.2, winds once around the torus is homologous to $\alpha + \beta$. In terms of the second triangulation of the torus (figure 4.3) we would have

$$\begin{aligned} \alpha &= P_1P_2 + P_2P_3 + P_3P_1 \\ \beta &= P_1P_7 + P_7P_4 + P_4P_1 \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} \gamma &= P_1P_8 + P_8P_6 + P_6P_1 \\ &= \alpha + \beta + \partial(P_1P_8P_2 + P_8P_9P_2 + P_2P_9P_3 + \cdots). \end{aligned} \tag{4.32}$$

Example: The Projective Plane. The projective plane $\mathbb{R}P^2$ can be regarded as a rectangle with diametrically opposite points identified. Suppose we decompose $\mathbb{R}P^2$ into eight triangles, as in figure 4.8.

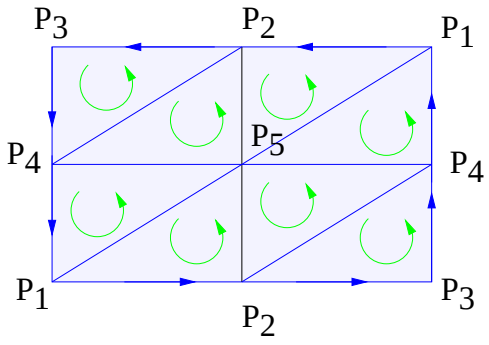

Figure 4.8: *A triangulation of the projective plane.*

Consider the "entire surface"

$$\sigma = P_1P_2P_5 + P_1P_5P_4 + \cdots \in C_2(\mathbb{R}P^2), \tag{4.33}$$

consisting of the sum of all eight 2-simplices with the orientation indicated in the figure. Let $\alpha = P_1P_2 + P_2P_3$ and $\beta = P_1P_4 + P_4P_3$ be the sides of the rectangle running along the bottom horizontal and left vertical sides of the figure, respectively. In each case they run from P_1 to P_3. Then

$$\begin{aligned} \partial(\sigma) &= P_1P_2 + P_2P_3 + P_3P_4 + P_4P_1 + P_1P_2 + P_2P_3 + P_3P_4 + P_1P_2 \\ &= 2(\alpha - \beta) \neq 0. \end{aligned} \tag{4.34}$$

Although $\mathbb{R}P^2$ has no actual edge that we can fall off, from the homological viewpoint it does have a boundary! This represents the conflict between local orientation of each of the 2-simplices and the global non-orientability of $\mathbb{R}P^2$. The surface σ of $\mathbb{R}P^2$ is not a two-cycle, therefore. Indeed $Z_2(\mathbb{R}P^2)$, and *a fortiori* $H_2(\mathbb{R}P^2)$, contain only the zero vector. The only one-cycle is $\alpha - \beta$ which runs from P_1 to P_1 via P_2, P_3 and P_4, but (4.34) shows that this is the boundary of $\frac{1}{2}\sigma$. Thus $H_2(\mathbb{R}P^2, \mathbb{R}) = \{\mathbf{0}\}$ and $H_1(\mathbb{R}P^2, \mathbb{R}) = \{\mathbf{0}\}$, while $H_0(\mathbb{R}P^2, \mathbb{R}) = \mathbb{R}$.

We can now see the advantage of restricting ourselves to integer coefficients. When we are not allowed fractions, the cycle $\gamma = (\alpha - \beta)$ is no longer a boundary, although $2(\alpha - \beta)$ is the boundary of σ. Thus, using the symbol $\mathbb{Z}_2$ to denote the additive group of the integers *modulo* two, we can write $H_1(\mathbb{R}P^2, \mathbb{Z}) = \mathbb{Z}_2$. This homology space is a set with only two members $\{0\gamma, 1\gamma\}$. The finite group $H_1(\mathbb{R}P^2, \mathbb{Z}) = \mathbb{Z}_2$ is said to be the *torsion* part of the homology — a confusing terminology because this torsion has nothing to do with the torsion tensor of Riemannian geometry.

We introduced real-number homology first, because the theory of vector spaces is simpler than that of modules, and more familiar to physicists. The torsion is, however, invisible to the real-number homology. We were therefore buying a simplification at the expense of throwing away information.

The Euler Character

The sum

$$\chi(M) \stackrel{\text{def}}{=} \sum_{p=0}^{d} (-1)^p \dim H_p(M, \mathbb{R}) \tag{4.35}$$

is called the *Euler character* of the manifold M. For example, the 2-sphere has $\chi(S^2) = 2$, the projective plane has $\chi(\mathbb{R}P^2) = 1$, and the n-torus has $\chi(T^n) = 0$. This number is manifestly a topological invariant because the individual $\dim H_p(M, \mathbb{R})$ are. We will show that that the Euler character is also equal to $V - E + F - \cdots$ where V is the number of vertices, E is the number of edges and F is the number of faces in the simplicial dissection. The dots are for higher dimensional spaces, where the alternating sum continues with $(-1)^p$ times the number of p-simplices. In other words, we are claiming that

$$\chi(M) = \sum_{p=0}^{d} (-1)^p \dim C_p(M). \tag{4.36}$$

It is not so obvious that this new sum is a topological invariant. The individual dimensions of the spaces of p-chains depend on the details of how we dissect M into simplices. If our claim is to be correct, the dependence must somehow drop out when we take the alternating sum.

A useful tool for working with alternating sums of vector-space dimensions is provided by the notion of an *exact sequence*. We say that a set of vector spaces V_p with maps $f_p : V_p \to V_{p+1}$ is an exact sequence if $\text{Ker}\,(f_p) = \text{Im}\,(f_{p-1})$. For example, if all cycles were boundaries then the set of spaces C_p with the maps ∂_p taking us from C_p to C_{p-1} would consitute an exact sequence—albeit with p decreasing rather than increasing, but this is irrelevent. When the homology is non-zero, however, we only have $\text{Im}\,(f_{p-1}) \subset \text{Ker}\,(f_p)$, and the number $\dim H_p = \dim(\text{Ker}\, f_p) - \dim(\text{Im}\, f_{p-1})$ provides a measure of how far this set inclusion falls short of being an equality.

Suppose that

$$\{\mathbf{0}\} \xrightarrow{f_0} V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} \dots \xrightarrow{f_{n-1}} V_n \xrightarrow{f_n} \{\mathbf{0}\} \tag{4.37}$$

is a finite-length exact sequence. Here, $\{\mathbf{0}\}$ is the vector space containing only the zero vector. Being linear, f_0 maps $\mathbf{0}$ to $\mathbf{0}$. Also f_n maps everything in V_n to $\mathbf{0}$. Since this last map takes everything to zero, and what is mapped to zero is the image of the penultimate map, we have $V_n = \text{Im}\, f_{n-1}$. Similarly, the fact that $\text{Ker}\, f_1 = \text{Im}\, f_0 = \{\mathbf{0}\}$ shows that $\text{Im}\, f_1 \subseteq V_2$ is an isomorphic image of V_1. This situation is represented pictorially in figure 4.9.

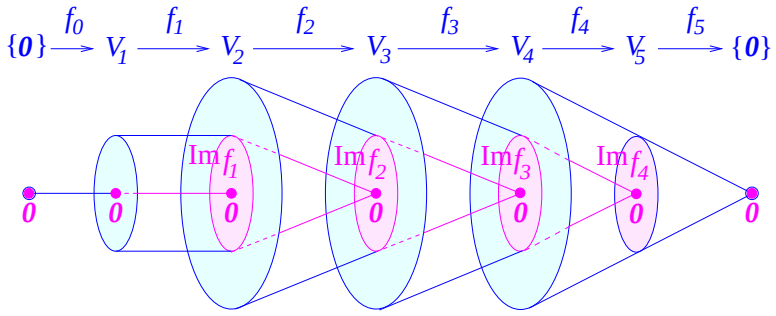

Figure 4.9: *A schematic representation of an exact sequence.*

Now the range-nullspace theorem tells us that

$$\begin{aligned} \dim V_p &= \dim(\text{Im}\, f_p) + \dim(\text{Ker}\, f_p) \\ &= \dim(\text{Im}\, f_p) + \dim(\text{Im}\, f_{p-1}). \end{aligned} \tag{4.38}$$

When we take the alternating sum of the dimensions, and use $\dim(\text{Im}\, f_0) = 0$ and $\dim(\text{Im}\, f_n) = 0$, we find that the sum telescopes to give

$$\sum_{p=0}^{n} (-1)^p \dim V_p = 0. \tag{4.39}$$

The vanishing of this alternating sum is one of the principal properties of an exact sequence.

Now, for our sequence of spaces C_p with the maps $\partial_p : C_p \to C_{p-1}$, we have $\dim(\text{Ker}\, \partial_p) = \dim(\text{Im}\, \partial_{p+1}) + \dim H_p$. Using this and the range-nullspace

theorem in the same manner as above, shows that

$$\sum_{p=0}^{d}(-1)^p \dim C_p(M) = \sum_{p=0}^{d}(-1)^p \dim H_p(M). \tag{4.40}$$

This confirms our claim.

Exercise 4.1: Count the number of vertices, edges, and faces in the triangulation we used to compute the homology groups of the real projective plane $\mathbb{R}P^2$. Verify that $V - E + F = 1$, and that this is the same number that we get by evaluating

$$\chi(\mathbb{R}P^2) = \dim H_0(\mathbb{R}P^2, \mathbb{R}) - \dim H_1(\mathbb{R}P^2, \mathbb{R}) + \dim H_2(\mathbb{R}P^2, \mathbb{R}).$$

Exercise 4.2: Show that the sequence

$$\{\mathbf{0}\} \to V \stackrel{\phi}{\to} W \to \{\mathbf{0}\}$$

of vector spaces being exact means that the map $\phi : V \to W$ is one-to-one and onto, and hence an isomorphism $V \cong W$.

Exercise 4.3: Show that a *short exact sequence*

$$\{\mathbf{0}\} \to A \stackrel{i}{\to} B \stackrel{\pi}{\to} C \to \{\mathbf{0}\}$$

of vector spaces is just a sophisticated way of asserting that $C \cong B/A$. More precisely, show that the map i is injective (one-to-one), so A can be considered to be a subspace of B. Then show that the map π is surjective (onto), and can be regarded as projecting B onto the equivalence classes B/A.

Exercise 4.4: Let $\alpha : A \to B$ be a linear map. Show that

$$\{\mathbf{0}\} \to \mathrm{Ker}\,\alpha \stackrel{i}{\to} A \stackrel{\alpha}{\to} B \stackrel{\pi}{\to} \mathrm{Coker}\,\alpha \to \{\mathbf{0}\}$$

is an exact sequence. (Recall that $\mathrm{Coker}\,\alpha \equiv B/\mathrm{Im}\,\alpha$.)

4.3.2 Relative homology

Mathematicians have invented powerful tools for computing homology. In this section we introduce one of them: the *exact sequence of a pair*. We

describe this tool in detail because a homotopy analogue of this exact sequence is used in physics to classify defects such as dislocations, vortices and monopoles. Homotopy theory is however harder and requires more technical apparatus than homology, so the ideas are easier to explain here.

We have seen that it is useful to think of complicated manifolds as being assembled out of simpler ones. We constructed the torus, for example, by gluing together edges of a rectangle. Another construction technique involves *shrinking* parts of a manifold to a point. Think, for example, of the unit 2-disc as a being circle of cloth with a drawstring sewn into its boundary. Now pull the string tight to form a spherical bag. The continuous functions on the resulting 2-sphere are those continuous functions on the disc that took the same value at all points on its boundary. Recall that we used this idea in 3.4.2, where we claimed that those spin textures in $\mathbb{R}^2$ that point in a fixed direction at infinity can be thought of as spin textures on the 2-sphere. We now extend this shrinking trick to homology.

Suppose that we have a chain complex consisting of spaces C_p and boundary operations ∂_p. We wiill denote this chain complex by (C, ∂). Another set of of spaces and boundary operations (C', ∂') is a *subcomplex* of (C, ∂) if each $C'_p \subseteq C_p$ and $\partial'_p(c) = \partial_p(c)$ for each $c \in C'_p$. This situation arises if we have a simplical complex S and a some subset S' that is itself a simplicial complex, and take $C'_p = C_p(S')$

Since each C'_p is subspace of C_p we can form the quotient spaces C_p/C'_p and make them into a chain complex by defining, for $c + C'_p \in C_p/C'_p$,

$$\overline{\partial}_p(c + C'_p) = \partial_p c + C'_{p-1}. \tag{4.41}$$

It easy to see that this operation is well defined (*i.e.* it gives the same output independent of the choice of representative in the equivalence class $c + C'_p$), that $\overline{\partial}_p : C_p \to C_{p-1}$ is a linear map, and that $\overline{\partial}_{p-1}\overline{\partial}_p = 0$. We have constructed a new chain complex $(C/C', \overline{\partial})$. We can therefore form its homology spaces in the usual way. The resulting vector space, or abelian group, $H_p(C/C')$ is the p-th *relative homology group of C modulo C'*. When C' and C arise from simplicial complexes $S' \subseteq S$, these spaces are what remains of the homology of S after every chain in S' has been shrunk to a point. In this case, it is customary to write $H_p(S, S')$ instead of $H_p(C/C')$, and similarly write the chain, cycle and boundary spaces as $C_p(S, S')$, $Z_p(S, S')$ and $\mathrm{B}_p(S, S')$ respectively.

Example: Constructing the two-sphere S^2 from the two-ball (or disc) B^2. We regard B^2 to be the triangular simplex $P_1P_2P_3$, and its boundary, the

one-sphere or circle S^1, to be the simplicial complex containing the points P_1, P_2, P_3 and the sides P_1P_2, P_2P_3, P_3P_1, but not the interior of the triangle. We wish to contract this boundary complex to a point, and form the relative chain complexes and their homology spaces. Of the spaces we quotient by, $C_0(S^1)$ is spanned by the points P_1, P_2, P_3, the 1-chain space $C_1(S^1)$ is spanned by the sides P_1P_2, P_2P_3, P_3P_1, while $C_2(S^1) = \{\mathbf{0}\}$. The space of relative chains $C_2(B^1, S^1)$ consists of multiples of $P_1P_2P_3 + C_2(S^1)$, and the boundary

$$\overline{\partial}_2\Big(P_1P_2P_3 + C_2(S^1)\Big) = (P_2P_3 + P_3P_1 + P_1P_2) + C_1(S^1) \tag{4.42}$$

is equivalent to zero because $P_2P_3 + P_3P_1 + P_1P_2 \in C_1(S^1)$. Thus $P_1P_2P_3 + C_2(S^1)$ is a non-bounding cycle and spans $H_2(B^2, S^1)$, which is therefore one dimensional. This space is isomorphic to the one-dimensional $H_2(S^2)$. Similarly $H_1(B^2, S^1)$ is zero dimensional, and so isomorphic to $H_1(S^2)$. This is because all chains in $C_1(B^2, S^1)$ are in $C_1(S^1)$ and therefore equivalent to zero.

A peculiarity, however, is that $H_0(B^2, S^1)$ is *not* isomorphic to $H_0(S^2) = \mathbb{R}$. Instead, we find that $H_0(B^2, S^1) = \{\mathbf{0}\}$ because all the points are equivalent to zero. This vanishing is characteristic of the zeroth relative homology space $H_0(S, S')$ for the simplicial triangulation of any connected manifold. It occurs because S being connected means that any point P in S can be reached by walking along edges from any other point, in particular from a point P' in S'. This makes P homologous to P', and so equivalent to to zero in $H_0(S, S')$.

Exact homology sequence of a pair

Homological algebra is full of miracles. Here we describe one of them. From the ingredients we have at hand, we can construct a semi-infinite sequence of spaces and linear maps between them

$$\begin{aligned}
\cdots \xrightarrow{\partial_{*p+1}} H_p(S') \xrightarrow{i_{*p}} \quad & H_p(S) \quad \xrightarrow{\pi_{*p}} \quad H_p(S, S') \xrightarrow{\partial_{*p}} \\
H_{p-1}(S') \xrightarrow{i_{*p-1}} \quad & H_{p-1}(S) \quad \xrightarrow{\pi_{*p-1}} \quad H_{p-1}(S, S') \xrightarrow{\partial_{*p-1}} \\
& \vdots \\
\xrightarrow{\partial_{*1}} H_0(S') \xrightarrow{i_{*0}} \quad & H_0(S) \quad \xrightarrow{\pi_{*0}} \quad H_0(S, S') \xrightarrow{\partial_{*0}} \{\mathbf{0}\}.
\end{aligned} \tag{4.43}$$

The maps i_{*p} and π_{*p} are induced by the natural injection $i_p : C_p(S') \to C_p(S)$ and projection $\pi_p : C_p(S) \to C_p(S)/C_p(S')$. It is only necessary to check that

$$\begin{aligned} \pi_{p-1}\partial_p &= \overline{\partial}_p \pi_p, \\ i_{p-1}\partial_p &= \partial_p i_p, \end{aligned} \tag{4.44}$$

to see that they are compatible with the passage from the chain spaces to the homology spaces. More discussion is required of the *connection map* ∂_{*p} that takes us from one row to the next in the displayed form of (4.43).

Let $h \in H_p(S, S')$, then $h = z + B_p(S, S')$ for some cycle $z \in Z(S, S')$, and in turn $z = c + C_p(S')$ for some $c \in C_p(S)$. (So *two* choices of representative of equivalence class are being made here.) Now $\overline{\partial}_p z = 0$ which means that $\partial_p c \in C_{p-1}(S')$. This fact, when combined with $\partial_{p-1}\partial_p = 0$, tells us that $\partial_p c \in Z_{p-1}(S')$. We now set

$$\partial_{*p}(h) = \partial_p c + B_{p-1}(S'). \tag{4.45}$$

This sounds rather involved, but let's say it again in words: an element of $H_p(S, S')$ is a relative p-cycle *modulo* S'. This means that its boundary is not necessarily zero, but may be a non-zero element of $C_{p-1}(S')$. Since this element is the boundary of something its own boundary vanishes, so it is $(p-1)$-cycle in $C_{p-1}(S')$ and hence a representative of a homology class in $H_{p-1}(S')$. This homology class is the output of the ∂_{*p} map.

The miracle is that the sequence of maps (4.43) is *exact*. It is an example of a standard homological algebra construction of a *long exact sequence* out of a family of short exact sequences, in this case out the sequences

$$\{\mathbf{0}\} \to C_p(S') \to C_p(S) \to C_p(S, S') \to \{\mathbf{0}\}. \tag{4.46}$$

Proving that the long sequence is exact is straightforward. All one must do is check each map to see that it has the properties required. This exercise in *diagram chasing* is left to the reader.

This long exact sequence is called the *exact homology sequence of a pair*. If we know that certain homology spaces are zero dimensional, it provides a powerful tool for computing other spaces in the sequence. As an illustration, consider the sequence of the pair B^{n+1} and S^n for $n > 0$:

$$\cdots \xrightarrow{i_{*p}} \underbrace{H_p(B^{n+1})}_{=\{\mathbf{0}\}} \xrightarrow{\pi_{*p}} H_p(B^{n+1}, S^n) \xrightarrow{\partial_{*p}} H_{p-1}(S^n)$$

$$\begin{array}{ccccc}
\xrightarrow{i_{*p-1}} \underbrace{H_{p-1}(B^{n+1})}_{=\{\mathbf{0}\}} & \xrightarrow{\pi_{*p-1}} & H_{p-1}(B^{n+1},S^n) & \xrightarrow{\partial_{*p-1}} & H_{p-2}(S^n) \\
& & \vdots & & \\
\xrightarrow{i_{*1}} \underbrace{H_1(B^{n+1})}_{=\{\mathbf{0}\}} & \xrightarrow{\pi_{*1}} & H_1(B^{n+1},S^n) & \xrightarrow{\partial_{*1}} & \underbrace{H_0(S^n)}_{=\mathbb{R}} \\
\xrightarrow{i_{*0}} \underbrace{H_0(B^{n+1})}_{=\mathbb{R}} & \xrightarrow{\pi_{*0}} & H_0(B^{n+1},S^n) & \xrightarrow{\partial_{*0}} & \{\mathbf{0}\}.
\end{array} \tag{4.47}$$

We have inserted here the easily established data that $H_p(B^{n+1}) = \{\mathbf{0}\}$ for $p > 0$ (which is a consequence of the $(n+1)$-ball being a contractible space), and that $H_0(B^{n+1})$ and $H_0(S^n)$ are one dimensional because they consist of a single connected component. We read off, from the $\{\mathbf{0}\} \to A \to B \to \{\mathbf{0}\}$ exact subsequences, the isomorphisms

$$H_p(B^{n+1},S^n) \cong H_{p-1}(S^n), \quad p > 1, \tag{4.48}$$

and from the exact sequence

$$\{\mathbf{0}\} \to H_1(B^{n+1},S^1) \to \mathbb{R} \to \mathbb{R} \to H_0(B^{n+1},S^n) \to \{\mathbf{0}\} \tag{4.49}$$

that $H_1(B^{n+1},S^n) = \{\mathbf{0}\} = H_0(B^{n+1},S^n)$. The first of these equalities holds because $H_1(B^{n+1},S^n)$ is the kernel of the isomorphism $\mathbb{R} \to \mathbb{R}$, and the second because $H_0(B^{n+1},S^n)$ is the range of a surjective null map.

In the case $n = 0$, we have to modify our last conclusion because $H_0(S^0) = \mathbb{R} \oplus \mathbb{R}$ is two dimensional. (Remember that $H_0(M)$ counts the number of disconnected components of M, and the zero-sphere S^0 consists of the two disconnected points P_1, P_2 lying in the boundary of the interval $B^1 = P_1P_2$.) As a consequence, the last five maps become

$$\{\mathbf{0}\} \to H_1(B^1,S^0) \to \mathbb{R} \oplus \mathbb{R} \to \mathbb{R} \to H_0(B^1,S^0) \to \{\mathbf{0}\}. \tag{4.50}$$

This tells us that $H_1(B^1,S^0) = \mathbb{R}$ and $H_0(B^1,S^0) = \{\mathbf{0}\}$.

Exact homotopy sequence of a pair

We have met the homotopy groups $\pi_n(M)$ in section 3.4.4. As we saw there, homotopy groups can be used to classify defects or solitons in physical systems in which some field takes values in the manifold M. When the system

has undergone spontaneous symmetry breaking from a larger symmetry G to a subgroup H, the relevant manifold is the coset G/H. The group $\pi_n(G)$ can be taken to be the set of continuous maps of an n-dimensional cube into G, with the surface of the cube mapping to the identity element $e \in G$. We similarly define the relative homotopy group $\pi_n(G, H)$ of G *modulo* H to be the set of continuous maps of the cube into G, with all-but-one face of the cube mapping to e, but with the remaining face mapping to the subgroup H. It can then be shown that $\pi_n(G/H) \simeq \pi_n(G, H)$ (the hard part is to show that any continuous map into G/H can be represented as the projection of some continuous map into G).

The short exact sequence

$$\{e\} \to H \xrightarrow{i} G \xrightarrow{\pi} G/H \to \{e\} \tag{4.51}$$

of group homomorphisms (where $\{e\}$ is the group consisting only of the identity element) then gives rise to the long exact sequence

$$\cdots \to \pi_n(H) \to \pi_n(G) \to \pi_n(G, H) \to \pi_{n-1}(H) \to \cdots \tag{4.52}$$

The derivation and utility of this exact sequence is very well described in the review article by Mermin cited in section 3.4.4. We have therefore contented ourselves with simply displaying the result so that the reader can see the similarity between the homology theorem and its homotopy-theory analogue.

4.4 De Rham's Theorem

We still have not related homology to cohomology. The link is provided by integration.

The integral provides a natural pairing of a p-chain c and a p-form ω: if $c = a_1 s_1 + a_2 s_2 + \cdots + a_n s_n$, where the s_i are simplices, we set

$$(c, \omega) = \sum_i a_i \int_{s_i} \omega. \tag{4.53}$$

The perhaps mysterious notion of "adding" geometric simplices is thus given a concrete interpretation in terms of adding real numbers.

Stokes' theorem now reads

$$(\partial c, \omega) = (c, d\omega), \tag{4.54}$$

suggesting that d and ∂ should be regarded as adjoints of each other. From this observation follows the key fact that the pairing between chains and forms descends to a pairing between homology classes and cohomology classes. In other words,

$$(z + \partial c, \omega + d\chi) = (z, \omega), \tag{4.55}$$

so it does not matter which representative of the equivalence classes we take when we compute the integral. Let us see why this is so:

Suppose $z \in Z_p$ and $\omega_2 = \omega_1 + d\eta$. Then

$$\begin{aligned}(z, \omega_2) = \int_z \omega_2 &= \int_z \omega_1 + \int_z d\eta \\ &= \int_z \omega_1 + \int_{\partial z} \eta \\ &= \int_z \omega_1 \\ &= (z, \omega_1)\end{aligned} \tag{4.56}$$

because $\partial z = 0$. Thus, all elements of the cohomology class of ω return the same answer when integrated over a cycle.

Similarly, if $\omega \in Z^p$ and $c_2 = c_1 + \partial a$ then

$$\begin{aligned}(c_2, \omega) &= \int_{c_1} \omega + \int_{\partial a} \omega \\ &= \int_{c_1} \omega + \int_a d\omega \\ &= \int_{c_1} \omega \\ &= (c_1, \omega),\end{aligned}$$

since $d\omega = 0$.

All this means that we can consider the equivalence classes of closed forms composing $H^p_{\rm DR}(M)$ to be elements of $(H_p(M))^*$, the dual space of $H_p(M)$ — hence the "co" in cohomology. The existence of the pairing does not automatically mean that $H^p_{\rm DR}$ *is* the dual space to $H_p(M)$, however, because there might be elements of the dual space that are not in $H^p_{\rm DR}$, and there might be distinct elements of $H^p_{\rm DR}$ that give identical answers when integrated over any cycle, and so correspond to the same element in $(H_p(M))^*$. This

does not happen, however, when the manifold is *compact*: De Rham showed that, for compact manifolds, $(H_p(M,\mathbb{R}))^* = H^p_{\rm DR}(M,\mathbb{R})$. We will not try to prove this, but be satisfied with some examples.

The statement $(H_p(M))^* = H^p_{\rm DR}(M)$ neatly summarizes de Rham's results, but, in practice, the more explicit statements below are more useful.

Theorem: (de Rham) Suppose that M is a compact manifold.

1) A closed p-form ω is exact if and only if
$$\int_{z_i} \omega = 0 \tag{4.57}$$
for all cycles $z_i \in Z_p$. It suffices to check this for one representative of each homology class.
2) If $z_i \in Z_p$, $i = 1, \ldots, \dim H_p$, is a basis for the p-th homology space, and α_i a set of numbers, one for each z_i, then there exists a closed p-form ω such that
$$\int_{z_i} \omega = \alpha_i. \tag{4.58}$$

If ω^i constitute a basis of the vector space $H^p(M)$ then the matrix of numbers
$$\Omega_i{}^j = (z_i, \omega^j) = \int_{z_i} \omega^j \tag{4.59}$$
is called the *period matrix*, and the $\Omega_i{}^j$ themselves are the *periods*.

Example: $H_1(T^2) = \mathbb{R} \oplus \mathbb{R}$ is two-dimensional. Since a finite-dimensional vector space and its dual have the same dimension, de Rham tells us that $H^1_{\rm DR}(T^2)$ is also two-dimensional. If we take as coordinates on T^2 the angles θ and ϕ, then the basis elements, or *generators*, of the cohomology spaces are the forms "$d\theta$" and "$d\phi$". We have inserted the quotes to stress that these expressions are not the d of a function. The angles θ and ϕ are *not* functions on the torus, since they are not single-valued. The homology basis 1-cycles can be taken as z_θ running from $\theta = 0$ to $\theta = 2\pi$ along $\phi = \pi$, and z_ϕ running from $\phi = 0$ to $\phi = 2\pi$ along $\theta = \pi$. Clearly, $\omega = \alpha_\theta d\theta/2\pi + \alpha_\phi d\phi/2\pi$ returns $\int_{z_\theta} \omega = \alpha_\theta$ and $\int_{z_\phi} \omega = \alpha_\phi$ for any $\alpha_\theta, \alpha_\pi$, so $\{d\theta/2\pi, d\phi/2\pi\}$ and $\{z_\theta, z_\phi\}$ are dual bases.

Example: We have earlier computed $H_2(\mathbb{R}P^2,\mathbb{R}) = \{\mathbf{0}\}$ and $H_1(\mathbb{R}P^2,\mathbb{R}) = \{\mathbf{0}\}$. De Rham therefore tells us that $H^2(\mathbb{R}P^2,\mathbb{R}) = \{\mathbf{0}\}$ and $H^1(\mathbb{R}P^2,\mathbb{R}) = \{\mathbf{0}\}$. From this we deduce that all closed one- and two-forms on the projective plane $\mathbb{R}P^2$ are exact.

Example: As an illustration of de Rham part 1), observe that it is easy to show that a closed one-forms ϕ can be written as df, provided that $\int_{z_i} \phi = 0$ for all cycles. We simply define $f = \int_{x_0}^{x} \phi$, and observe that the proviso ensures that f is not multivalued.
Example: A more subtle problem is to show that, given a two-form ω on S^2, with $\int_{S^2} \omega = 0$, then there is a globally defined χ such that $\omega = d\chi$. We begin by covering S^2 by two open sets D_+ and D_- which have the form of caps such that D_+ includes all of S^2 except for a neighbourhood of the south pole, while D_- includes everything except a neighbourhood of the north pole, and the intersection, $D_+ \cap D_-$, has the topology of an annulus, or *cingulum*, encircling the equator.

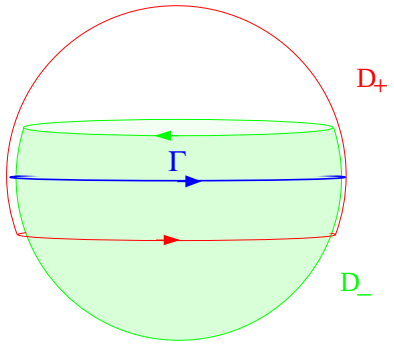

Figure 4.10: *A covering the sphere by two contractable caps.*

Since both D_+ and D_- are contractable, there are one-forms χ_+ and χ_- such that $\omega = d\chi_+$ in D_+ and $\omega = d\chi_-$ in D_-. Thus,

$$d(\chi_+ - \chi_-) = 0, \quad \text{in} \quad D_+ \cap D_-. \tag{4.60}$$

Dividing the sphere into two disjoint sets with a common (but oppositely oriented) boundary $\Gamma \in D_+ \cap D_-$ we have

$$0 = \int_{S^2} \omega = \oint_{\Gamma} (\chi_+ - \chi_-), \tag{4.61}$$

and this is true for any such curve Γ. Thus, by the previous example,

$$\phi \equiv (\chi_+ - \chi_-) = df \tag{4.62}$$

for some smooth function f defined in $D_+ \cap D_-$. We now introduce a *partition of unity* subordinate to the cover of S^2 by D_+ and D_-. This partition is a

pair of non-negative smooth functions, $\rho_\pm$, such that ρ_+ is non-zero only in D_+, ρ_- is non-zero only in D_-, and $\rho_+ + \rho_- = 1$. Now

$$f = \rho_+ f - (-\rho_-) f, \tag{4.63}$$

and $f_- = \rho_+ f$ is a function defined everywhere on D_-. Similarly $f_+ = (-\rho_-) f$ is a function on D_+. Notice the interchange of $\pm$ labels! This is not a mistake. The function f is not defined outside $D_+ \cap D_-$, but we can define $\rho_- f$ everywhere on D_+ because f gets multiplied by zero wherever we have no specific value to assign to it.

We now observe that

$$\chi_+ + df_+ = \chi_- + df_-, \quad \text{in} \quad D_+ \cap D_-. \tag{4.64}$$

Thus $\omega = d\chi$, where χ is defined everywhere by the rule

$$\chi = \begin{cases} \chi_+ + df_+, & \text{in } D_+, \\ \chi_- + df_-, & \text{in } D_-. \end{cases} \tag{4.65}$$

It does not matter which definition we take in the cingular region $D_+ \cap D_-$, because the two definitions coincide there.

The methods of this example, a special case of the *Mayer-Vietoris principle*, can be extended to give a proof of de Rham's claims.

4.5 Poincaré Duality

De Rham's theorem does not require that our manifold M be orientable. Our next results do, however, require orientablity. We therefore assume throughout this section that M is a compact, orientable, D-dimensional manifold.

We begin with the observation that if the forms ω_1 and ω_2 are closed then so is $\omega_1 \wedge \omega_2$. Furthermore if one or both of ω_1, ω_2 is exact then the product $\omega_1 \wedge \omega_2$ is also exact. It follows that the cohomology class $[\omega_1 \wedge \omega_2]$ of $\omega_1 \wedge \omega_2$ depends only on the cohomology classes $[\omega_1]$ and $[\omega_2]$. The wedge product thus induces a map

$$H^p(M, \mathbb{R}) \times H^q(M, \mathbb{R}) \stackrel{\wedge}{\to} H^{p+q}(M, \mathbb{R}), \tag{4.66}$$

which is called the "cup product" of the cohomology classes. It is written as

$$[\omega_1 \wedge \omega_2] = [\omega_1] \cup [\omega_2], \tag{4.67}$$

and gives the cohomology the structure of a graded-commutative ring, denoted by $H^\bullet(M, R)$

More significant for us than the ring structure is that, given $\omega \in H^D(M, \mathbb{R})$, we can obtain a real number by forming $\int_M \omega$ (This is the point at which we need orientability. We only know how to integrate over orientable chains, and so cannot even define $\int_M \omega$ when M is not orientable.) and can combine this integral with the cup product to make any cohomology class $[f] \in H^{D-p}(M, \mathbb{R})$ into an element F of $(H^p(M, \mathbb{R}))^*$. We do this by setting

$$F([g]) = \int_M f \wedge g \tag{4.68}$$

for each $[g] \in H^p(M, \mathbb{R})$. Furthermore, it is possible to show that we can get *any* element F of $(H^p(M, \mathbb{R}))^*$ in this way, and the corresponding $[f]$ is *unique*. But de Rham has already given us a way of identifying the elements of $(H^p(M, \mathbb{R}))^*$ with the cycles in $H_p(M, \mathbb{R})$! There is, therefore, a 1-1 onto map

$$H_p(M, \mathbb{R}) \leftrightarrow H^{D-p}(M, \mathbb{R}). \tag{4.69}$$

In particular the dimensions of these two spaces must coincide

$$b_p(M) = b_{D-p}(M). \tag{4.70}$$

This equality of Betti numbers is called *Poincaré duality*. Poincaré originally conceived of it geometrically. His idea was to construct from each simplicial triangulation S of M a new "dual" triangulation S', where, in two dimensions for example, we place a new vertex at the centre of each triangle, and join the vertices by lines through each side of the old triangles to make new cells — each new cell containing one of the old vertices. If we are lucky, this process will have the effect of replacing each p-simplex by a $(D-p)$-simplex, and so set up a map between $C_p(S)$ and $C_{D-p}(S')$ that turns the homolgy "upside down." The new cells are not always simplices, however, and it is hard to make this construction systematic. Poincaré's original recipe was flawed.

Our present approach to Poincaré's result is asserting that for each basis p-cycle class $[z_i^p]$ there is a unique (up to cohomology) $(D-p)$-form ω_i^{D-p} such that

$$\int_{z_i^p} f = \int_M \omega_i^{D-p} \wedge f. \tag{4.71}$$

We can construct this ω_i^{D-p} "physically" by taking a representative cycle z_i^p in the homology class $[z_i^p]$ and thinking of it as a surface with a conserved

unit $(d-p)$-form current flowing in its vicinity. An example would be the two-form topological current running along the one-dimensional worldline of a Skyrmion. (See the discussion surrounding equation (3.63).) The ω_i^{D-p} form a basis for $H^{D-p}(M,\mathbb{R})$. We can therefore expand $f \sim f^i\omega_i^{D-p}$, and similarly for the closed p-form g, to obtain

$$\int_M g \wedge f = f^i g^j I(i,j) \tag{4.72}$$

where the matrix

$$I(i,j) \equiv I(z_i^p, z_j^{D-p}) = \int_M \omega_i^{D-p} \wedge \omega_j^p \tag{4.73}$$

is called the *intersection form*. From the definition we have

$$I(i,j) = (-1)^{p(D-p)} I(j,i). \tag{4.74}$$

Less obvious is that $I(i,j)$ is an *integer* that reports the number of times (counted with orientation) that the cycles z_i^p and z_j^{D-p} intersect. This latter fact can be understood from our construction of the ω_i^p as unit currents localized near the z_i^{D-p} cycles. The integrand in (4.73) is non-zero only in the neighbourhood of the intersections of z_i^p with z_j^{D-p}, and at each intersection constitutes a D-form that integrates up to give ± 1.

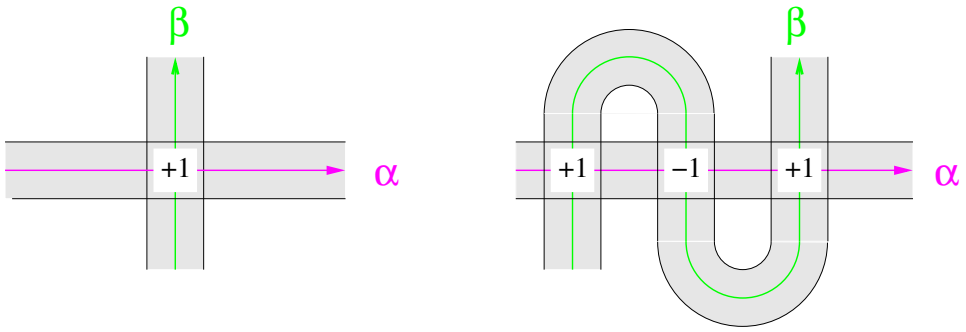

Figure 4.11: *The intersection of two cycles:* $I(\alpha,\beta) = 1 = 1 - 1 + 1$.

This claim is illustrated in the left-hand part of figure 4.11, which shows a region surrounding the intersection of the α and β one-cycles on the 2-torus. The co-ordinate system has been chosen so that the α cycle runs along the

x axis and the β cycle along then y axis. Each cycle is surrounded by the narrow shaded regions $-w < y < w$ and $-w < x < w$, respectively. To construct suitable forms ω_α and ω_β we select a smooth function $f(x)$ that vanishes for $|x| \geq w$ and such that $\int f\,dx = 1$. In the local chart we can then set

$$\begin{aligned} \omega_\alpha &= f(y)\,dy, \\ \omega_\beta &= -f(x)\,dx, \end{aligned}$$

both these forms being closed. The intersection number is given by the integral

$$I(\alpha,\beta) = \int \omega_\alpha \wedge \omega_\beta = \iint f(x)f(y)\,dxdy = 1. \tag{4.75}$$

The right-hand part of figure 4.11 illustrates why this intersection number depends only on the homology classes of the two one-cycles, and not on their particular instantiation as curves.

We can more conveniently re-express (4.72) terms of the *periods* of the forms

$$f_i \equiv \int_{z_i^p} f = I(i,k)f^k, \quad g_j \equiv \int_{z_j^{D-p}} g = I(j,l)g^l, \tag{4.76}$$

as

$$\int_M f \wedge g = \sum_{i,j} K(i,j) \int_{z_i^p} f \int_{z_j^{D-p}} g, \tag{4.77}$$

where

$$K(i,j) = I^{-1}(i,k)I^{-1}(j,l)I(k,l) = I^{-1}(j,i) \tag{4.78}$$

is the transpose of the inverse of the intersection-form matrix. The decomposition (4.77) of the integral of the product of a pair of closed forms into a bilinear form in their periods is one of the two principal results of this section, the other being (4.70).

In simple cases we can obtain the decomposition (4.77) by more direct methods. Suppose, for example, that we label the cycles generating the homology group $H_1(T^2)$ of the 2-torus as α and β, and that a and b are closed ($da = db = 0$), but not necessarily exact, one-forms. We will show that

$$\int_{T^2} a \wedge b = \int_\alpha a \int_\beta b - \int_\alpha b \int_\beta a. \tag{4.79}$$

To do this, we cut the torus along the cycles α and β and open it out into a rectangle with sides of length L_x and L_y. The cycles α and β will form the sides of the rectangle and we will take them as lying parallel to the x and y axes, respectively. Functions on the *torus* now become functions on the *rectangle*. Not all functions on the rectangle descend from functions on the torus, however. Only those functions that satisfy the periodic boundary conditions $f(0,y) = f(L_x,y)$ and $f(x,0) = f(x,L_y)$ can be considered (mathematicians would say "can be *lifted*") to be functions on the torus.

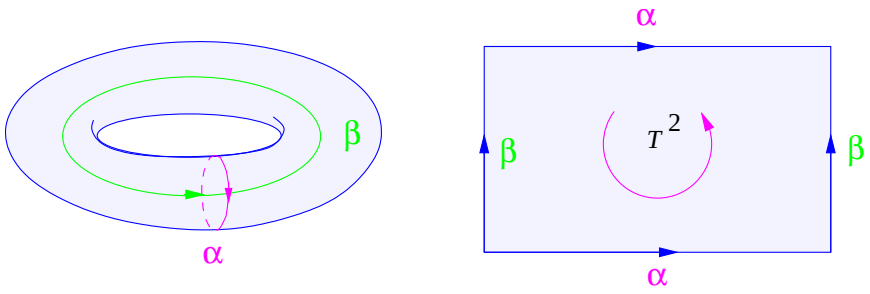

Figure 4.12: *Cut-open torus*

Since the rectangle (but not the torus) is retractable, we can write $a = df$ where f is a function on the rectangle — but not necessarily a function on the torus, *i.e.*, f will not, in general, be periodic. Since $a \wedge b = d(fb)$, we can now use Stokes' theorem to evaluate

$$\int_{T^2} a \wedge b = \int_{T^2} d(fb) = \int_{\partial T^2} fb. \tag{4.80}$$

The two integrals on the two vertical sides of the rectangle can be combined to a single integral over the points of the one-cycle β:

$$\int_{\text{vertical}} fb = \int_{\beta} [f(L_x,y) - f(0,y)]b. \tag{4.81}$$

We now observe that $[f(L_x,y) - f(0,y)]$ is a constant, and so can be taken out of the integral. It is a constant because all paths from the point $(0,y)$ to (L_x,y) are homologous to the one-cycle α, so the difference $f(L_x,y) - f(0,y)$ is equal to $\int_\alpha a$. Thus

$$\int_{\beta} [f(L_x,y) - f(0,y)]b = \int_{\alpha} a \int_{\beta} b. \tag{4.82}$$

Similarly, the contributions of the two horizontal sides is

$$\int_\alpha [f(x,0) - f((x,L_y)]b = -\int_\beta a \int_\alpha b. \tag{4.83}$$

On putting the contributions of both pairs of sides together, the claimed result follows.

4.6 Characteristic Classes

A supply of elements of $H^{2m}(M,\mathbb{R})$ and $H^{2m}(M,\mathbb{Z})$ is provided by the *characteristic classes* associated with connections on vector bundles over the manifold M.

Recall that connections appear in covariant derivatives

$$\nabla_\mu \equiv \partial_\mu + A_\mu, \tag{4.84}$$

and are to be thought of as matrix-valued one-forms $A = A_\mu dx^\mu$. In the quantum mechanics of charged particles the covariant derivative that appears in the Schrödinger equation is

$$\nabla_\mu = \frac{\partial}{\partial x^\mu} - ieA_\mu^{\text{Maxwell}}. \tag{4.85}$$

Here e is the charge of the particle on whose wavefunction the derivative acts, and A_μ^{Maxwell} is the usual electromagnetic vector potential. The matrix-valued connection one-form is therefore

$$A = -ieA_\mu^{\text{Maxwell}} dx^\mu. \tag{4.86}$$

In this case the matrix is one-by-one.

In a non-abelian gauge theory with gauge group G the connection becomes

$$A = i\hat{\lambda}_a A_\mu^a dx^\mu \tag{4.87}$$

The $\hat{\lambda}_a$ are hermitian matrices that have commutation relations $[\hat{\lambda}_a, \hat{\lambda}_b] = if_{ab}^c \hat{\lambda}_c$, where the f_{ab}^c are the structure constants of the Lie algebra of G. The $\hat{\lambda}_a$ therefore form a representation of the Lie algebra, and this representation plays the role of the "charge" of the non-abelian gauge particle.

For covariant derivatives acting on a tangent vector field $f^a\mathbf{e}_a$ on a Riemann n-manifold, where the $\mathbf{e}_a$ are an orthonormal vielbein frame, we have

$$A = \omega_{ab\mu}dx^\mu, \tag{4.88}$$

where, for each μ, the coefficients $\omega_{ab\mu} = -\omega_{ba\mu}$ can be thought of as the entries in a skew symmetric n-by-n matrix. These matrices are elements of the Lie algebra $\mathfrak{o}(n)$ of $\mathrm{O}(n)$.

In all these cases we define the curvature two-form to be $F = dA + A^2$, where a combined matrix and wedge product is to be understood in A^2. In exercises 2.19 and 2.20 you used the Bianchi identity to show that the gauge-invariant $2n$-forms $\mathrm{tr}\,(F^n)$ were closed. The integrals of these forms over cycles provide numbers that are topological invariants of the bundle. For example, in four-dimensional QCD, the integral

$$c_2 = -\frac{1}{8\pi^2}\int_\Omega \mathrm{tr}\,(F^2), \tag{4.89}$$

over a compactified four-dimensional manifold Ω is an integer that a mathematician would call the second Chern number of the non-abelian gauge bundle, and that a physicist would call the *instanton number* of the gauge field configuration.

In this section we will show that the integrals of such characteristic classes are indeed topological invariants. We also explain something of what these invariants are measuring, and illustrate why, when suitably normalized, certain of them are integer valued.

4.6.1 Topological invariance

Suppose that we have been given a connection A and slightly deform it $A \to A + \delta A$, then

$$\delta F = d(\delta A) + \delta A\, A + A\,\delta A. \tag{4.90}$$

Using the Bianchi identity $dF = FA - AF$, we find that

$$\begin{aligned}
\delta\,\mathrm{tr}(F^n) &= n\,\mathrm{tr}(\delta F\, F^{n-1}) \\
&= n\,\mathrm{tr}(d(\delta A)F^{n-1}) + n\,\mathrm{tr}(\delta A\, AF^{n-1}) + n\,\mathrm{tr}(A\,\delta AF^{n-1}) \\
&= n\,\mathrm{tr}(d(\delta A)F^{n-1}) + n\,\mathrm{tr}(\delta A\, AF^{n-1}) - n\,\mathrm{tr}(\delta A\, F^{n-1}A) \\
&= d\left\{n\,\mathrm{tr}(\delta A\, F^{n-1})\right\}.
\end{aligned} \tag{4.91}$$

The last line of (4.91) is equal to the penultimate line because all but the first and last terms arising from the dF's in $d\left\{\mathrm{tr}(\delta A\, F^{n-1})\right\}$ cancel in pairs. A globally defined change in A therefore changes $\mathrm{tr}(F^n)$ by the d of something, and so does not change its cohomology class, or its integral over a cycle.

At first sight, this invariance under deformation suggests that all the $\mathrm{tr}(F^n)$ are exact forms — they can apparently all be written as $\mathrm{tr}(F^n) = d\omega_{2n-1}(A)$ for some $(2n-1)$-form $\omega_{2n-1}(A)$. To find $\omega_{2n-1}(A)$ all we have to do is deform the connection to zero by setting $A_t = t\,A$ and

$$F_t = dA_t + A_t^2 = tdA + t^2A^2. \tag{4.92}$$

Then $\delta A_t = A\delta t$, and

$$\frac{d}{dt}\mathrm{tr}(F_t^n) = d\left\{n\,\mathrm{tr}(AF_t^{n-1})\right\}. \tag{4.93}$$

Integrating up from $t = 0$, we find

$$\mathrm{tr}(F^n) = d\left\{n\int_0^1 \mathrm{tr}(AF_t^{n-1})\,dt\right\}. \tag{4.94}$$

For example

$$\begin{aligned}\mathrm{tr}(F^2) &= d\left\{2\int_0^1 \mathrm{tr}(A(tdA + t^2A^2)\,dt\right\}\\ &= d\left\{\mathrm{tr}\left(AdA + \frac{2}{3}A^3\right)\right\}.\end{aligned} \tag{4.95}$$

You should recognize here the $\omega_3(A) = \mathrm{tr}(AdA + \frac{2}{3}A^3)$ Chern-Simons form of exercise 2.19. The naïve conclusion — that all the $\mathrm{tr}(F^n)$ are exact — is false, however. What the computation actually shows is that when $\int \mathrm{tr}(F^n) \neq 0$ we cannot find a globally defined one-form A representing the connection or gauge field. With no global A, we cannot globally deform A to zero.

Consider, for example, an Abelian U(1) gauge field on the two-sphere S^2. When the first Chern-number

$$c_1 = \frac{1}{2\pi i}\int_{S^2} F \tag{4.96}$$

is non-zero, there can be no globally defined one-form A such that $F = dA$. Glance back, however, at figure 4.10 on page 141. There we see that

the retractability of the spherical caps $D_\pm$ guarantees that there are one-forms $A_\pm$ defined on $D_\pm$ such that $F = dA_\pm$ in $D_\pm$. In the cingular region $D_+ \cap D_-$ where they are both defined, A_+ and A_- will be related by a gauge transformation. For a U(1) gauge field, the matrix g appearing in the general gauge transformation rule

$$A \to A^g \equiv g^{-1}Ag + g^{-1}dg, \tag{4.97}$$

of exercise 2.20 becomes the phase $e^{i\chi} \in \mathrm{U}(1)$. Consequently

$$A_+ = A_- + e^{-i\chi}de^{i\chi} = A_- + id\chi \quad \text{in} \quad D_+ \cap D_-. \tag{4.98}$$

The U(1) group element $e^{i\chi}$ is required to be single valued in $D_+ \cap D_-$, but the angle χ may be multivalued. We now write c_1 as the sum of integrals over the north and south hemispheres of S^2, and use Stokes theorem to reduce this sum to a single integral over the hemispheres' common boundary, the equator Γ.

$$\begin{aligned} c_1 &= \frac{1}{2\pi i}\int_{\text{north}} F + \frac{1}{2\pi i}\int_{\text{south}} F \\ &= \frac{1}{2\pi i}\int_{\text{north}} dA_+ + \frac{1}{2\pi i}\int_{\text{south}} dA_- \\ &= \frac{1}{2\pi i}\int_\Gamma A_+ - \frac{1}{2\pi i}\int_\Gamma A_- \\ &= \frac{1}{2\pi}\int_\Gamma d\chi \end{aligned} \tag{4.99}$$

We see that c_1 is the integer counting the winding of χ as we circle Γ. An integer cannot be continuously reduced to zero, and if we attempt to deform $A \to tA \to 0$, we will violate the required single-valuedness of the U(1) group element $e^{i\chi}$.

Although the Chern-Simons forms $\omega_{2n-1}(A)$ cannot be defined globally, they are still very useful in physics. They occur as *Wess-Zumino terms* describing the low energy properties of various quantum field theories, the prototype being the Skyrme-Witten model of Hadrons.[2]

[2]E. Witten, *Nucl. Phys.* **B223** (1983) 422; *ibid.* **B223** (1983) 433.

4.6.2 Chern characters and Chern classes

Any gauge-invariant polynomial (with exterior multiplication of forms understood) in F provides a closed, topologically invariant, differential form. Certain combinations, however, have additional desirable properties, and so have been given names.

The form

$$\mathrm{ch}_n(F) = \mathrm{tr}\left\{\frac{1}{n!}\left(\frac{i}{2\pi}F\right)^n\right\} \tag{4.100}$$

is called the n-th *Chern character*. It is convenient to think of this $2n$-form as being the n-th term in a generating-function expansion

$$\mathrm{ch}(F) \stackrel{\mathrm{def}}{=} \mathrm{tr}\left\{\exp\left(\frac{i}{2\pi}F\right)\right\} = \mathrm{ch}_0(F) + \mathrm{ch}_1(F) + \mathrm{ch}_2(F) + \cdots, \tag{4.101}$$

where $\mathrm{ch}_0(F) \equiv \mathrm{tr}\, I$ is the dimension of the space on which the $\hat{\lambda}_a$ act. This formal sum of forms of different degree is called the *total Chern character*. The $n!$ normalization is chosen because it makes the Chern character behave nicely when we combine vector bundles.

Given two vector bundles over the same manifold, having fibres U_x and V_x over the point x, we can make a new bundle with the direct sum $U_x \oplus V_x$ as fibre over x. This resulting bundle is called the *Whitney sum* of the bundles. Similarly we can make a tensor-product bundle whose fibre over x is $U_x \otimes V_x$.

Let us use the notation $\mathrm{ch}(U)$ to represent the Chern character of the bundle with fibres U_x, and $U \oplus V$ to denote the Whitney sum. Then we have

$$\mathrm{ch}(U \oplus V) = \mathrm{ch}(U) + \mathrm{ch}(V), \tag{4.102}$$

and

$$\mathrm{ch}(U \otimes V) = \mathrm{ch}(U) \wedge \mathrm{ch}(V). \tag{4.103}$$

The second of these formulæ comes about because if $\hat{\lambda}_a^{(1)}$ is a Lie algebra element acting on $V^{(1)}$ and $\hat{\lambda}_a^{(2)}$ the corresponding element acting on $V^{(2)}$, then they act on the tensor product $V^{(1)} \otimes V^{(2)}$ as

$$\hat{\lambda}_a^{(1\otimes 2)} = \hat{\lambda}_a^{(1)} \otimes I + I \otimes \hat{\lambda}_a^{(2)}, \tag{4.104}$$

where I is the identity operator, and for matrices A, B,

$$\mathrm{tr}\left\{\exp\left(A \otimes I + I \otimes B\right)\right\} = \mathrm{tr}\left\{\exp A \otimes \exp B\right\} = \mathrm{tr}\left\{\exp A\right\} \mathrm{tr}\left\{\exp B\right\}. \tag{4.105}$$

In terms of the individual $\mathrm{ch}_n(V)$ equations (4.102) and (4.103) read

$$\mathrm{ch}_n(U \oplus V) = \mathrm{ch}_n(U) + \mathrm{ch}_n(V), \tag{4.106}$$

and

$$\mathrm{ch}_n(U \otimes V) = \sum_{m=0}^{n} \mathrm{ch}_{n-m}(U) \wedge \mathrm{ch}_m(V). \tag{4.107}$$

Related to the Chern characters are the *Chern classes*. These are wedge-product polynomials in the Chern characters, and are defined, *via* the matrix expansion

$$\det(I + A) = 1 + \mathrm{tr}\, A + \frac{1}{2}\Big((\mathrm{tr}\, A)^2 - \mathrm{tr}\, A^2\Big) + \ldots, \tag{4.108}$$

by the generating function for the total Chern class

$$c(F) = \det\left(I + \frac{i}{2\pi}F\right) = 1 + c_1(F) + c_2(F) + \cdots. \tag{4.109}$$

Thus

$$c_1(F) = \mathrm{ch}_1(F), \quad c_2(F) = \frac{1}{2}\mathrm{ch}_1(F) \wedge \mathrm{ch}_1(F) - \mathrm{ch}_2(F), \tag{4.110}$$

and so on.

For matrices A and B we have $\det(A \oplus B) = \det(A)\det(B)$, and this leads to

$$c(U \oplus V) = c(U) \wedge c(V). \tag{4.111}$$

Although the Chern classes are more complicated in appearance than the Chern characters, they are introduced because their integrals over cycles are *integers*, and this property remains true of integer-coefficient sums of products of Chern-classes. The cohomology classes $[c_n(F)]$ are therefore elements of the integer cohomology ring $H^\bullet(M, \mathbb{Z})$. This property does not hold for the Chern characters, whose integrals over cycles can be fractions. The cohomology classes $[\mathrm{ch}_n(F)]$ are therefore only elements of $H^\bullet(M, \mathbb{Q})$.

When we integrate products of Chern classes of total degree $2m$ over a closed $2m$-dimensional orientable manifold we get integer *Chern numbers*. These integers can be related to generalized winding numbers, and characterize the extent to which the gauge transformations that relate the connection fields in different patches serve to *twist* the vector bundle. Unfortunately it requires a considerable amount of machinery (the Schubert calculus of complex Grassmannians) to explain these integers.

Pontryagin and Euler classes

When the fibres of a vector bundle are vector spaces over $\mathbb{R}$, the complex skew-hermitian matrices $i\hat{\lambda}_a$ are replaced by real skew symmetric matrices. The Lie algebra of the n-by-n matrices $i\hat{\lambda}_a$ was a subalgebra of $\mathfrak{u}(n)$. The Lie algebra of the n-by-n real, skew symmetric, matrices is a subalgebra of $\mathfrak{o}(n)$. Now the trace of an odd power of any skew symmetric matrix is zero. As a consequence, Chern characters and Chern classes containing an odd number of F's all vanish. The remaining real $4n$-forms are known as *Pontryagin classes*. The precise definition is

$$p_k(V) = (-1)^k c_{2k}(V). \tag{4.112}$$

Pontryagin classes help to classify bundles whose gauge transformations are elements of $\mathrm{O}(n)$. If we restrict ourselves to gauge transformations that lie in $\mathrm{SO}(n)$, as we would when considering the tangent bundle of an *orientable* Riemann manifold, then we can make a gauge-invariant polynomial out of the skew-symmetric matrix-valued F by forming its *Pfaffian*.

Recall (or see exercise **??.??**) that the Pfaffian of a skew symmetric $2n$-by-$2n$ matrix $\mathbf{A}$ with entries a_{ij} is

$$\mathrm{Pf}\,\mathbf{A} = \frac{1}{2^n n!}\epsilon_{i_1,\ldots i_{2n}} a_{i_1 i_2}\cdots a_{i_{2n-1} i_{2n}}. \tag{4.113}$$

The *Euler class* of the tangent bundle of a $2n$-dimensional orientable manifold is defined *via* its skew-symmetric Riemann-curvature form

$$\mathbf{R} = \frac{1}{2} R_{ab,\mu\nu} dx^\mu dx^\nu \tag{4.114}$$

to be

$$e(\mathbf{R}) = \mathrm{Pf}\left(\frac{1}{2\pi}\mathbf{R}\right). \tag{4.115}$$

In four dimensions, for example, this becomes the 4-form

$$e(\mathbf{R}) = \frac{1}{32\pi^2}\epsilon_{abcd} R_{ab} R_{cd}. \tag{4.116}$$

The generalized *Gauss-Bonnet theorem* asserts, for an oriented, even-dimensional, manifold without boundary, that the Euler character is given by

$$\chi(M) = \int_M e(\mathbf{R}). \tag{4.117}$$

We will not prove this theorem, but in section 7.3.6 we will illustrate the strategy that leads to Chern's influential proof.

Exercise 4.5: Show that

$$c_3(F) = \frac{1}{6}\Big((\mathrm{ch}_1(F))^3 - 6\,\mathrm{ch}_1(F)\mathrm{ch}_2(F) + 12\,\mathrm{ch}_3(F)\Big).$$

4.7 Hodge Theory and the Morse Index

The Laplacian, when acting on a scalar function ϕ in $\mathbb{R}^3$ is simply div (grad ϕ), but when acting on a vector $\mathbf{v}$ it becomes

$$\nabla^2\mathbf{v} = \mathrm{grad}\,(\mathrm{div}\,\mathbf{v}) - \mathrm{curl}\,(\mathrm{curl}\,\mathbf{v}). \tag{4.118}$$

Is there a general construction that would have allowed us to write down this second expression? What about the Laplacian on other types of fields?

The Laplacian acting on any vector or tensor field $\mathbf{T}$ in $\mathbb{R}^n$ is given, in general curvilinear co-ordinates, by $\nabla^2\mathbf{T} = g^{\mu\nu}\nabla_\mu\nabla_\nu\mathbf{T}$ where ∇_μ is the flat-space covariant derivative. This is the unique co-ordinate independent object that reduces in Cartesian co-ordinates to the ordinary Laplacian acting on the individual components of $\mathbf{T}$. The proof that the rather different-seeming (4.118) holds for vectors is that it too is constructed out of co-ordinate independent operations and in Cartesian co-ordinates reduces to the ordinary Laplacian acting on the individual components of $\mathbf{v}$. It must therefore coincide with the covariant derivative definition. Why it should work out this way is not exactly obvious. Now div, grad and curl can all be expressed in differential form language, and therefore so can the scalar and vector Laplacian. Moreover, when we let the Laplacian act on any p-form the general pattern becomes clear. The differential form definition of the Laplacian, and the exploration of its consequences, was the work of William Hodge in the 1930's. His theory has natural applications to the topology of manifolds.

4.7.1 The Laplacian on p-forms

Suppose that M is an oriented, compact, D-dimensional manifold without boundary. We can make the space $\Omega^p(M)$ of p-form fields on M into an L^2

Hilbert space by introducing the positive-definite inner product

$$\langle a, b\rangle_p = \langle b, a\rangle_p = \int_M a \star b = \frac{1}{p!}\int d^D x \sqrt{g}\, a_{i_1 i_2 \ldots i_p} b^{i_1 i_2 \ldots i_p}. \tag{4.119}$$

Here the subscript p denotes the order of the forms in the product, and should not to be confused with the p we have elsewhere used to label the norm in L^p Banach spaces. The presence of the $\sqrt{g}$ and the Hodge $\star$ operator tells us that this inner product depends on both the metric on M and the global orientation.

We can use our new product to define a "hermitian adjoint" $\delta \equiv d^\dagger$ of the exterior differential operator d. The "..." are because this is not quite an adjoint operator in the normal sense — d takes us from one vector space to another — but it is constructed in an analogous manner. We define δ by requiring that

$$\langle da, b\rangle_{p+1} = \langle a, \delta b\rangle_p, \tag{4.120}$$

where a is an arbitrary p-form and b and arbitrary $(p+1)$-form. Now recall that $\star$ takes p-forms to $(D-p)$ forms, and so $d\star b$ is a $(D-p)$ form. Acting twice on a $(D-p)$-form with $\star$ gives us back the original form multiplied by $(-1)^{p(D-p)}$. We use this to compute

$$\begin{aligned} d(a\star b) &= da\star b + (-1)^p a(d\star b) \\ &= da\star b + (-1)^p(-1)^{p(D-p)} a\star(\star d\star b) \\ &= da\star b - (-1)^{Dp+1} a\star(\star d\star b). \end{aligned} \tag{4.121}$$

In obtaining the last line we have observed that $p(p-1)$ is an even integer and so $(-1)^{p(1-p)} = 1$. Now, using Stokes' theorem, and the absence of a boundary to discard the integrated-out part, we conclude that

$$\int_M (da)\star b = (-1)^{Dp+1}\int_M a\star(\star d\star b), \tag{4.122}$$

or

$$\langle da, b\rangle_{p+1} = (-1)^{Dp+1}\langle a, (\star d\star) b\rangle_p \tag{4.123}$$

and so $\delta b = (-1)^{Dp+1}(\star d\star)b$. This was for δ acting on a $(p-1)$ form. Acting on a p form we have

$$\delta = (-1)^{Dp+D+1}\star d\star. \tag{4.124}$$

Observe how the sequence of maps in $\star d \star$ works:

$$\Omega^p(M) \xrightarrow{\star} \Omega^{D-p}(M) \xrightarrow{d} \Omega^{D-p+1}(M) \xrightarrow{\star} \Omega^{p-1}(M). \tag{4.125}$$

The net effect is that δ takes a p-form to a $(p-1)$-form. Observe also that $\delta^2 \propto \star d^2 \star = 0$.

We now define a second-order partial differential operator Δ_p to be the combination

$$\Delta_p = \delta d + d\delta, \tag{4.126}$$

acting on p-forms This maps a p-form to a p-form. A slightly tedious calculation in cartesian co-ordinates will show that, for flat space,

$$\Delta_p = -\nabla^2 \tag{4.127}$$

on each component of a p-form. This Δ_p is therefore the natural definition for (minus) the Laplacian acting on differential forms. It is usually called the *Laplace-Beltrami* operator.

Using $\langle a, db\rangle = \langle \delta a, b\rangle$ we have

$$\langle (\delta d + d\delta)a, b\rangle_p = \langle \delta a, \delta b\rangle_{p-1} + \langle da, db\rangle_{p+1} = \langle a, (\delta d + d\delta)b\rangle_p, \tag{4.128}$$

and so we deduce that Δ_p is self-adjoint on $\Omega^p(M)$. The middle terms in (4.128) are both positive, so we also see that Δ_p is a positive operator — *i.e.* all its eigenvalues are positive or zero.

Suppose that $\Delta_p a = 0$, then (4.128) for $a = b$ becomes that

$$0 = \langle \delta a, \delta a\rangle_{p-1} + \langle da, da\rangle_{p+1}. \tag{4.129}$$

Because both these inner products are positive or zero, the vanishing of their sum requires them to be individually zero. Thus $\Delta_p a = 0$ implies that $da = \delta a = 0$. By analogy with harmonic functions, we call a form that is annihilated by Δ_p a *harmonic form*. Recall that a form a is closed if $da = 0$. We correspondingly say that a is *co-closed* if δa=0. A differential form is therefore harmonic if and only if it is both closed and co-closed.

When a self-adjoint operator A is Fredholm (*i.e* the solutions of the equation $Ax = y$ are governed by the Fredholm alternative) the vector space on which it acts is decomposed into a direct sum of the kernel and range of the operator

$$V = \mathrm{Ker}\,(A) \oplus \mathrm{Im}\,(A). \tag{4.130}$$

It may be shown that our Laplace-Beltrami Δ_p is a Fredholm operator, and so for any p-form ω there is an η such that ω can be written as

$$\begin{aligned} \omega &= (d\delta + \delta d)\eta + \gamma \\ &= d\alpha + \delta\beta + \gamma, \end{aligned} \tag{4.131}$$

where $\alpha = \delta\eta$, $\beta = d\eta$, and γ is harmonic. This result is known as the *Hodge decomposition* of ω. It is a form-language generalization of the of the Hodge-Weyl and Helmholtz-Hodge decompositions of chapter **??**. It is easy to see that α, β and γ are uniquely determined by ω. If they were not then we could find some α, β and γ such that

$$0 = d\alpha + \delta\beta + \gamma \tag{4.132}$$

with non-zero $d\alpha$, $\delta\beta$ and γ. To see that this is not possible, take the d of (4.132) and then the inner product of the result with β. Because $d(d\alpha) = d\gamma = 0$, we end up with

$$\begin{aligned} 0 &= \langle \beta, d\delta\beta \rangle \\ &= \langle \delta\beta, \delta\beta \rangle. \end{aligned} \tag{4.133}$$

Thus $\delta\beta = 0$. Now apply δ to the two remaining terms of (4.132) and take an inner product with α. Because $\delta\gamma = 0$, we find $\langle d\alpha, d\alpha \rangle = 0$, and so $d\alpha = 0$. What now remains of (4.132) asserts that $\gamma = 0$.

Suppose that ω is closed. Then our strategy of taking the d of the decomposition

$$\omega = d\alpha + \delta\beta + \gamma, \tag{4.134}$$

followed by an inner product with β leads to $\delta\beta = 0$. A closed form can thus be decomposed as

$$\omega = d\alpha + \gamma \tag{4.135}$$

with α and γ unique. Each cohomology class in $H^p(M)$ therefore contains a unique harmonic representative. Since any harmonic function is closed, and hence a representative of some cohomology class, we conclude that there is a 1-1 correspondence between p-form solutions of Laplace's equation and elements of $H^p(M)$. In particular

$$\dim(\mathrm{Ker}\,\Delta_{\mathrm{p}}) = \dim\left(H^p(M)\right) = b_p. \tag{4.136}$$

Here b_p is the p-th Betti number. From this we immediately deduce that

$$\chi(M) = \sum_{p=0}^{D} (-1)^p \dim(\mathrm{Ker}\,\Delta_{\mathrm{p}}), \tag{4.137}$$

where $\chi(M)$ is the Euler character of M. There is therefore an intimate relationship between the null-spaces of the second-order partial differential operators Δ_p and the global topology of the manifold in which they live. This is an example of an *index theorem*.

Just as for the ordinary Laplace operator, Δ_p has a complete set of eigenfunctions with associated eigenvalues λ. Because the the manifold is compact and hence has finite volume, the spectrum will be discrete. Remarkably, the topological influence we uncovered above is restricted to the zero-eigenvalue spaces. Suppose that we have a p-form eigenfunction u_λ for Δ_p:

$$\Delta_p u_\lambda = \lambda u_\lambda. \tag{4.138}$$

Then

$$\begin{aligned} \lambda\, du_\lambda &= d\,\Delta_p u_\lambda \\ &= d(d\delta + \delta d)u_\lambda \\ &= (d\delta)du_\lambda \\ &= (\delta d + d\delta)du_\lambda \\ &= \Delta_{p+1} du_\lambda. \end{aligned} \tag{4.139}$$

Thus, provided it is not identically zero, du_λ is an $(p+1)$-form eigenfunction of $\Delta_{(p+1)}$ with eigenvalue λ. Similarly, δu_λ is a $(p-1)$-form eigenfunction also with eigenvalue λ.

Can du_λ be zero? Yes! It will certainly be zero if u_λ itself is the d of something. What is less obvious is that it will be zero *only* if it is the d of something. To see this suppose that $du_\lambda = 0$ and $\lambda \neq 0$. Then

$$\lambda u_\lambda = (\delta d + d\delta)u_\lambda = d(\delta u_\lambda). \tag{4.140}$$

Thus $du_\lambda = 0$ implies that $u_\lambda = d\eta$, where $\eta = \delta u_\lambda/\lambda$. We see that for λ non-zero, the operators d and δ map the λ eigenspaces of Δ into one another, and the kernel of d acting on p-form eigenfunctions is precisely the image of d acting on $(p-1)$-form eigenfunctions. In other words, when restricted to positive λ eigenspaces of Δ, the cohomology is trivial.

The set of spaces V_p^λ together with the maps $d : V_p^\lambda \to V_{p+1}^\lambda$ therefore constitute an exact sequence when $\lambda \neq 0$, and so the alternating sum of their dimension must be zero. We have therefore established that

$$\sum_p (-1)^p \dim V_p^\lambda = \begin{cases} \chi(M), & \lambda = 0, \\ 0, & \lambda \neq 0. \end{cases} \tag{4.141}$$

All the topology resides in the null-spaces, therefore.

Exercise 4.6: Show that if ω is closed and co-closed then so is $\star\omega$. Deduce that in a for a compact orientable D-manifold we have $b_p = b_{D-p}$. This observation therefore gives another way of understanding Poincaré duality.

4.7.2 Morse Theory

Suppose, as in the previous section, M is a D-dimensional compact manifold without boundary and $V : M \to \mathbb{R}$ a smooth function. The global topology of M imposes some constraints on the possible maxima, minima and saddle points of V. Suppose that P is a stationary point of V. Taking co-ordinates such that P is at $x^\mu = 0$, we can expand

$$V(x) = V(0) + \frac{1}{2} H_{\mu\nu} x^\mu x^\nu + \dots . \tag{4.142}$$

Here, the matrix $H_{\mu\nu}$ is the *Hessian*

$$H_{\mu\nu} = \left. \frac{\partial^2 V}{\partial x^\mu \partial x^\nu} \right|_0 . \tag{4.143}$$

We can change co-ordinates so as reduce the Hessian to a canonical form with only $\pm 1, 0$ on the diagonal:

$$H_{\mu\nu} = \begin{pmatrix} -I_m & & \\ & I_n & \\ & & 0_{D-m-n} \end{pmatrix} . \tag{4.144}$$

If there are no zero's on the diagonal then the stationary points is said to be *non-degenerate*. The the number m of downward-bending directions is then called the *index* of V at P. If P were a local maximum, then $m = D$, $n = 0$. If it were a local minimum then $m = 0$, $n = D$. When all its stationary points are non-degenerate, V is said to be a *Morse function*. This is the

generic case. Degenerate stationary points can be regarded as arising from the merging of two or more non-degenerate points.

The *Morse index theorem* asserts that if V is a Morse function, and if we define N_0 to be the number of stationary points with index 0 (*i.e.* local minima), and N_1 to be the number of stationary points with index 1 *etc.*, then

$$\sum_{m=0}^{D}(-1)^m N_m = \chi(M). \tag{4.145}$$

Here $\chi(M)$ is the Euler character of M. Thus, a function on the two-dimensional torus, which has $\chi = 0$, can have a local maximum, a local minimum and two saddle points, but cannot have only one local maximum, one local minimum and no saddle points. On a two-sphere ($\chi = 2$), if V has one local maximum and one local minimum it can have no saddle points.

Closely related to the Morse index theorem is the *Poincaré-Hopf theorem.* It counts the isolated zeros of a tangent-vector field X on a compact D-manifold and, among other things, explains why we cannot comb a hairy ball. An *isolated zero* is a point z_n at which X becomes zero, and that has a neighbourhood in which there is no other zero. If there are only finitely many zeros then each of them will be isolated. We can define a *vector field index* at z_n by surrounding it with a small $(D-1)$-sphere on which X does not vanish. The direction of X at each point on this sphere then provides a map from the sphere to itself. The index $i(z_n)$ is defined to be the winding number (Brouwer degree) of this map. The index can be any integer, but in the special case that X is the gradient of a Morse function we have $i(z_n) = (-1)^{m_n}$ where m is the Morse index at z_n.

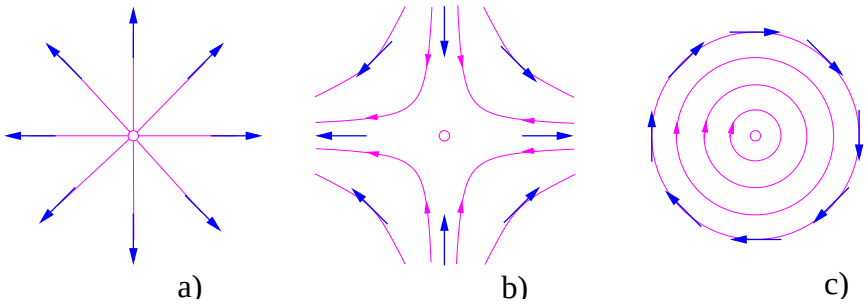

Figure 4.13: *Two-dimensional vector-fields and their streamlines near zeros with indices a)* $i(z_a) = +1$, *b)* $i(z_b) = -1$, *c)* $i(z_c) = +1$.

The Poincaré-Hopf theorem now states that, for a compact manifold without boundary, and for a tangent vector field with only finitely many zeros,

$$\sum_{\text{zeros } n} i(z_n) = \chi(M). \tag{4.146}$$

A tangent-vector field must therefore always have at least one zero unless $\chi(M) = 0$. Since the two-sphere has $\chi = 2$, it cannot be combed.

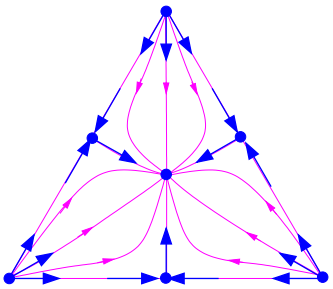

Figure 4.14: *Gradient vector field and streamilines in a two-simplex.*

If one is prepared to believe that $\sum_{\text{zeros}} i(z_n)$ is the same integer for all tangent vector fields X on M, it is simple to show that this integer must be equal to the Euler character of M. Consider, for ease of visualization, a two-manifold. Triangulate M and take X to be the gradient field of a function with local minima at each vertices, saddle points on the edges, and local maxima at the centre of each face (see figure 4.14). It must be clear that this particular field X has

$$\sum_{\text{zeros } n} i(z_n) = V - E + F = \chi(M). \tag{4.147}$$

In the case of a two-dimensional oriented surface equipped with a smooth metric, it is also simple to demonstrate the invariance of the index sum. Consider two vector fields X and Y. Triangulate M so that all zeros of both fields lie in the interior of the faces of the simplices. The metric allows us to compute the angle θ between X and Y wherever they are both non-zero, and in particular on the edges of the simplices. For each two-simplex σ we compute the total change $\Delta\theta$ in the angle as we circumnavigate its boundary. This change is an integral multiple of 2π, with the integer counting the difference

$$\sum_{\text{zeros of } X \in \sigma} i(z_n) - \sum_{\text{zeros of } Y \in \sigma} i(z_n) \tag{4.148}$$

of the indices of the zeros within σ. On summing over all triangles σ, each edge is traversed twice, once in each direction, so $\sum_\sigma \Delta\theta$ vanishes . The total index of X is therefore the same as that of Y.

This pairwise cancellation argument can be extended to non-orientable surfaces, such as the projective plane, In this case the edges constituting the homological "boundary" of the closed surface are traversed twice in the *same* direction, but the angle θ at a point on one edge is paired with $-\theta$ at the corresponding point of the other edge.

Supersymmetric Quantum Mechanics

Edward Witten gave a beautiful proof of the Morse index theorem for an orientable manifold by re-interpreting the Laplace-Beltrami operator as the Hamiltonian of *supersymmetric quantum mechanics* on M. Witten's idea had a profound impact, and led to quantum physics serving as a rich source of inspiration and insight for mathematicians. We have seen most of the ingredients of this re-interpretation in previous chapters. Indeed you should have experienced a sense of *déjà vu* when you saw d and δ mapping eigenfunctions of one differential operator into eigenfunctions of a related operator.

We begin with an novel way to think of the calculus of differential forms. We introduce a set of fermion annihilation and creation operators ψ^μ and $\psi^{\dagger\mu}$ which anti-commute, $\psi^\mu\psi^\nu = -\psi^\nu\psi^\mu$, and obey

$$\{\psi^{\dagger\mu}, \psi^\nu\} \equiv \psi^{\dagger\mu}\psi^\nu + \psi^\nu\psi^{\dagger\mu} = g^{\mu\nu}. \tag{4.149}$$

Here μ runs from 1 to D. As is usual when we are given such operators, we also introduce a *vacuum state* $|0\rangle$ which is killed by all the annihilation operators: $\psi^\mu|0\rangle = 0$. The states

$$(\psi^{\dagger 1})^{p_1}(\psi^{\dagger 2})^{p_2}\ldots(\psi^{\dagger n})^{p_n}|0\rangle, \tag{4.150}$$

with each of the p_i taking the value one or zero, then constitute a basis for 2^D-dimensional space. We call $p = \sum_i p_i$ the *fermion number* of the state. We now assume that $\langle 0|0\rangle = 1$ and use the anti-commutation relations to show that

$$\langle 0|\psi^{\mu_p}\ldots\psi^{\mu_2}\psi^{\mu_1}\ldots\psi^{\dagger\nu_1}\psi^{\dagger\nu_2}\ldots\psi^{\dagger\nu_q}|0\rangle$$

is zero unless $p = q$, in which case it is equal to

$$g^{\mu_1\nu_1}g^{\mu_2\nu_2}\ldots g^{\mu_p\nu_p} \pm (\text{permutations}).$$

We now make the correspondence

$$\frac{1}{p!}f_{\mu_1\mu_2\ldots\mu_p}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\ldots\psi^{\dagger\mu_p}|0\rangle \leftrightarrow \frac{1}{p!}f_{\mu_1\mu_2\ldots\mu_p}(x)dx^{\mu_1}dx^{\mu_2}\ldots dx^{\mu_p}, \tag{4.151}$$

to identify p-fermion states with p-forms. We think of $f_{\mu_1\mu_2\ldots\mu_p}(x)$ as being the wavefunction of a particle moving on M, with the subscripts informing us there are fermions occupying the states μ_i. It is then natural to take the inner product of

$$|a\rangle = \frac{1}{p!}a_{\mu_1\mu_2\ldots\mu_p}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\ldots\psi^{\dagger\mu_p}|0\rangle \tag{4.152}$$

and

$$|b\rangle = \frac{1}{q!}b_{\mu_1\mu_2\ldots\mu_q}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\ldots\psi^{\dagger\mu_q}|0\rangle \tag{4.153}$$

to be

$$\begin{aligned}\langle a,b\rangle &= \int_M d^Dx\sqrt{g}\,\frac{1}{p!q!}\,a^*_{\mu_1\mu_2\ldots\mu_p}b_{\nu_1\nu_2\ldots\nu_q}\langle 0|\psi^{\mu_p}\ldots\psi^{\mu_1}\psi^{\dagger\nu_1}\ldots\psi^{\dagger\nu_q}|0\rangle \\ &= \delta_{pq}\int_M d^Dx\sqrt{g}\,\frac{1}{p!}\,a^*_{\mu_1\mu_2\ldots\mu_p}b^{\mu_1\mu_2\ldots\mu_p}.\end{aligned} \tag{4.154}$$

This coincides the Hodge inner product of the corresponding forms.

If we lower the index by setting ψ_μ to be $g_{\mu\nu}\psi^\mu$ then the action of $X^\mu\psi_\mu$ on a p-fermion state coincides with the action of the interior multiplication i_X on the corresponding p-form. All the other operations of the exterior calculus can also be expressed in terms of the ψ's. In particular, in Cartesian co-ordinates where $g_{\mu\nu}=\delta_{\mu\nu}$, we can identify d with $\psi^{\dagger\mu}\partial_\mu$. To find the operator that corresponds to the Hodge δ, we compute

$$\delta = d^\dagger = (\psi^{\dagger\mu}\partial_\mu)^\dagger = \partial^\dagger_\mu\psi^\mu = -\partial_\mu\psi^\mu = -\psi^\mu\partial_\mu. \tag{4.155}$$

The hermitian adjoint of ∂_μ is here being taken with respect to the standard $L^2(\mathbb{R}^D)$ inner product. This computation becomes more complicated when when $g_{\mu\nu}$ becomes position dependent. The adjoint $\partial^\dagger_\mu$ then involves the derivative of $\sqrt{g}$, and ψ and ∂_μ no longer commute. For this reason, and because such complications are inessential for what follows, we will delay discussing this general case until the end of this section.

Having found a simple formula for δ, it is now automatic to compute

$$d\delta + \delta d = -\{\psi^{\dagger\mu},\psi^\nu\}\,\partial_\mu\partial_\nu = -\delta^{\mu\nu}\partial_\mu\partial_\nu = -\nabla^2. \tag{4.156}$$

This much easier than deriving the same result by using $\delta = (-1)^{Dp+D+1} \star d \star$.

Witten's fermionic formalism simplifies a number of compuations involving δ, but his real innovation was to consider a *deformation* of the exterior calculus by introducing the operators

$$d_t = e^{-tV(x)} d\, e^{tV(x)}, \quad \delta_t = e^{tV(x)} \delta\, e^{-tV(x)}, \tag{4.157}$$

and

$$\Delta_t = d_t \delta_t + \delta_t d_t. \tag{4.158}$$

Here $V(x)$ is the Morse function whose stationary points we are seeking to count.

The deformed derivative continues to obey $d_t^2 = 0$, and $d\omega = 0$ if and only if $d_t e^{-tV} \omega = 0$. Similarly, if $\omega = d\eta$ then $e^{-tV}\omega = d_t e^{-tV}\eta$. The cohomology of d and d_t are therefore transformed into each other by multiplication by e^{-tV}. Since the exponential function is never zero, this correspondence is invertible and the mapping is an isomorphism. In particular, the Betti numbers b_p, the dimensions of $\mathrm{Ker}\,(d_t)_p/\mathrm{Im}\,(d_t)_{p-1}$, are t independent. Further, the t-deformed Laplace-Beltrami operator remains Fredholm with only positive or zero eigenvalues. We can make a Hodge decomposition

$$\omega = d_t \alpha + \delta_t \beta + \gamma, \tag{4.159}$$

where $\Delta_t \gamma = 0$, and concude that

$$\dim\left(\mathrm{Ker}\,(\Delta_t)_p\right) = b_p \tag{4.160}$$

as before. The non-zero eigenvalue spaces will also continue to form exact sequences. Nothing seems to have changed! Why do we introduce d_t then? The motivation is that when t becomes large we can use our knowledge of quantum mechanics to compute the Morse index.

To do this, we expand out

$$\begin{aligned} d_t &= \psi^{\dagger\mu}(\partial_\mu + t\partial_\mu V) \\ \delta_t &= -\psi^\mu(\partial_\mu - t\partial_\mu V) \end{aligned} \tag{4.161}$$

and find

$$d_t\delta_t + \delta_t d_t = -\nabla^2 + t^2|\nabla V|^2 + t[\psi^{\dagger\mu}, \psi^\nu]\,\partial^2_{\mu\nu} V. \tag{4.162}$$

This can be thought of as a Schrödinger Hamiltonian on M containing a potential and a fermionic term. When t is large and positive the potential

$t^2|\nabla V|^2$ will be large everywhere except near those points where $\nabla V = 0$. The wavefunctions of all low-energy states, and in particular all zero-energy states, will therefore be concentrated at precisely the stationary points we are investigating. Let us focus on a particular stationary point, which we will take as the origin of our co-ordinate system, and identify any zero-energy state localized there. We first rotate the coordinate system about the origin so that the Hessian matrix $\partial^2_{\mu\nu} V|_0$ becomes diagonal with eigenvalues λ_n. The Schrödinger problem can then be approximated by a sum of harmonic oscillator hamiltonians

$$\Delta_{p,t} \approx \sum_{i=1}^{D} \left\{ -\frac{\partial^2}{\partial x_i^2} + t^2 \lambda_i^2 x_i^2 + t\lambda_i [\psi^{\dagger i}, \psi^i] \right\}. \tag{4.163}$$

The commutator $[\psi^{\dagger i}, \psi^i]$ takes the value $+1$ if the i'th fermion state is occupied, and -1 if it is not. The spectrum of the approximate Hamiltonian is therefore

$$t \sum_{i=1}^{D} \{|\lambda_i|(1 + 2n_i) \pm \lambda_i\}. \tag{4.164}$$

Here the n_i label the harmonic oscillator states. The lowest energy states will have all the $n_i = 0$. To get a state with zero energy we must arrange for the $\pm$ sign to be negative (no fermion in state i) whenever λ_i is positive, and to be positive (fermion state i occupied) whenever λ_i is negative. The fermion number "p" of the zero-energy state is therefore equal to the the number of negative λ_i — *i.e.* to the index of the critical point! We can, in this manner, find one zero-energy state for each critical point. All other states have energies proportional t, and therefore large. Since the number of zero energy states having fermion number p is the Betti number b_p, the harmonic oscillator approximation suggests that $b_p = N_p$.

If we could trust our computation of the energy spectrum, we would have established the Morse theorem

$$\sum_{p=0}^{D} (-1)^p N_p = \sum_{p=0}^{D} (-1)^p b_p = \chi(M), \tag{4.165}$$

by having the two sums agree term by term. Our computation is only approximate, however. While there can be no more zero-energy states than those we have found, some states that appear to be zero modes may instead

have small positive energy. This might arise from tunnelling between the different potential minima, or from the higher-order corrections to the harmonic oscillator potentials, both effects we have neglected. We can therefore only be confident that

$$N_p \geq b_p. \tag{4.166}$$

The remarkable thing is that, for the Morse index, *this does not matter*! If one of our putative zero modes gains a small positive energy, it is now in the non-zero eigenvalue sector of the spectrum. The exact-sequence property therefore tells us that one of the other putative zero modes must also be a not-quite-zero mode state with exactly the same energy. This second state will have a fermion number that differs from the first by plus or minus one. Our error in counting the zero energy states therefore cancels out when we take the alternating sum. Our unreliable estimate $b_p \approx N_p$ has thus provided us with an *exact* computation of the Morse index.

We have described Witten's argument as if the manifold M were flat. When the manifold M is not flat, however, the curvature will not affect our computations. Once the parameter t is large the low-energy eigenfunctions will be so tightly localized about the critical points that they will be hard-pressed to detect the curvature. Even if the curvature can effect an infintesimal energy shift, the exact-sequence argument again shows that this does not affect the alternating sum.

The Weitzenböck Formula

Although we we were able to evade them when proving the Morse index theorem, it is interesting to uncover the workings of the nitty-gritty Riemann tensor index machinary that lie concealed behind the polished facade of Hodge's d, δ calculus.

Let us assume that our manifold M is equipped with a torsion-free connection $\Gamma^{\mu}{}_{\nu\lambda} = \Gamma^{\mu}{}_{\lambda\nu}$, and use this connection to define the action of an operator $\hat{\nabla}_\mu$ by specifying its commutators with c-number functions f, and with the ψ^μ and $\psi^{\dagger\mu}$'s:

$$\begin{aligned} [\hat{\nabla}_\mu, f] &= \partial_\mu f, \\ [\hat{\nabla}_\mu, \psi^{\dagger\nu}] &= -\Gamma^{\nu}{}_{\mu\lambda}\psi^{\dagger\lambda}, \\ [\hat{\nabla}_\mu, \psi^{\nu}] &= -\Gamma^{\nu}{}_{\mu\lambda}\psi^{\lambda}. \end{aligned} \tag{4.167}$$

We also set $\hat{\nabla}_\mu|0\rangle = 0$. These rules allow us to compute the action of $\hat{\nabla}_\mu$ on $f_{\mu_1\mu_2\ldots\mu_p}(x)\psi^{\dagger\mu_1}\ldots\psi^{\dagger\mu_p}|0\rangle$. For example

$$\begin{aligned}\hat{\nabla}_\mu\left(f_\nu\psi^{\dagger\nu}|0\rangle\right) &= \left([\hat{\nabla}_\mu, f_\nu\psi^{\dagger\nu}] + f_\nu\psi^{\dagger\nu}\hat{\nabla}_\mu\right)|0\rangle \\ &= \left([\hat{\nabla}_\mu, f_\nu]\psi^{\dagger\nu} + f_\alpha[\hat{\nabla}_\mu, \psi^{\dagger\alpha}]\right)|0\rangle \\ &= (\partial_\mu f_\nu - f_\alpha\Gamma^\alpha{}_{\mu\nu})\psi^{\dagger\nu}|0\rangle \\ &= (\nabla_\mu f_\nu)\,\psi^{\dagger\nu}|0\rangle, \end{aligned} \tag{4.168}$$

where

$$\nabla_\mu f_v = \partial_\mu f_\nu - \Gamma^\alpha{}_{\mu\nu} f_\alpha, \tag{4.169}$$

is the usual covariant derivative acting on the componenents of a covariant vector.

The metric $g^{\mu\nu}$ counts as a c-number function, and so $[\hat{\nabla}_\alpha, g^{\mu\mu}]$ is not zero, but is instead $\partial_\alpha g^{\mu\nu}$. This might be disturbing—being able pass the metric through a covariant derivative is a basic compatibilty condition in Riemann geometry—but all is not lost. $\hat{\nabla}_\mu$ (with a caret) is not quite the same beast as ∇_μ. We proceed as follows:

$$\begin{aligned}\partial_\alpha g^{\mu\nu} &= [\hat{\nabla}_\alpha, g^{\mu\mu}] \\ &= [\hat{\nabla}_\alpha, \{\psi^{\dagger\mu}, \psi^\nu\}] \\ &= [\hat{\nabla}_\alpha, \psi^{\dagger\mu}\psi^\nu] + [\hat{\nabla}_\alpha, \psi^\nu\psi^{\dagger\mu},] \\ &= -\{\psi^{\dagger\mu}, \psi^\lambda\}\Gamma^\nu{}_{\alpha\lambda} - \{\psi^{\dagger\nu}, \psi^\lambda\}\Gamma^\mu{}_{\alpha\lambda} \\ &= -g^{\mu\lambda}\,\Gamma^\nu{}_{\alpha\lambda} - g^{\nu\lambda}\,\Gamma^\mu{}_{\alpha\lambda}. \end{aligned} \tag{4.170}$$

We conclude that

$$\partial_\alpha g^{\mu\nu} + g^{\mu\lambda}\Gamma^\nu{}_{\alpha\lambda} + g^{\lambda\nu}\Gamma^\mu{}_{\alpha\lambda} \equiv \nabla_\alpha g^{\mu\nu} = 0. \tag{4.171}$$

Metric compatibility is therefore satisfied, and the connection is therefore the standard Riemannian

$$\Gamma^\alpha{}_{\mu\nu} = \frac{1}{2}g^{\alpha\lambda}\left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}\right). \tag{4.172}$$

Knowing this, we can compute the adjoint of $\hat{\nabla}_\mu$:

$$\begin{aligned}\left(\hat{\nabla}_\mu\right)^\dagger &= -\frac{1}{\sqrt{g}}\hat{\nabla}_\mu\sqrt{g} \\ &= -\left(\hat{\nabla}_\mu + \partial_\mu \ln\sqrt{g}\right) \\ &= -(\hat{\nabla}_\mu + \Gamma^\nu{}_{\nu\mu}). \end{aligned} \tag{4.173}$$

That $\Gamma^{\nu}{}_{\nu\mu}$ is the logarithmic derivative of $\sqrt{g}$ is a standard identity for the Riemann connection (see exercise 2.14). The resultant formula for $(\hat{\nabla}_\mu)^\dagger$ can be used to verify that the second and third equations in (4.167) are compatible with each other.

We can also compute $[[\hat{\nabla}_\mu, \hat{\nabla}_\nu], \psi^\alpha]$ and from it deduce that

$$[\hat{\nabla}_\mu, \hat{\nabla}_\nu] = R_{\sigma\lambda\mu\nu}\psi^{\dagger\sigma}\psi^\lambda, \tag{4.174}$$

where

$$R^{\alpha}{}_{\beta\mu\nu} = \partial_\mu\Gamma^{\alpha}{}_{\beta\nu} - \partial_\nu\Gamma^{\alpha}{}_{\beta\mu} + \Gamma^{\alpha}{}_{\lambda\mu}\Gamma^{\lambda}{}_{\beta\nu} - \Gamma^{\alpha}{}_{\lambda\nu}\Gamma^{\lambda}{}_{\beta\mu} \tag{4.175}$$

is the Riemann curvature tensor.

We now define d to be

$$d = \psi^{\dagger\mu}\hat{\nabla}_\mu. \tag{4.176}$$

Its action coincides with the usual d because the symmetry of the $\Gamma^{\alpha}_{\mu\nu}$'s ensures that their contributions cancel. From this we find that δ is

$$\begin{aligned}
\delta &\equiv \left(\psi^{\dagger\mu}\hat{\nabla}_\mu\right)^\dagger \\
&= \hat{\nabla}^\dagger_\mu\,\psi^\mu \\
&= -(\hat{\nabla}_\mu + \Gamma^{\nu}{}_{\mu\nu})\psi^\mu \\
&= -\psi^\mu(\hat{\nabla}_\mu + \Gamma^{\nu}{}_{\mu\nu}) + \Gamma^{\mu}{}_{\mu\nu}\psi^\nu \\
&= -\psi^\mu\hat{\nabla}_\mu.
\end{aligned} \tag{4.177}$$

The Laplace-Beltrami operator can now be worked out as

$$\begin{aligned}
d\delta + \delta d &= -\left(\psi^{\dagger\mu}\hat{\nabla}_\mu\psi^\nu\hat{\nabla}_\nu + \psi^\nu\hat{\nabla}_\nu\psi^{\dagger\mu}\hat{\nabla}_\mu\right) \\
&= -\left(\{\psi^{\dagger\mu}, \psi^\nu\}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^{\sigma}{}_{\mu\nu}\hat{\nabla}_\sigma) + \psi^\nu\psi^{\dagger\mu}[\hat{\nabla}_\nu, \hat{\nabla}_\mu]\right) \\
&= -\left(g^{\mu\nu}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^{\alpha}{}_{\mu\nu}\hat{\nabla}_\sigma) + \psi^\nu\psi^{\dagger\mu}\psi^{\dagger\sigma}\psi^\lambda R_{\sigma\lambda\nu\mu}\right)
\end{aligned} \tag{4.178}$$

By making use of the symmetries $R_{\sigma\lambda\nu\mu} = R_{\nu\mu\sigma\lambda}$ and $R_{\sigma\lambda\nu\mu} = -R_{\sigma\lambda\mu\nu}$ we can tidy up the curvature term to get

$$d\delta + \delta d = -g^{\mu\nu}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^{\sigma}{}_{\mu\nu}\hat{\nabla}_\sigma) - \psi^{\dagger\alpha}\psi^\beta\psi^{\dagger\mu}\psi^\nu R_{\alpha\beta\mu\nu}. \tag{4.179}$$

This result is called the *Weitzenböck formula*. An equivalent formula can be derived directly from (4.124), but only with a great deal more effort. The part

without the curvature tensor is called the *Bochner Laplacian*. It is normally written as $B = -g^{\mu\nu}\nabla_\mu\nabla_\nu$ with ∇_μ being understood to be acting on the index ν, and therefore tacitly containing the extra $\Gamma^\sigma_{\mu\nu}$ that must be made explicit when we define the action of $\hat{\nabla}_\mu$ *via* commutators. The Bochner Laplacian can also be written as

$$B = \hat{\nabla}^\dagger_\mu \, g^{\mu\nu} \, \hat{\nabla}_\nu \tag{4.180}$$

which shows that it is a positive operator.

Chapter 5

Groups and Group Representations

Groups appear in physics as symmetries of the system we are studying. Often the symmetry operation involves a linear transformation, and this naturally leads to the idea of finding sets of matrices having the same multiplication table as the group. These sets are called *representations* of the group. Given a group, we endeavour to find and classify all possible representations.

5.1 Basic Ideas

We begin with a rapid review of basic group theory.

5.1.1 Group Axioms

A *group* G is a set with a binary operation that assigns to each ordered pair (g_1, g_2) of elements a third element, g_3, usually written with multiplicative notation as $g_3 = g_1 g_2$. The binary operation, or *product*, obeys the following rules:

i) Associativity: $g_1(g_2 g_3) = (g_1 g_2) g_3$.

ii) Existence of an identity: There is an element[1] $e \in G$ such that $eg = g$ for all $g \in G$.

[1]The symbol "e" is often used for the identity element, from the German *Einheit*, meaning "unity."

iii) Existence of an inverse: For each $g \in G$ there is an element g^{-1} such that $g^{-1}g = e$.

From these axioms there follow some conclusions that are so basic that they are often included in the axioms themselves, but since they are not independent, we state them as corollaries.

Corollary i): $gg^{-1} = e$.
Proof: Start from $g^{-1}g = e$, and multiply on the right by g^{-1} to get $g^{-1}gg^{-1} = eg^{-1} = g^{-1}$, where we have used the left identity property of e at the last step. Now multiply on the left by $(g^{-1})^{-1}$, and use associativity to get $gg^{-1} = e$.

Corollary ii): $ge = g$.
Proof: Write $ge = g(g^{-1}g) = (gg^{-1})g = eg = g$.

Corollary iii): The identity e is unique.
Proof: Suppose there is another element e_1 such that $e_1 g = eg = g$. Multiply on the right by g^{-1} to get $e_1 e = e^2 = e$, but $e_1 e = e_1$, so $e_1 = e$.

Corollary iv): The inverse of a given element g is unique.
Proof: Let $g_1 g = g_2 g = e$. Use the result of corollary (i), that any left inverse is also a right inverse, to multiply on the right by g_1^{-1}, and so find that $g_1 = g_2$.

Two elements g_1 and g_2 are said to *commute* if $g_1 g_2 = g_2 g_1$. If the group has the property that $g_1 g_2 = g_2 g_1$ for all $g_1, g_2 \in G$, it is said to be *Abelian*, otherwise it is *non-Abelian*.

If the set G contains only finitely many elements, the group G is said to be *finite*. The number of elements in the group, $|G|$, is called the *order* of the group.

Examples of Groups:

1) The integers $\mathbb{Z}$ under addition. The binary operation is $(n, m) \mapsto n+m$, and "0" plays the role of the identity element. This is not a finite group.
2) The integers modulo n under addition. $(m, m') \mapsto m+m'$, $\bmod n$. This group is denoted by $\mathbb{Z}_n$.
3) The non-zero integers modulo p (a prime) under *multiplication* $(m, m') \mapsto mm'$, $\bmod p$. Here "1" is the identity element. If the modulus is not a prime number, we do not get a group (why not?). This group is sometimes denoted by $(\mathbb{Z}_p)^\times$.

4) The set of numbers $\{2, 4, 6, 8\}$ under multication modulo 10. Here, the number "6" plays the role of the identity!
5) The set of functions

$$f_1(z) = z, \quad f_2(z) = \frac{1}{1-z}, \quad f_3(z) = \frac{z-1}{z}$$
$$f_4(z) = \frac{1}{z}, \quad f_5(z) = 1-z, \quad f_6(z) = \frac{z}{z-1}$$

with $(f_i, f_j) \mapsto f_i \circ f_j$. Here the "$\circ$" is a standard notation for composition of functions: $(f_i \circ f_j)(z) = f_i(f_j(z))$.
6) The set of rotations in three dimensions, equivalently the set of 3-by-3 real matrices O, obeying $O^T O = I$, and $\det O = 1$. This is the group SO(3). SO(n) is defined analogously as the group of rotations in n dimensions. If we relax the condition on the determinant we get the *orthogonal group* O(n). Both SO(n) and O(n) are examples of *Lie groups*. A Lie group a group that is also a manifold M, and whose multiplication law is a smooth function $M \times M \to M$.
7) Groups are often specified by giving a list of *generators* and *relations*. For example the *cyclic group* of order n, denoted by C_n, is specified by giving the generator a and relation $a^n = e$. Similarly, the *dihedral group* D_n has two generators a, b and relations $a^n = e$, $b^2 = e$, $(ab)^2 = e$. This group has order $2n$.

5.1.2 Elementary Properties

Here are the basic properties of groups that we need:

i) *Subgroups*: If a subset of elements of a group forms a group, it is called a subgroup. For example, $\mathbb{Z}_{12}$ has a subgroup of consisting of $\{0, 3, 6, 9\}$. All groups have at least two subgroups: the trivial subgroups G itself, and $\{e\}$. Any other subgroups are called *proper* subgroups.
ii) *Cosets*: Given a subgroup $H \subseteq G$, having elements $\{h_1, h_2, \ldots\}$, and an element $g \in G$, we form the (left) *coset* $gH = \{gh_1, gh_2, \ldots\}$. If two cosets $g_1 H$ and $g_2 H$ intersect, they coincide. (Proof: if $g_1 h_1 = g_2 h_2$, then $g_2 = g_1(h_1 h_2^{-1})$ and so $g_1 H = g_2 H$.) If H is a finite group, each coset has the same number of distinct elements as H. (Proof: if $gh_1 = gh_2$ then left multiplication by g^{-1} shows that $h_1 = h_2$.) If the

order of G is also finite, the group G is decomposed into an integer number of cosets,

$$G = g_1 H + g_2 H + \cdots, \tag{5.1}$$

where "+"denotes the union of disjoint sets. From this we see that the order of H must divide the order of G. This result is called *Lagrange's theorem*. The set whose elements are the cosets is denoted by G/H.

iii) *Normal subgroups and quotient groups*: A subgroup H of G is said to be *normal*, or *invariant*, if $g^{-1}Hg = H$ for all $g \in G$. Given a normal subgroup H, we can define a multiplication rule on the coset space cosets $G/H \equiv \{g_1 H, g_2 H, \ldots\}$ by taking a representative element from each of $g_i H$, and $g_j H$, taking the product of these elements, and defining $(g_i H)(g_j H)$ to be the coset in which this product lies. This coset is independent of the representative elements chosen (this would not be so if the subgroup was not normal). The resulting group is called the *quotient group* G/H. (Note that the symbol "G/H" is used to denote both the set of cosets, and, when it exists, the group whose elements are these cosets.)

iv) *Simple groups*: A group G with no normal subgroups is said to be *simple*. The finite simple groups have been classified. They fall into various infinite families (Cyclic groups, Alternating groups, 16 families of Lie type) together with 26 *sporadic groups*, the largest of which, the *Monster*, has order 808,017,424,794,512,875,886,459,904,961,710,757,005, 754, 368,000,000,000. The mysterious "Monstrous moonshine" links its representation theory to the elliptic modular function $J(\tau)$ and to string theory.

iv) *Conjugacy and Conjugacy Classes*: Two group elements g_1, g_2 are said to be *conjugate* in G if there is an element $g \in G$ such that $g_2 = g^{-1} g_1 g$. If g_1 is conjugate to g_2, we write $g_1 \sim g_2$. Conjugacy is an *equivalence relation*,[2] and, for finite groups, the resulting *conjugacy classes* have order that divide the order of G. To see this, consider the conjugacy class containing an element g. Observe that the set H of elements $h \in G$ such that $h^{-1}gh = g$ forms a subgroup. The set of elements

[2] An equivalence relation, $\sim$, is a binary relation that is
i) *Reflexive*: $A \sim A$.
ii) *Symmetric*: $A \sim B \iff B \sim A$.
iii) *Transitive*: $A \sim B,\ B \sim C \implies A \sim C$
Such a relation breaks a set up into disjoint *equivalence classes.*

conjugate to g can be identified with the coset space G/H. The order of G divided by the order of the conjugacy class is therefore $|H|$.

Example: In the rotation group SO(3), the conjugacy classes are the sets of rotations through the same angle, but about different axes.

Example: In the group U(n), of n-by-n unitary matrices, the conjugacy classes are the set of matrices possessing the same eigenvalues.

Example: Permutations. The permutation group on n objects, S_n, has order $n!$. Suppose we consider permutations π_1, π_2 in S_8 such that π_1 that maps

$$\pi_1 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 1 & 5 & 4 & 7 & 6 & 8 \end{pmatrix},$$

and π_2 maps

$$\pi_2 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 1 \end{pmatrix}.$$

The product $\pi_2 \circ \pi_1$ then takes

$$\pi_2 \circ \pi_1 : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 4 & 2 & 6 & 5 & 8 & 7 & 1 \end{pmatrix}.$$

We can write these partitions out more compactly by using Paolo Ruffini's cycle notation:

$$\pi_1 = (123)(45)(67)(8), \quad \pi_2 = (12345678), \quad \pi_2 \circ \pi_1 = (132468)(5)(7).$$

In this notation, each number is mapped to the one immediately to its right, with the last number in each bracket, or *cycle*, wrapping round to map to the first. Thus $\pi_1(1) = 2$, $\pi_1(2) = 3$, $\pi_1(3) = 1$. The "8", being both first and last in its cycle, maps to itself: $\pi_1(8) = 8$. Any permutation with this cycle pattern, $(***)(**)(**)(*)$, is in the same conjugacy class as π_1. We say that π_1 possesses one 1-cycle, two 2-cycles, and one 3-cycle. The class $(r_1, r_2, \ldots r_n)$ having r_1 1-cycles, r_2 2-cycles *etc.*, where $r_1 + 2r_2 + \cdots + nr_n = n$, contains

$$N_{(r_1,r_2,\ldots)} = \frac{n!}{1^{r_1}(r_1!)\, 2^{r_2}\,(r_2!) \cdots n^{r_n}\,(r_n!)}$$

elements. The *sign* of the permutation,

$$\operatorname{sgn} \pi = \epsilon_{\pi(1)\pi(2)\pi(3)\ldots\pi(n)}$$

is equal to

$$\operatorname{sgn}\pi = (+1)^{r_1}(-1)^{r_2}(+1)^{r_3}(-1)^{r_4}\cdots.$$

We have, for any two permutations π_1, π_2

$$\operatorname{sgn}(\pi_1)\operatorname{sgn}(\pi_2) = \operatorname{sgn}(\pi_1 \circ \pi_2),$$

so the *even* ($\operatorname{sgn}\pi = +1$) permutations form an invariant subgroup called the *Alternating group*, A_n. The group A_n is simple for $n \geq 5$, and Ruffini (1801) showed that this simplicity prevents the solution of the general quintic by radicals. His work was ignored, however, and later independently rediscovered by Abel (1824) and Galois (1829).

If we write out the group elements in some order $\{e, g_1, g_2, \ldots\}$, and then multiply on the left

$$g\{e, g_1, g_2, \ldots\} = \{g, gg_1, gg_2, \ldots\}$$

then the ordered list $\{g, gg_1, gg_2, \ldots\}$ is a permutation of the original list. Any group is therefore a subgroup of $S_{|G|}$. This is called *Cayley's Theorem*.

Exercise 5.1: Let H_1, H_2 be two subgroups of a group G. Show that $H_1 \cap H_2$ is also a subgroup.

Exercise 5.2: Let G be any group.

a) The subset $Z(G)$ of G consisting of those $g \in G$ that commute with all other elements of the group is called the *center* of the group. Show that $Z(G)$ is a subgroup of G.
b) If g is an element of G, the set $C_G(g)$ of elements of G that commute with g is called the *centeralizer* of g in G. Show that it is a subgroup of G.
c) If H is a subgroup, the set of elements of G that commute with all elements of H is the *centralizer* $C_G(H)$ of H in G. Show that it is a subgroup of G.
d) If H is a subgroup, the set $N_G(H) \subset G$ consisting of those g such that $g^{-1}Hg = H$ is called the *normalizer* of H in G. Show that $N_G(H)$ is a subgroup of G, and that H is a normal subgroup of $N_G(H)$.

Exercise 5.3: Show that the set of powers a^n of an element $a \in G$ form a subgroup. Let p be prime. Recall that the set $\{1, 2, \ldots p-1\}$ forms the group $(\mathbb{Z}_p)^\times$ under multiplication modulo p. By appealing to Lagrange's theorem, prove *Fermat's little theorem* that for any prime p and integer a, we have $a^{p-1} = 1, \bmod p$.

Exercise 5.4: Use Fermat's theorem from the previous excercise to establish the mathematical identity underlying the RSA algorithm for public-key cryptography: Let p, q be prime and $N = pq$. First use Euclid's algorithm for the HCF of two numbers to show that if the integer e is co-prime to[3] $(p-1)(q-1)$, then there is an integer d such that

$$de = 1, \bmod (p-1)(q-1).$$

Then show that if,

$$C = M^e, \bmod N, \qquad \text{(encryption)}$$

then

$$M = C^d, \bmod N. \qquad \text{(decryption)}.$$

The numbers e and N can be made known to the public, but it is hard to find the secret decoding key, d, unless the factors p and q of N are known.

Exercise 5.5: Consider the group $\mathcal{G}$ with multiplication table shown in table 5.1.

$\mathcal{G}$	I	A	B	C	D	E
I	I	A	B	C	D	E
A	A	B	I	E	C	D
B	B	I	A	D	E	C
C	C	D	E	I	A	B
D	D	E	C	B	I	A
E	E	C	D	A	B	I

Table 5.1: *Multiplication table of $\mathcal{G}$. To find AB look in row A column B.*

This group has proper a subgroup $\mathcal{H} = \{I, A, B\}$, and corresponding (left) cosets are $I\mathcal{H} = \{I, A, B\}$ and $C\mathcal{H} = \{C, D, E\}$.

(i) Construct the conjugacy classes of this group.
(ii) Show that $\{I, A, B\}$ and $\{C, D, E\}$ are indeed the left cosets of $\mathcal{H}$.
(iii) Determine whether $\mathcal{H}$ is a normal subgroup.
(iv) If so, construct the group multiplication table for the corresponding quotient group.

[3]Has no factors in common with.

Exercise 5.6: Let H and K, be groups. Make the cartesian product $G = H \times K$ into a group by introducing a multiplication rule for elements of the Cartesian product by setting:

$$(h_1, k_1) * (h_2, k_2) = (h_1 h_2, k_1 k_2).$$

Show that G, equipped with $*$ as its product, satsifies the group axioms. The resultant group is called the *direct product* of H and K.

Exercise 5.7: If F and G are groups, a map $\varphi : F \to G$ that preserves the group structure, *i.e.* if $\varphi(g_1)\varphi(g_2) = \varphi(g_1 g_2)$, is called a group homomorphism. If φ is such a homomorphism show that $\varphi(e_F) = e_G$, where e_F, and e_G are the identity element in F, G respectively.

Exercise 5.8:. If $\varphi : F \to G$ is a group homomorphism, and if we define $\mathrm{Ker}(\varphi)$ as the set of elements $f \in F$ that map to e_G, show that $\mathrm{Ker}(\varphi)$ is a normal subgroup of F.

5.1.3 Group Actions on Sets

Groups usually appear in physics as symmetries: they act on a physical object to change it in some way, perhaps while leaving some other property invariant.

Suppose X is a set. We call its elements "points." A *group action* on X is a map $g \in G : X \to X$ that takes a point $x \in X$ to a new point that we denote by $gx \in X$, and such that $g_2(g_1 x) = (g_1 g_2)x$, and $ex = x$. There is some standard vocabulary for group actions:

i) Given a a point $x \in X$ we define the *orbit* of x to be the set $Gx \equiv \{gx : g \in G\} \subseteq X$.

ii) The action of the group is *transitive* if any orbit is the whole of X.

iii) The action is *effective*, or *faithful*, if the map $g : X \to X$ being the identity map implies that $g = e$. Another way of saying this is that the action is effective if the map $G \to \mathrm{Map}\,(X \to X)$ is one-to-one. If the action of G is *not* faithful, the set of $g \in G$ that act as the identity map forms an invariant subgroup H of G, and the quotient group G/H has a faithful action.

iv) The action is *free* if the existence of an x such that $gx = x$ implies that $g = e$. In this case, we also say that g acts without fixed points.

If the group acts freely and transitively, then having chosen a fiducial point x_0, we can uniquely label every point in X by the group element g such that $x = gx_0$. (If g_1 and g_2 both take $x_0 \to x$, then $g_1^{-1}g_2x_0 = x_0$. By the free action property we deduce that $g_1^{-1}g_2 = e$, and $g_1 = g_2$.). In this case we might, for some purposes, identify X with G.

Suppose the group acts transitively, but not freely. Let H be the set of elements that leaves x_0 fixed. This is clearly a subgroup of G, and if $g_1x_0 = g_2x_0$ we have $g_1^{-1}g_2 \in H$, or $g_1H = g_2H$. The space X can therefore be identified with the space of cosets G/H. Such sets are called *quotient spaces* or *Homogeneous spaces.* Many spaces of significance in physics can be though of as cosets in this way.
Example: The rotation group SO(3) acts transitively on the two-sphere S^2. The SO(2) subgroup of rotations about the z axis, leaves the north pole of the sphere fixed. We can therefore identify $S^2 \simeq \mathrm{SO}(3)/\mathrm{SO}(2)$.

Many phase transitions are a result of *spontaneous symmetry breaking*. For example the water $\to$ ice transition results in the continuous translation invariance of the liquid water being broken down to the discrete translation invariance of the crystal lattice of the solid ice. When a system with symmetry group G spontaneously breaks the symmetry to a subgroup H, the set of inequivalent ground states can be identified with the homogeneous space G/H.

5.2 Representations

An n-dimensional *representation* of a group is formally defined to be a homomorphism from G to a subgroup of $\mathrm{GL}(n, \mathbb{C})$, the group of invertible n-by-n matrices with complex entries. In effect, it is a set of n-by-n matrices that obeys the group multiplication rules

$$D(g_1)D(g_2) = D(g_1g_2), \quad D(g^{-1}) = [D(g)]^{-1}. \tag{5.2}$$

Given such a representation, we can form another one $D'(g)$ by conjugation with any fixed invertible matrix C

$$D'(g) = C^{-1}D(g)C. \tag{5.3}$$

If $D'(g)$ is obtained from $D(g)$ in this way, we say that they are *equivalent* representations and write $D \sim D'$. We can think of D and D' as being

matrices representing the same linear map, but in different bases. Our task in the rest of this chapter is to find and classify all representations of a finite group G up to equivalence.

Real and pseudo-real representations

We can form a new representation from $D(g)$ by setting

$$D'(g) = D^*(g),$$

where $D^*(g)$ denotes the matrix whose entries are the complex conjugates of those in $D(g)$. Suppose $D^* \sim D$. It may then be possible to find a basis in which the matrices have only real entries. In this case we say the representation is *real*. It may be, however, be that $D^* \sim D$ but we cannot find a basis in which the matrices become real. In this case we say that D is *pseudo-real*.

Example: Consider the defining representation of SU(2) (the group of 2-by-2 unitary matrices with unit determinant.) Such matrices are necessarily of the form

$$U = \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix}, \tag{5.4}$$

where a and b are complex numbers with $|a|^2 + |b|^2 = 1$. They are therefore specified by *three* real parameters, and so the group manifold is three dimensional. Now

$$\begin{aligned} \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix}^* &= \begin{pmatrix} a^* & -b \\ b^* & a \end{pmatrix}, \\ &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \\ &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \end{aligned} \tag{5.5}$$

and so $U \sim U^*$. It is not possible to find a basis in which all SU(2) matrices are simultaneously real, however. If such a basis existed we could specify the matrices by only two real parameters—but we have seen that we need three real numbers to describe all possible SU(2) matrices.

Direct Sum and Direct Product

We can obtain new representations from old by combining them.

Given two representations $D^{(1)}(g)$, $D^{(2)}(g)$, we can form their *direct sum* $D^{(1)} \oplus D^{(2)}$ as the block-diagonal matrix

$$\begin{pmatrix} D^{(1)}(g) & 0 \\ 0 & D^{(2)}(g) \end{pmatrix}. \tag{5.6}$$

We are particularly interested in taking a representation and breaking it up as a direct sum of *irreducible* representations.

Given two representations $D^{(1)}(g)$, $D^{(2)}(g)$, we can combine them in a different way by taking their *direct product* $D^{(1)} \otimes D^{(2)}$, the natural action of the group on the tensor product of the representation spaces. In other words, if $D^{(1)}(g)\mathbf{e}_j^{(1)} = \mathbf{e}_i^{(1)} D_{ij}^{(1)}(g)$ and $D^{(2)}(g)\mathbf{e}_j^{(2)} = \mathbf{e}_i^{(2)} D_{ij}^{(2)}(g)$ we define

$$[D^{(1)} \otimes D^{(2)}](g)(\mathbf{e}_i^{(1)} \otimes \mathbf{e}_j^{(2)}) = (\mathbf{e}_k^{(1)} \otimes \mathbf{e}_l^{(2)}) D_{ki}^{(1)}(g) D_{lj}^{(2)}(g). \tag{5.7}$$

We think of $D_{ki}^{(1)}(g) D_{lj}^{(2)}(g)$ being the entries in the direct-product matrix matrix

$$[D^{(1)}(g) \otimes D^{(2)}(g)]_{kl,ij},$$

whose rows and columns are indexed by *pairs* of numbers. The dimension of the product representation is therefore the product of the dimensions of its factors.

Exercise 5.9: Show that if $D(g)$ is a representation, then so is

$$D'(g) = [D(g^{-1})]^T,$$

where the superscript T denotes the transposed matrix.

Exercise 5.10: Show that a map that assigns every element of a group G to the 1-by-1 identity matrix is a representation. It is, not unreasonably, called the *trivial* representation.

Exercise 5.11: A representation $D : G \to \mathrm{GL}(n, \mathbb{C})$ that assigns an element $g \in G$ to the n-by-n identity matrix I_n if and only if $g = e$ is said to be *faithful*. Let D be a non-trivial, but non-faithful, representation of G by n-by-n matrices. Let $H \subset G$ consist of those elements h such that $D(h) = I_n$. Show that H is a normal subgroup of G, and that D projects to a faithful representation of the quotient group G/H.

Exercise 5.12: Let A and B be linear maps from $U \to U$ and C and D be linear maps from $V \to V$. Then the direct products $A \otimes C$ and $B \otimes D$ are linear maps from $U \otimes V \to U \otimes V$. Show that

$$(A \otimes C)(B \otimes D) = (AB) \otimes (CD).$$

Show also that

$$(A \oplus C)(B \oplus D) = (AB) \oplus (CD).$$

Exercise 5.13: Let A and B be m-by-m and n-by-n matrices respectively, and let I_n denote the n-by-n unit matrix. Show that:

i) $\mathrm{tr}(A \oplus B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.
ii) $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\,\mathrm{tr}(B)$.
iii) $\exp(A \oplus B) = \exp(A) \oplus \exp(B)$.
iv) $\exp(A \otimes I_n + I_m \otimes B) = \exp(A) \otimes \exp(B)$.
v) $\det(A \oplus B) = \det(A)\det(B)$.
vi) $\det(A \otimes B) = (\det(A))^n (\det(B))^m$.

5.2.1 Reducibility and Irreducibility

The "atoms" of representation theory are those representations that cannot, by a clever choice of basis, be decomposed into, or *reduced* to, a direct sum of smaller representations. Such a representation is said to be *irreducible*. It is not easy to tell by just looking at a representation whether is is reducible or not. We need to develop some tools. We begin with a more powerful definition of irreducibilty.

We first introduce the notion of an *invariant subspace*. Suppose we have a set $\{A_\alpha\}$ of linear maps acting on a vector space V. A subspace $U \subseteq V$ is an invariant subspace for the set if $x \in U \Rightarrow A_\alpha x \in U$ for all A_α. The set $\{A_\alpha\}$ is *irreducible* if the only invariant subspaces are V itself and $\{0\}$. Conversely, if there is a non-trivial invariant subspace, then the set[4] of operators is *reducible*.

If the A_α's posses a non-trivial invariant subspace U, and we decompose $V = U \oplus U'$, where U' is a complementary subspace, then, in a basis adapted to this decomposition, the matrices A_α take the block-partitioned form of figure 5.1.

[4]Irreducibility is a property of the set as a whole. Any individual matrix always has a non-trivial invariant subspace because it possesses at least one eigenvector.

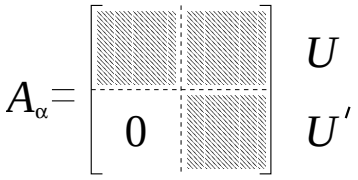

Figure 5.1: *Block partitioned reducible matrices.*

If we can find a[5] complementary subspace U' which is also invariant, then we have the block partitioned form of figure 5.2.

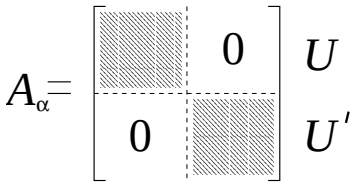

Figure 5.2: *Completely reducible matrices.*

We say that such matrices are *completely reducible*. When our linear operators are unitary with respect to some inner product, we can take the complementary subspace to be the *orthogonal complement*. This, by unitarity, is automatically be invariant. Thus, unitarity and reducibility implies complete reducibility.

Schur's Lemma

The most useful results concerning irreducibility come from:
Schur's Lemma: Suppose we have two sets of linear operators $A_\alpha : U \to U$, and $B_\alpha : V \to V$, that act irreducibly on their spaces, and an *intertwining operator* $\Lambda : U \to V$ such that

$$\Lambda A_\alpha = B_\alpha \Lambda, \tag{5.8}$$

for all α, then *either*
a) $\Lambda = 0$,
or

[5]Remember that complementary subspaces are not unique.

b) Λ is 1-1 and onto (and hence invertible), in which case U and V have the same dimension and $A_\alpha = \Lambda^{-1} B_\alpha \Lambda$.

The proof is straightforward: The relation (5.8) shows that $\mathrm{Ker}\,(\Lambda) \subseteq U$ and $\mathrm{Im}(\Lambda) \subseteq V$ are invariant subspaces for the sets $\{A_\alpha\}$ and $\{B_\alpha\}$ respectively. Consequently, either $\Lambda = 0$, or $\mathrm{Ker}\,(\Lambda) = \{0\}$ and $\mathrm{Im}(\Lambda) = V$. In the latter case Λ is 1-1 and onto, and hence invertible.

Corollary: If $\{A_\alpha\}$ acts irreducibly on an n-dimensional vector space, and there is an operator Λ such that

$$\Lambda A_\alpha = A_\alpha \Lambda, \tag{5.9}$$

then either $\Lambda = 0$ or $\Lambda = \lambda I$. To see this observe that (5.9) remains true if Λ is replaced by $(\Lambda - xI)$. Now $\det(\Lambda - xI)$ is a polynomial in x of degree n, and, by the fundamental theorem of algebra, has at least one root, $x = \lambda$. Since its determinant is zero, $(\Lambda - \lambda I)$ is not invertible, and so must vanish by Schur's lemma.

5.2.2 Characters and Orthogonality

Unitary Representations of Finite Groups

Let G be a finite group and let $g \mapsto D(g)$ be a representation of G by matrices acting on a vector space V. Let $(\mathbf{x}, \mathbf{y})$ denote a positive-definite, conjugate-symmetric, sesquilinear inner product of two vectors in V. From $(\ ,\ )$ we construct a new inner product $\langle\ ,\ \rangle$ by averaging over the group

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{g \in G} (D(g)\mathbf{x}, D(g)\mathbf{y}). \tag{5.10}$$

It is easy to see that this new inner product remains positive definite, and in addition has the property that

$$\langle D(g)\mathbf{x}, D(g)\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{5.11}$$

This means that the maps $D(g) : V \to V$ are unitary with respect to the new product. If we change basis to one that is orthonormal with respect to this new product then the $D(g)$ become unitary matrices, with $D(g^{-1}) = D^{-1}(g) = D^\dagger(g)$, where $D^\dagger_{ij}(g) = D^*_{ji}(g)$ denotes the conjugate-transposed matrix.

We conclude that representations of finite groups can always be taken to be unitary. This leads to the important consequence that for such representations reducibility implies complete reducibility. **Warning**: In this construction it is essential that the sum over the $g \in G$ converge. This is guaranteed for a finite group, but may not work for infinite groups. In particular, non-compact Lie groups, such as the Lorentz group, have no finite dimensional unitary representations.

Orthogonality of the Matrix Elements

Now let $D^J(g) : V_J \to V_J$ be the matrices of an irreducible representation or *irrep*. Here J is a label which distinguishes inequivalent irreps from one another. We will use the symbol $\dim J$ to denote the dimension of the representation vector space V_J.

Let D^K be an irrep that is either identical to D^J or inequivalent, and let M_{ij} be a matrix possessing the appropriate number of rows and columns for product $D^J M D^K$ to be defined, but otherwise arbitrary. The sum

$$\Lambda = \sum_{g\in G} D^J(g^{-1}) M D^K(g) \tag{5.12}$$

obeys $D^J(g)\Lambda = \Lambda D^K(g)$ for any g. Consequently, Schur's lemma tells us that

$$\Lambda_{il} = \sum_{g\in G} D^J_{ij}(g^{-1}) M_{jk} D^K_{kl}(g) = \lambda(M)\delta_{il}\delta^{JK}. \tag{5.13}$$

We have written $\lambda(M)$ to stress that the number λ depends on the chosen matrix M. Now take M to be zero everywhere except for one entry of unity in row j column k. Then we have

$$\sum_{g\in G} D^J_{ij}(g^{-1}) D^K_{kl}(g) = \lambda_{jk}\delta_{il}, \delta^{JK} \tag{5.14}$$

where we have relabelled λ to indicate its dependence on the location (j,k) of the non-zero entry in M. We can find the constants λ_{jk} by assuming that $K = J$, setting $i = l$, and summing over i. We find

$$|G|\delta_{jk} = \lambda_{jk} \dim J. \tag{5.15}$$

Putting these results together we find that

$$\frac{1}{|G|}\sum_{g\in G} D^J_{ij}(g^{-1}) D^K_{kl}(g) = (\dim J)^{-1}\delta_{jk}\delta_{il}\delta^{JK}. \tag{5.16}$$

When our matrices $D(g)$ are unitary, we can write this as

$$\frac{1}{|G|}\sum_{g\in G}\left(D^J_{ij}(g)\right)^* D^K_{kl}(g) = (\dim J)^{-1}\delta_{ik}\delta_{jl}\delta^{JK}. \tag{5.17}$$

If we consider complex-valued functions $G \to \mathbb{C}$ as forming a vector space, then the D^J_{ij} are elements of this space and are mutually orthogonal with respect to its natural inner product.

There can be no more orthogonal functions on G than the dimension of the function space itself, which is $|G|$. We therefore have a constraint

$$\sum_J (\dim J)^2 \le |G| \tag{5.18}$$

that places a limit on how many inequivalent representations can exist. In fact, as you will show later, the equality holds: the sum of the squares of the dimensions of the inequivalent irreducible representations is equal to the order of G, and consequently the matrix elements form a complete orthonormal set of functions on G.

Class functions and characters

Because

$$\operatorname{tr}(C^{-1}DC) = \operatorname{tr} D, \tag{5.19}$$

the trace of a representation matrix is the same for equivalent representations. Further, because

$$\operatorname{tr} D(g_1^{-1}gg_1) = \operatorname{tr}\left(D^{-1}(g_1)D(g)D(g_1)\right) = \operatorname{tr} D(g), \tag{5.20}$$

the trace is the same for all group elements in a conjugacy class. The *character*,

$$\chi(g) \stackrel{\text{def}}{=} \operatorname{tr} D(g), \tag{5.21}$$

is therefore said to be a *class function*.

By taking the trace of the matrix-element orthogonality relation we see that the characters $\chi^J = \operatorname{tr} D^J$ of the irreducible representations obey

$$\frac{1}{|G|}\sum_{g\in G}\left(\chi^J(g)\right)^*\chi^K(g) = \frac{1}{|G|}\sum_i d_i\left(\chi^J_i\right)^*\chi^K_i = \delta^{JK}, \tag{5.22}$$

where d_i is the number of elements in the i-th conjugacy class.

The completeness of the matrix elements as functions on G implies that the characters form a complete orthogonal set of functions on the space of conjugacy classes equipped with inner product

$$\langle \chi^1, \chi^2 \rangle \stackrel{\text{def}}{=} \frac{1}{|G|} \sum_i d_i \left(\chi_i^1\right)^* \chi_i^2. \tag{5.23}$$

Conseqently there are exactly as many inequivalent irreducible representations as there are conjugacy classes in the group.

Given a reducible representation, $D(g)$, we can find out exactly which irreps J it contains, and how many times, n_J, they occur. We do this forming the *compound character*

$$\chi(g) = \operatorname{tr} D(g) \tag{5.24}$$

and observing that if we can find a basis in which

$$D(g) = \underbrace{(D^1(g) \oplus D^1(g) \oplus \cdots)}_{n_1 \text{ terms}} \oplus \underbrace{(D^2(g) \oplus D^2(g) \oplus \cdots)}_{n_2 \text{ terms}} \oplus \cdots, \tag{5.25}$$

then

$$\chi(g) = n_1 \chi^1(g) + n_2 \chi^2(g) + \cdots \tag{5.26}$$

From this we find

$$n_J = \langle \chi, \chi^J \rangle = \frac{1}{|G|} \sum_i d_i \, (\chi_i)^* \chi_i^J. \tag{5.27}$$

There are extensive tables of group characters. Table 5.2 shows, for example, the characters of the group S_4 of permutations on 4 objects.

	Typical element and class size				
S_4	(1)	(12)	(123)	(1234)	(12)(34)
Irrep	1	6	8	6	3
A_1	1	1	1	1	1
A_2	1	-1	1	-1	1
E	2	0	-1	0	2
T_1	3	1	0	-1	-1
T_2	3	-1	0	1	-1

Table 5.2: *Character table of* S_4

Since $\chi^J(e) = \dim J$ we see that the irreps A_1 and A_2 are one dimensional, that E is two dimensional, and that $T_{1,2}$ are both three dimensional. Also we confirm that the sum of the squares of the dimensions

$$1 + 1 + 2^2 + 3^2 + 3^2 = 24 = 4!$$

is equal to the order of the group.

As a further illustration of how to read table 5.2, let us verify the orthonormality of the characters of the representations T_1 and T_2. We have

$$\langle \chi^{T_1}, \chi^{T_2} \rangle = \frac{1}{|G|} \sum_i d_i \left(\chi_i^{T_1}\right)^* \chi_i^{T_2} = \frac{1}{24}[1{\cdot}3{\cdot}3 - 6{\cdot}1{\cdot}1 + 8{\cdot}0{\cdot}0 - 6{\cdot}1{\cdot}1 + 3{\cdot}1{\cdot}1] = 0,$$

while

$$\langle \chi^{T_1}, \chi^{T_1} \rangle = \frac{1}{|G|} \sum_i d_i \left(\chi_i^{T_1}\right)^* \chi_i^{T_1} = \frac{1}{24}[1{\cdot}3{\cdot}3 + 6{\cdot}1{\cdot}1 + 8{\cdot}0{\cdot}0 + 6{\cdot}1{\cdot}1 + 3{\cdot}1{\cdot}1] = 1.$$

The sum giving $\langle \chi^{T_2}, \chi^{T_2} \rangle = 1$ is identical to this.

Exercise 5.14: Let D^1 and D^2 be representations with characters $\chi^1(g)$ and $\chi^2(g)$ respectively. Show that the character of the direct product representation $D^1 \otimes D^2$ is given by

$$\chi^{1\otimes 2}(g) = \chi^1(g)\chi^2(g).$$

5.2.3 The Group Algebra

Given a finite group G, we construct a vector space $\mathbb{C}(G)$ whose basis vectors are in one-to-one correspondence with the elements of the group. We denote the vector corresponding to the group element g by the boldface symbol $\mathbf{g}$. A general element of $\mathbb{C}(G)$ is therefore a formal sum

$$\mathbf{x} = x_1\mathbf{g}_1 + x_2\mathbf{g}_2 + \cdots + x_{|G|}\mathbf{g}_{|G|}. \tag{5.28}$$

We take products of these sums by using the group multiplication rule. If $g_1 g_2 = g_3$ we set $\mathbf{g}_1\mathbf{g}_2 = \mathbf{g}_3$, and require the product to be distributive with respect to vector-space addition. Thus

$$\mathbf{g}\mathbf{x} = x_1\mathbf{g}\mathbf{g}_1 + x_2\mathbf{g}\mathbf{g}_2 + \cdots + x_{|G|}\mathbf{g}\mathbf{g}_{|G|}. \tag{5.29}$$

The resulting mathematical structure is called the *group algebra*. It was introduced by Frobenius.

The group algebra, considered as a vector space, is automatically a representation. We define the natural action of G on $\mathbb{C}(G)$ by setting

$$D(g)\mathbf{g}_i = \mathbf{g}\,\mathbf{g}_i = \mathbf{g}_j D_{ji}(g). \tag{5.30}$$

The matrices $D_{ji}(g)$ make up the *regular representation.* Because the list $\mathbf{g}\,\mathbf{g}_1, \mathbf{g}\,\mathbf{g}_2, \ldots$ is a permutation of the list $\mathbf{g}_1, \mathbf{g}_2, \ldots$, their entries consist of 1's and 0's, with exactly one non-zero entry in each row and each column.

Exercise 5.15: Show that the character of the regular representation has $\chi(e) = |G|$, and $\chi(g) = 0$, for $g \neq e$.

Exercise 5.16: Use the previous exercise to show that the number of times an n dimensional irrep occurs in the regular representation is n. Deduce that $|G| = \sum_J (\dim J)^2$, and from this construct the completeness proof for the representations and characters.

Projection Operators

A representation D^J of the group automatically provides a representation of the group algebra. We simply set

$$D^J(x_1\mathbf{g}_1 + x_2\mathbf{g}_2 + \cdots) \stackrel{\text{def}}{=} x_1 D^J(g_1) + x_2 D^J(g_2) + \cdots. \tag{5.31}$$

Certain linear combinations of group elements turn out to be very useful because the corresponding matrices can be used to project out vectors with desirable symmetry properties.

Consider the elements

$$\mathbf{e}^J_{\alpha\beta} = \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* \mathbf{g} \tag{5.32}$$

of the group algebra. These have the property that

$$\begin{aligned} \mathbf{g}_1 \mathbf{e}^J_{\alpha\beta} &= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* (\mathbf{g}_1 \mathbf{g}) \\ &= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g_1^{-1} g)\right]^* \mathbf{g} \end{aligned}$$

$$
\begin{aligned}
&= \left[D^J_{\alpha\gamma}(g_1^{-1})\right]^* \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\gamma\beta}(g)\right]^* \mathbf{g} \\
&= \mathbf{e}^J_{\gamma\beta} D^J_{\gamma\alpha}(g_1). \qquad (5.33)
\end{aligned}
$$

In going from the first to the second line we have changed summation variables from $g \to g_1^{-1}g$, and going from the second to the third line we have used the representation property to write $D^J(g_1^{-1}g) = D^J(g_1^{-1})D^J(g)$.

From $\mathbf{g}_1 \mathbf{e}^J_{\alpha\beta} = \mathbf{e}^J_{\gamma\beta} D^J_{\gamma\alpha}(g_1)$ and the matrix-element orthogonality, it follows that

$$
\begin{aligned}
\mathbf{e}^J_{\alpha\beta}\, \mathbf{e}^K_{\gamma\delta} &= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* \mathbf{g}\, \mathbf{e}^K_{\gamma\delta} \\
&= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* D^K_{\epsilon\gamma}(g) \mathbf{e}^K_{\epsilon\delta} \\
&= \delta^{JK}\delta_{\alpha\epsilon}\delta_{\beta\gamma}\, \mathbf{e}^K_{\epsilon\delta} \\
&= \delta^{JK}\delta_{\beta\gamma}\, \mathbf{e}^J_{\alpha\delta}. \qquad (5.34)
\end{aligned}
$$

For each J, this multiplication rule of the $\mathbf{e}^J_{\alpha\beta}$ is identical to that of matrices having zero entries everywhere except for the (α, β)-th, which is a "1." There are $(\dim J)^2$ of these $\mathbf{e}^J_{\alpha\beta}$ for each n-dimensional representation J, and they are linearly independent. Because $\sum_J (\dim J)^2 = |G|$, they form a basis for the algebra. In particular every element of G can be reconstructed as

$$
\mathbf{g} = \sum_J D^J_{ij}(g) \mathbf{e}^J_{ij}. \qquad (5.35)
$$

We can also define the useful objects

$$
\mathbf{P}^J = \sum_i \mathbf{e}^J_{ii} = \frac{\dim J}{|G|} \sum_{g\in G} \left[\chi^J(g)\right]^* \mathbf{g}. \qquad (5.36)
$$

They have the property

$$
\mathbf{P}^J \mathbf{P}^K = \delta^{JK} \mathbf{P}^K, \quad \sum_J \mathbf{P}^J = \mathbf{I}, \qquad (5.37)
$$

where $\mathbf{I}$ is the identity element of $\mathbb{C}(G)$. The $\mathbf{P}^J$ are therefore projection operators composing a resolution of the identity. Their utility resides in the fact that when $D(g)$ is a reducible representation acting on a linear space

$$
V = \bigoplus_J V_J, \qquad (5.38)
$$

then setting $\mathbf{g} \to D(g)$ in the formula for $\mathbf{P}^J$ results in a projection matrix from V onto the irreducible component V_J. To see how this comes about, let $\mathbf{v} \in V$ and, for any fixed p, set

$$\mathbf{v}_i = \mathbf{e}^J_{ip}\mathbf{v}, \tag{5.39}$$

where $\mathbf{e}^J_{ip}\mathbf{v}$ should be understood as shorthand for $D(\mathbf{e}^J_{ip})\mathbf{v}$. Then

$$D(g)\mathbf{v}_i = \mathbf{g}\mathbf{e}^J_{ip}\mathbf{v} = \mathbf{e}^J_{jp}\mathbf{v}D^J_{ji}(g) = \mathbf{v}_j D^J_{ji}(g). \tag{5.40}$$

We see the $\mathbf{v}_i$, if not all zero, are basis vectors for V_J. Since $\mathbf{P}^J$ is a sum of the $\mathbf{e}^J_{ij}$, the vector $\mathbf{P}^J\mathbf{v}$ is a sum of such vectors, and therefore lies in V_J. The advantage of using $\mathbf{P}^J$ over any individual $\mathbf{e}^J_{ip}$ is that $\mathbf{P}^J$ can be computed from character table, *i.e.* its construction does not require knowledge of the irreducible representation matrices.

The algebra of classes

If a conjugacy class C_i consists of the elements $\{g_1, g_2, \ldots g_{d_i}\}$, we can define $\mathbf{C}_i$ to be the corresponding element of the group algebra:

$$\mathbf{C}_i = \frac{1}{d_i}(\mathbf{g}_1 + \mathbf{g}_2 + \cdots \mathbf{g}_{d_i}). \tag{5.41}$$

(The factor of $1/d_i$ is a conventional normalization.) Because conjugation merely permutes the elements of a conjugacy class, we have $\mathbf{g}^{-1}\mathbf{C}_i\mathbf{g} = \mathbf{C}_i$ for all $\mathbf{g} \in \mathbb{C}(G)$. The $\mathbf{C}_i$ therefore commute with every element of $\mathbb{C}(G)$. Conversely any element of $\mathbb{C}(G)$ that commutes with everything in $\mathbb{C}(G)$ must be a linear combination $\mathbf{C} = c_1\mathbf{C}_1 + c_2\mathbf{C}_2 + \ldots$. The subspace of $\mathbb{C}(G)$ consisting of sums of the classes is therefore the *centre* $Z[\mathbb{C}(G)]$ of the group algebra. Because the product $\mathbf{C}_i\mathbf{C}_j$ commutes with everything, it lies in $Z[\mathbb{C}(G)]$ and so there are constants ${c_{ij}}^k$ such that

$$\mathbf{C}_i\mathbf{C}_j = \sum_k {c_{ij}}^k \mathbf{C}_k. \tag{5.42}$$

We can regard the $\mathbf{C}_i$ as being linear maps from $Z[\mathbb{C}(G)]$ to itself, whose associated matrices have entries $(\mathbf{C}_i)^k{}_j = {c_{ij}}^k$. These matrices commute, and can be simultaneously diagonalized. We will leave it as exercise for the reader to demonstrate that

$$\mathbf{C}_i\mathbf{P}^J = \left(\frac{\chi^J_i}{\chi^J_0}\right)\mathbf{P}^J. \tag{5.43}$$

Here $\chi_0^J \equiv \chi_{\{e\}}^J = \dim J$. The common eigenvectors of the $\mathbf{C}_i$ are therefore the projection operators $\mathbf{P}^J$, and the eigenvalues $\lambda_i^J = \chi_i^J/\chi_0^J$ are, up to normalization, the characters. Equation (5.43) provides a convenient method for computing the characters from knowledge only of the coefficients ${c_{ij}}^k$ appearing in the class multiplication table. Once we have found the eigenvalues λ_i^J, we recover the χ_i^J by noting that χ_0^J is real and positive, and that $\sum_i d_i |\chi_i^J|^2 = |G|$.

Exercise 5.17: Use Schur's lemma to show that for an irrep $D^J(g)$ we have

$$\frac{1}{d_i}\sum_{g\in C_i} D^J_{jk}(g) = \frac{1}{\dim J}\delta_{jk}\chi_i^J,$$

and hence establish (5.43).

5.3 Physics Applications

5.3.1 Quantum Mechanics

When a group $G = \{g_i\}$ acts on a mechanical system, then G will act as set of linear operators $D(g)$ on the Hilbert space $\mathcal{H}$ of the corresponding quantum system. Thus $\mathcal{H}$ will be a representation[6] space for G. If the group is a symmetry of the system then the $D(g)$ will commute with the hamiltonian $\hat{H}$. If this is so, and if we can decompose

$$\mathcal{H} = \bigoplus_{\text{irreps } J} \mathcal{H}_J \tag{5.44}$$

into $\hat{H}$-invariant irreps of G then Schur's lemma tells us that in each $\mathcal{H}_J$ the hamiltonian $\hat{H}$ will act as a multiple of the identity operator. In other words every state in $\mathcal{H}_J$ will be an eigenstate of $\hat{H}$ with a common energy E_J.

This fact can greatly simplify the task of finding the energy levels. If an irrep J occurs only once in the decomposition of $\mathcal{H}$ then we can find the eigenstates directly by applying the projection operator $\mathbf{P}^J$ to vectors in $\mathcal{H}$.

[6]The rules of quantum mechanics only require that $D(g_1)D(g_2) = e^{i\phi(g_1,g_2)}D(g_1g_2)$. A set of matrices that obeys the group multiplication rule "up to a phase" is called a *projective* (or *ray*) representation. In many cases, however, we can choose the $D(g)$ so that ϕ is not needed. This is the case in all the examples we discuss.

If the irrep occurs n_J times in the decomposition, then $\mathbf{P}^J$ will project to the reducible subspace

$$\underbrace{\mathcal{H}_J \oplus \mathcal{H}_J \oplus \cdots \mathcal{H}_J}_{n_J \text{ copies}} = \mathcal{M} \otimes \mathcal{H}_J.$$

Here $\mathcal{M}$ is an n_J dimensional *multiplicity space*. The hamiltonian $\hat{H}$ will act in $\mathcal{M}$ as an n_J-by-n_J matrix. In other words, if the vectors

$$|n, i\rangle \equiv |n\rangle \otimes |i\rangle \in \mathcal{M} \otimes \mathcal{H}_J \tag{5.45}$$

form a basior $\mathcal{M} \otimes \mathcal{H}_J$, with n labelling which copy of $\mathcal{H}_J$ the vector $|n, i\rangle$ lies in, then

$$\begin{aligned} \hat{H}|n, i\rangle &= |m, i\rangle H^J_{mn}, \\ D(g)|n, i\rangle &= |n, j\rangle D^J_{ji}(g). \end{aligned} \tag{5.46}$$

Diagonalizing H^J_{nm} provides us with n_j $\hat{H}$-invariant copies of $\mathcal{H}_J$ and gives us the energy eigenstates.

Consider, for example, the molecule C_{60} (buckminsterfullerine) consisting of 60 carbon atoms in the form of a soccer ball. The chemically active electrons can be treated in a tight-binding approximation in which the Hilbert space has dimension 60 — one π-orbital basis state for each each carbon atom. The geometric symmetry group of the molecule is $Y_h = Y \times \mathbb{Z}_2$, where Y is the rotational symmetry group of the icosohedron (a subgroup of SO(3)) and $\mathbb{Z}_2$ is the parity inversion $\sigma : \mathbf{r} \mapsto -\mathbf{r}$. The characters of Y are displayed in table 5.3.

	Typical element and class size				
Y	e	C_5	C_5^2	C_2	C_3
Irrep	1	12	12	15	20
A	1	1	1	1	1
T_1	3	τ^{-1}	$-\tau$	-1	0
T_2	3	$-\tau$	τ^{-1}	-1	0
G	4	-1	-1	0	1
H	5	0	0	1	-1

Table 5.3: *Character table for the group Y.*

In this table $\tau = \frac{1}{2}(\sqrt{5}-1)$ denotes the golden mean. The class C_5 is the set of $2\pi/5$ rotations about an axis through the centres of a pair of antipodal pentagonal faces, the class C_3 is the set of of $2\pi/3$ rotations about an axis through the centres of a pair of antipodal hexagonal faces, and C_2 is the set of π rotatations through the midpoints of a pair of antipodal edges, each lying between two adjacent hexagonal faces.

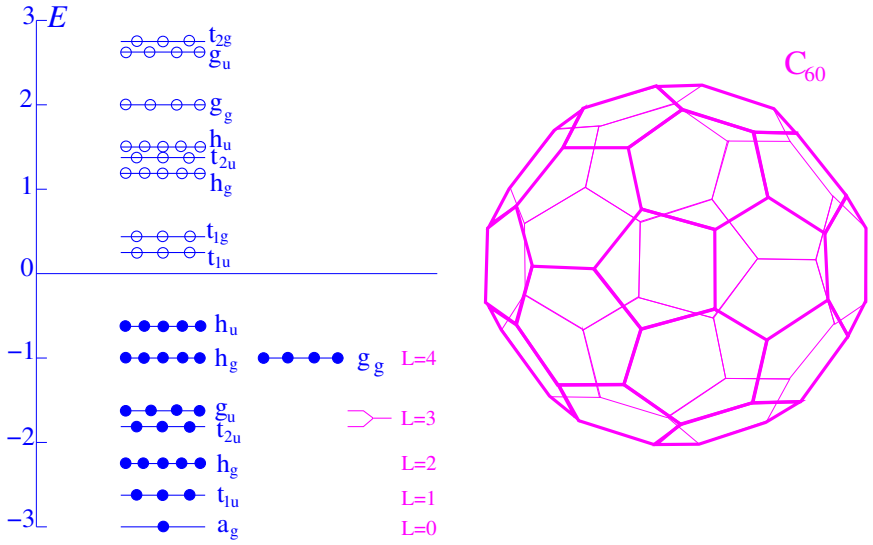

Figure 5.3: *A sketch of the tight-binding electronic energy levels of* C_{60}.

The geometric symmetry group acts on the 60-dimensional Hilbert space by permuting the basis states concurrently with their associated atoms. Figure 5.3 shows how the 60 states are disposed into energy levels.[7] Each level is labelled by a lower case letter specifying the irrep of Y, and by a subscript g or u standing for *gerade* (German for *even*) or *ungerade* (German for *odd*) that indicates whether the wavefunction is even or odd under the inversion $\sigma : \mathbf{r} \mapsto -\mathbf{r}$.

The buckyball is roughly spherical, and the lowest 25 states can be thought as being derived from the angular-momentum eigenstates with $L = 0, 1, 2, 3, 4$, that classify the energy levels for an electron moving on a perfect sphere. In the many-electron ground-state, the 30 single-particle states with energy below $E < 0$ are each occupied by pairs of spin up/down electrons. The 30 states with $E > 0$ are empty.

[7] After R. C. Haddon, L. E. Brus, K. Raghavachari, *Chem. Phys. Lett.* **125** (1986) 459.

To explain, for example, why three copies of T_1 appear, and why two of these are T_{1u} and one T_{1g}, we must investigate the manner in which the 60-dimensional Hilbert space decomposes into irreducible representations of 120-element group Y_h. Problem 5.23 leads us through this computation, and shows that no irrep of Y_h occurs more that three times. In finding the energy levels, we therefore never have to diagonalize a bigger than 3-by-3 matrix.

The equality of the energies of the h_g and g_g levels at $E = -1$ is an *accidental degeneracy*. It is not required by the symmetry, and will presumably disappear in a more sophisticated calculation. The appearance of many "accidental" degeneracies in an energy spectrum hints that there may be a *hidden symmetry* that arises from something beyond geometry. For example, in the Schrödinger spectrum of the hydrogen atom all states with the same principal quantum number n have the same energy although they correspond to different irreps $L = 1, \ldots, n-1$ of O(3). This degeneracy occurs because the classical Kepler-orbit problem has symmetry group O(4), rather than the naïvely expected O(3) rotational symmetry.

5.3.2 Vibrational spectrum of H_2O

The small vibrations of a mechanical system with n degrees of freedom are governed by a Lagrangian of the form

$$L = \frac{1}{2}\dot{\mathbf{x}}^T M \dot{\mathbf{x}} - \frac{1}{2}\mathbf{x}^T V \mathbf{x} \tag{5.47}$$

where M and V are symmetric n-by-n matrices, and with M being positive definite. This Lagrangian leads to the equations of motion

$$M\ddot{\mathbf{x}} = V\mathbf{x} \tag{5.48}$$

We look for normal mode solutions $\mathbf{x}(t) \propto e^{i\omega_i t}\mathbf{x}_i$, where the vectors $\mathbf{x}_i$ obey

$$-\omega_i^2 M\mathbf{x}_i = V\mathbf{x}_i. \tag{5.49}$$

The normal-mode frequencies are solutions of the secular equation

$$\det\left(V - \omega^2 M\right) = 0, \tag{5.50}$$

and modes with distinct frequencies are orthogonal with respect to the inner product defined by M,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T M \mathbf{y}. \tag{5.51}$$

We are interested in solving this problem for vibrations about the equilibrium configuration of a molecule. Suppose this equilibrium configuration has a symmetry group G. This gives rise to an n-dimensional representation on the space of $\mathbf{x}$'s in which

$$g : \mathbf{x} \mapsto D(g)\mathbf{x}, \tag{5.52}$$

leaves both the intertia matrix M and the potential matrix V unchanged.

$$[D(g)]^T M D(g) = M, \qquad [D(g)]^T V D(g) = V. \tag{5.53}$$

Consequently, if we have an eigenvector $\mathbf{x}_i$ with frequency ω_i,

$$-\omega_i^2 M\mathbf{x}_i = V\mathbf{x}_i \tag{5.54}$$

we see that $D(g)\mathbf{x}_i$ also satisfies this equation. The frequency eigenspaces are therefore left invariant by the action of $D(g)$, and barring accidental degeneracy, there will be a one-to-one correspondence between the frequency eigenspaces and the irreducible representations occurring in $D(g)$.

Consider, for example, the vibrational modes of the water molecule H_2O. This familiar molecule has symmetry group C_{2v} which is generated by two elements: a rotation a through π about an axis through the oxygen atom, and a reflection b in the plane through the oxygen atom and bisecting the angle between the two hydrogens. The product ab is a reflection in the plane defined by the equilibrium position of the three atoms. The relations are $a^2 = b^2 = (ab)^2 = e$, and the characters are displayed in table 5.4.

	class and size			
C_{2v}	e	a	b	ab
Irrep	1	1	1	1
A_1	1	1	1	1
A_2	1	1	-1	-1
B_1	1	-1	1	-1
B_2	1	-1	-1	1

Table 5.4: *Character table of* C_{2v}.

The group C_{2v} is Abelian, so all the representations are one dimensional.

To find out what representations occur when C_{2v} acts, we need to find the character of its action $D(g)$ on the nine-dimensional vector

$$\mathbf{x} = (x_O, y_O, z_O, x_{H_1}, y_{H_1}, z_{H_1}, x_{H_2}, y_{H_2}, z_{H_2}). \tag{5.55}$$

Here the coordinates $x_{H_2}, y_{H_2}, z_{H_2}$ *etc.* denote the *displacements* of the labelled atom from its equilibrium position.

We take the molecule as lying in the xy plane, with the z pointing towards us.

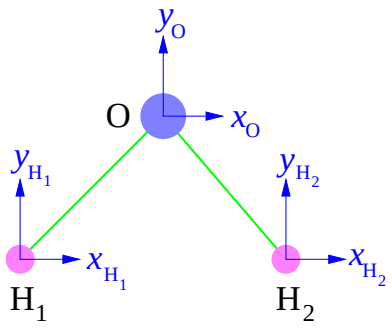

Figure 5.4: *Water Molecule.*

The effect of the symmetry operations on the atomic displacements is

$$\begin{aligned} D(a)\mathbf{x} &= (-x_O, +y_O, -z_O, -x_{H_2}, +y_{H_2}, -z_{H_2}, -x_{H_1}, +y_{H_1}, -z_{H_1}) \\ D(b)\mathbf{x} &= (-x_O, +y_O, +z_O, -x_{H_2}, +y_{H_2}, +z_{H_2}, -x_{H_1}, +y_{H_1}, +z_{H_1}) \\ D(ab)\mathbf{x} &= (+x_O, +y_O, -z_O, +x_{H_1}, +y_{H_1}, -z_{H_1}, +x_{H_2}, +y_{H_2}, -z_{H_2}). \end{aligned}$$

Notice how the transformations $D(a)$, $D(b)$ have interchanged the displacement co-ordinates of the two hydrogen atoms. In calculating the character of a transformation we need look only at the effect on atoms that are left fixed — those that are moved have matrix elements only in non-diagonal positions. Thus, when computing the compound characters for a b, we can focus on the oxygen atom. For ab we need to look at all three atoms. We find

$$\begin{aligned} \chi^D(e) &= 9, \\ \chi^D(a) &= -1 + 1 - 1 = -1, \\ \chi^D(b) &= -1 + 1 + 1 = 1, \\ \chi^D(ab) &= 1 + 1 - 1 + 1 + 1 - 1 + 1 + 1 - 1 = 3. \end{aligned}$$

By using the orthogonality relations, we find the decomposition

$$\begin{pmatrix} 9 \\ -1 \\ 1 \\ 3 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + 2\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} + 3\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \tag{5.56}$$

or

$$\chi^D = 3\chi^{A_1} + \chi^{A_2} + 2\chi^{B_1} + 3\chi^{B_2}. \tag{5.57}$$

Thus, the nine-dimensional representation decomposes as

$$D = 3A_1 \oplus A_2 \oplus 2B_1 \oplus 3B_2. \tag{5.58}$$

How do we exploit this? First we cut out the junk. Out of the nine modes, six correspond to easily identified zero-frequency motions – three of translation and three rotations. A translation in the x direction would have $x_O = x_{H_1} = x_{H_2} = \xi$, all other entries being zero. This displacement vector changes sign under both a and b, but is left fixed by ab. This behaviour is characteristic of the representation B_2. Similarly we can identify A_1 as translation in y, and B_1 as translation in z. A rotation about the y axis makes $z_{H_1} = -z_{H_2} = \phi$. This is left fixed by a, but changes sign under b and ab, so the y rotation mode is A_2. Similarly, rotations about the x and z axes correspond to B_1 and B_2 respectively. All that is left for genuine vibrational modes is $2A_1 \oplus B_2$.

We now apply the projection operator

$$P^{A_1} = \frac{1}{4}[(\chi^{A_1}(e))^* D(e) + (\chi^{A_1}(a))^* D(b) + (\chi^{A_1}(b))^* D(b) + (\chi^{A_1}(ab))^* D(ab)] \tag{5.59}$$

to $\mathbf{v}_{H_1,x}$, a small displacement of H_1 in the x direction. We find

$$\begin{aligned} P^{A_1}\mathbf{v}_{H_1,x} &= \frac{1}{4}(\mathbf{v}_{H_1,x} - \mathbf{v}_{H_2,x} - \mathbf{v}_{H_2,x} + \mathbf{v}_{H_1,x}) \\ &= \frac{1}{2}(\mathbf{v}_{H_1,x} - \mathbf{v}_{H_2,x}). \end{aligned} \tag{5.60}$$

This mode is an eigenvector for the vibration problem.

If we apply P^{A_1} to $\mathbf{v}_{H_1,y}$ and $\mathbf{v}_{O,y}$ we find

$$\begin{aligned} P^{A_1}\mathbf{v}_{H_1,y} &= \frac{1}{2}(\mathbf{v}_{H_1,y} + \mathbf{v}_{H_2,y}), \\ P^{A_1}\mathbf{v}_{O,y} &= \mathbf{v}_{O,y}, \end{aligned} \tag{5.61}$$

but we are not quite done. These modes are contaminated by the y translation direction zero mode, which is also in an A_1 representation. After we make our modes orthogonal to this, there is only one left, and this has $y_{H_1} = y_{H_2} = -y_O m_O/(2m_H) = a_1$, all other components vanishing.

We can similarly find vectors corresponding to B_2 as

$$\begin{aligned} P^{B_2}\mathbf{v}_{H_1,x} &= \frac{1}{2}(\mathbf{v}_{H_1,x} + \mathbf{v}_{H_2,x}) \\ P^{B_2}\mathbf{v}_{H_1,y} &= \frac{1}{2}(\mathbf{v}_{H_1,y} - \mathbf{v}_{H_2,y}) \\ P^{B_2}\mathbf{v}_{O,x} &= \mathbf{v}_{O,x} \end{aligned}$$

and these need to be cleared of both translations in the x direction and rotations about the z axis, both of which transform under B_2. Again there is only one mode left and it is

$$y_{H_1} = -y_{H_2} = \alpha x_{H_1} = \alpha x_{H_2} = \beta x_0 = a_2 \tag{5.62}$$

where α is chosen to ensure that there is no angular momentum about O, and β to make the total x linear momentum vanish. We have therefore found three true vibration eigenmodes, two transforming under A_1 and one under B_2 as advertised earlier. The eigenfrequencies, of course, depend on the details of the spring constants, but now that we have the eigenvectors we can just plug them in to find these.

5.3.3 Crystal Field Splittings

A quantum mechanical system has a symmetry G if the hamiltonian $\hat{H}$ obeys

$$D^{-1}(g)\hat{H}D(g) = \hat{H}, \tag{5.63}$$

for some group action $D(g) : \mathcal{H} \to \mathcal{H}$ on the Hilbert space. If follows that the eigenspaces, $\mathcal{H}_\lambda$, of states with a common eigenvalue, λ, are invariant subspaces for the representation $D(g)$.

We often need to understand how a degeneracy is lifted by perturbations that break G down to a smaller subgroup H. An n-dimensional irreducible representation of G is automatically a representation of any subgroup of G, but in general it is no longer be irreducible. Thus the n-fold degenerate level is split into multiplets, one for each of the irreducible representations

of H contained in the original representation. The manner in which an originally irreducible representation decomposes under restriction to a subgroup is known as the *branching rule* for the representation.

A physically important case is given by the breaking of the full SO(3) rotation symmetry of an isolated atomic hamiltonian by a crystal field Suppose the crystal has octohedral symmetry. The characters of the octohedral group are displayed in table 5.5.

	Class(size)				
O	e	$C_3(8)$	$C_4^2(3)$	$C_2(6)$	$C_4(6)$
A_1	1	1	1	1	1
A_2	1	1	1	-1	-1
E	2	-1	2	0	0
F_2	3	0	-1	1	-1
F_1	3	0	-1	-1	1

Table 5.5: *Character table of the octohedral group O.*

The classes are lableled by the rotation angles, C_2 being a twofold rotation axis ($\theta = \pi$), C_3 a threefold axis ($\theta = 2\pi/3$), *etc.*.

The chacter of the $J = l$ representation of $SO(3)$ is

$$\chi^l(\theta) = \frac{\sin(2l+1)\theta/2}{\sin\theta/2}, \tag{5.64}$$

and the first few χ^l's evaluated on the rotation angles of the classes of O are dsiplayed in table 5.6.

	Class(size)				
l	e	$C_3(8)$	$C_4^2(3)$	$C_2(6)$	$C_4(6)$
0	1	1	1	1	1
1	3	0	-1	-1	-1
2	5	-1	1	1	-1
3	7	1	-1	-1	-1
4	9	0	1	1	1

Table 5.6: *Characters evaluated on rotation classes*

The 9-fold degenerate $l = 4$ multiplet therefore decomposes as

$$\begin{pmatrix} 9 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \\ 2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \\ -1 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \tag{5.65}$$

or

$$\chi^4_{SO(3)} = \chi^{A_1} + \chi^E + \chi^{F_1} + \chi^{F_2}. \tag{5.66}$$

The octohedral crystal field splits the nine states into four multiplets with symmetries A_1, E, F_1, F_2 and degeneracies 1, 2, 3 and 3, respectively.

We have considered only the simplest case here, ignoring the complications introduced by reflection symmetries, and by 2-valued spinor representations of the rotation group.

5.4 Further Exercises and Problems

We begin with some technologically important applications of group theory to cryptography and number theory.

Exercise 5.18: The set $\mathbb{Z}_n$ forms a group under multiplication only when n is a prime number. Show, however, that the subset $\mathrm{U}(\mathbb{Z}_n) \subset \mathbb{Z}_n$ of elements of $\mathbb{Z}_n$ that are co-prime to n is a group. It is the *group of units* of the ring $\mathbb{Z}_n$.

Exercise 5.19: *Cyclic groups*. A group G is said to be *cyclic* if its elements consist of powers a^n of of an element a, called the *generator*. The group will be of finite order $|G| = m$ if $a^m = a^0 = e$ for some $m \in \mathbb{Z}^+$.

a) Show that a group of prime order is necessarily cyclic, and that any element other than the identity can serve as its generator. (Hint: Let a be any element other than e and consider the subgroup consisting of powers a^m.)
b) Show that any subgroup of a cyclic group is itself cyclic.

Exercise 5.20: *Cyclic groups and cryptography*. In a large cyclic group G it can be relatively easy to compute a^x, but to recover x given $h = a^x$ one might have to compute a^y and compare it with h for every $1 < y < |G|$. If $|G|$ has several hundred digits, such a brute force search could take longer than the age of the universe. Rather more efficient algorithms for this *discrete logarithm problem* exist, but the difficulty is still sufficient for it to be useful in cryptopgraphy.

a) *Diffie-Hellman key exchange*. This algorithm allows Alice and Bob to establish a secret key that can be used with a conventional cypher without Eve, who is listening to their conversation, being able to reconstruct it. Alice choses a random element $g \in G$ and an integer x between 1 and $|G|$ and computes g^x. She sends g and g^x to Bob, but keeps x to herself. Bob chooses an integer y and computes g^y and $g^{xy} = (g^x)^y$. He keeps y secret and sends g^y to Alice, who computes $g^{xy} = (g^y)^x$. Show that, although Eve knows g, g^y and g^x, she cannot obtain Alice and Bob's secret key g^{xy} without solving the discrete logarithm problem.

b) *ElGamal public key encryption*. This algorithm, based on Diffie-Hellman, was invented by the Egyptian cryptographer Taher Elgamal. It is a component of PGP and and other modern encryption packages. To use it, Alice first chooses a random integer x in the range 1 to $|G|$ and computes $h = a^x$. She publishes a description of G, together with the elements h and a, as her public key. She keeps the integer x secret. To send a message m to Alice, Bob chooses an integer y in the same range and computes $c_1 = a^y$, $c_2 = mh^y$. He transmits c_1 and c_2 to Alice, but keeps y secret. Alice can recover m from c_1, c_2 by computing $c_2(c_1^x)^{-1}$. Show that, although Eve knows Alice's public key and has overheard c_1 and c_2, she nonetheless cannot decrypt the message without solving the discrete logarithm problem.

Popular choices for G are subgroups of $(\mathbb{Z}_p)^\times$, for large prime p. $(\mathbb{Z}_p)^\times$ is itself cyclic (can you prove this?), but is unsuitable for technical reasons.

Exercise 5.21: *Modular arithmetic and number theory*. An integer a is said to be a *quadratic residue* mod p if there is an r such that $a = r^2 \pmod p$. Let p be an odd prime. Show that if $r_1^2 = r_2^2 \pmod p$ then $r_1 = \pm r_2 \pmod p$, and that $r \neq -r \pmod p$. Deduce that exactly *one half* of the $p-1$ non-zero elements of $\mathbb{Z}_p$ are quadratic residues.

Now consider the *Legendre symbol*

$$\left(\frac{a}{p}\right) \stackrel{\text{def}}{=} \begin{cases} 0, & a = 0, \\ 1, & a \text{ a quadratic residue (mod } p), \\ -1 & a \text{ not a quadratic residue (mod } p). \end{cases}$$

Show that

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) = \left(\frac{ab}{p}\right),$$

and so the Legendre symbol forms a one-dimensional representation of the multiplicative group $(\mathbb{Z}_p)^\times$. Combine this fact with the character orthogonality

theorem to give an alternative proof that precisely half the $p-1$ elements of $(\mathbb{Z}_p)^\times$ are quadratic residues. (Hint: To show that the product of two non-residues is a residue, observe that the set of residues is a normal subgroup of $(\mathbb{Z}_p)^\times$, and consider the multiplication table of the resulting quotient group.)

Exercise 5.22: *More practice with modular arithmetic.* Again let p be an odd prime. Prove *Euler's theorem* that

$$a^{(p-1)/2} \pmod p = \left(\frac{a}{p}\right).$$

(Hint: Begin by showing that the usual school-algebra proof that an equation of degree n can have no more than n solutions remains valid for arithmetic modulo a prime number, and so $a^{(p-1)/2} = 1 \pmod p$ can have no more than$(p-1)/2$ roots. Cite Fermat's little theorem to show that these roots must be the quadratic residues. Cite Fermat again to show that the quadratic non-residues must then have $a^{(p-1)/2} = -1 \pmod p$.)

The harder-to-prove *law of quadratic reciprocity* asserts that for p, q odd primes, we have

$$(-1)^{(p-1)(q-1)/4}\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right).$$

Problem 5.23: *Buckyball spectrum.* Consider the symmetry group of the C_{60} buckyball molecule of figure 5.3.

a) Starting from the character table of the orientation-preserving icosohedral group Y (table 5.3), and using the fact that the $\mathbb{Z}_2$ parity inversion $\sigma : \mathbf{r} \to -\mathbf{r}$ combines with $g \in Y$ so that $D^{J_g}(\sigma g) = D^{J_g}(g)$, whilst $D^{J_u}(\sigma g) = -D^{J_u}(g)$, write down the character table of the extended group $Y_h = Y \times \mathbb{Z}_2$ that acts as a symmetry on the C_{60} molecule. There are now ten conjugacy classes, and the ten representations will be labelled A_g, A_u, *etc.* Verify that your character table has the expected row-orthogonality properties.

b) By counting the number of atoms left fixed by each group operation, compute the compound character of the action of Y_h on the C_{60} molecule. (Hint: Examine the pattern of panels on a regulation soccer ball, and deduce that four carbon atoms are left unmoved by operations in the class σC_2.)

c) Use your compound character from part b), to show that the 60-dimensional Hillbert space decomposes as

$$\mathcal{H}_{C_{60}} = A_g \oplus T_{1g} \oplus 2T_{1u} \oplus T_{2g} \oplus 2T_{2u} \oplus 2G_g \oplus 2G_u \oplus 3H_g \oplus 2H_u,$$

consistent with the energy-levels sketched in figure 5.3.

Problem 5.24: *The Frobenius-Schur Indicator.* Recall that a real or pseudo-real representation is one such that $D(g) \sim D^*(g)$, and for unitary matrices D we have $D^*(g) = [D^T(g)]^{-1}$. In this unitary case $D(g)$ being real or pseudo-real is equivalent to the statement that there exists an invertible matrix F such that

$$FD(g)F^{-1} = [D^T(g)]^{-1}.$$

We can rewrite this statement as $D^T(g)FD(g) = F$, and so F can be interpreted as the matrix representing a G-invariant quadratic form.

i) Use Schur's lemma to show that when D is irreducible the matrix F is unique up to an overall constant. In other words, $D^T(g)F_1D(g) = F_1$ and $D^T(g)F_2D(g) = F_2$ for all $g \in G$ implies that $F_2 = \lambda F_1$. Deduce that for irreducible D we have $F^T = \pm F$.

ii) By reducing F to a suitable canonical form, show that F is symmetric ($F = F^T$) in the case that $D(g)$ is a real representation, and F is skew symmetric ($F = -F^T$) when $D(g)$ is a pseudo-real representation.

iii) Now let G be a *finite* group. For any matrix U, the sum

$$F_U = \frac{1}{|G|} \sum_{g \in G} D^T(g) U D(g)$$

is a G-invariant matrix. Deduce that F_U is always zero when $D(g)$ is neither real nor pseudo-real, and, by specializing both U and the indices on F_U, show that in the real or pseudo-real case

$$\sum_{g \in G} \chi(g^2) = \pm \sum_{g \in G} \chi(g)\chi(g),$$

where $\chi(g) = \operatorname{tr} D(g)$ is the character of the irreducible representation $D(g)$. Deduce that the *Frobenius-Schur indicator*

$$\varkappa \stackrel{\text{def}}{=} \frac{1}{|G|} \sum_{g \in G} \chi(g^2)$$

takes the value $+1$, -1, or 0 when $D(g)$ is, respectively, real, pseudo-real, or not real.

iv) Show that the identity representation occurs in the decomposition of the tensor product $D(g) \otimes D(g)$ of an irrep with itself if, and only if, $D(g)$ is real or pseudo-real. Given a basis $\mathbf{e}_i$ for the vector space V on which $D(g)$ acts, show the matrix F can be used to construct the basis for the identity-representation subspace V^{id} in the decomposition

$$V \otimes V = \bigoplus_{\text{irreps } J} V^J.$$

Problem 5.25: *Induced Representations.* Suppose we know a representation $D^W(h) : W \to W$ for a subgroup $H \subset G$. From this representation we can construct an *induced representation* $\mathrm{Ind}_H^G(D^W)$ for the larger group G. The construction cleverly combines the coset space G/H with the representation space W to make a (usually reducible) representation space $\mathrm{Ind}_H^G(W)$ of dimension $|G/H| \times \dim W$.

Recall that there is a natural action of G on the coset space G/H. If $x = \{g_1, g_2, \ldots\} \in G/H$ then gx is the coset $\{gg_1, gg_2, \ldots\}$. We select from each coset $x \in G/H$ a representative element a_x, and observe that the product ga_x can be decomposed as $ga_x = a_{gx}h$, where a_{gx} is the selected representative from the coset gx and h is some element of H. Next we introduce a basis $|n, x\rangle$ for $\mathrm{Ind}_H^G(W)$. We use the symbol "0" to label the coset $\{e\}$, and take $|n, 0\rangle$ to be the basis vectors for W. For $h \in H$ we can therefore set

$$D(h)|n, 0\rangle \stackrel{\mathrm{def}}{=} |m, 0\rangle D^W_{mn}(h).$$

We also define the result of the action of a_x on $|n, 0\rangle$ to be the vector $|n, x\rangle$:

$$D(a_x)|n, 0\rangle \stackrel{\mathrm{def}}{=} |n, x\rangle.$$

We may now obtain the the action of a general element of G on the vectors $|n, x\rangle$ by requiring $D(g)$ to be representation, and so computing

$$\begin{aligned} D(g)|n, x\rangle &= D(g)D(a_x)|n, 0\rangle \\ &= D(ga_x)|n, 0\rangle \\ &= D(a_{gx}h)|n, 0\rangle \\ &= D(a_{gx})D(h)|n, 0\rangle \\ &= D(a_{gx})|m, 0\rangle D^W_{mn}(h) \\ &= |m, gx\rangle D^W_{mn}(h). \end{aligned}$$

i) Confirm that the action $D(g)|n, x\rangle = |m, gx\rangle D^W_{mn}(h)$, with h obtained from g and x *via* the decomposition $ga_x = a_{gx}h$, does indeed define a representation of G. Show also that if we set $|f\rangle = \sum_{n,x} f_n(x)|n, x\rangle$, then the action of g on the components takes

$$f_n(x) \mapsto D^W_{nm}(h) f_m(g^{-1}x).$$

ii) Let $f(h)$ be a class function on H. Let us extend it to a function on G by setting $f(g) = 0$ if $g \notin H$, and define

$$\mathrm{Ind}_H^G[f](s) = \frac{1}{|H|} \sum_{g \in G} f(g^{-1}sg).$$

Show that $\mathrm{Ind}_H^G[f](s)$ is a class function on G, and further show that if χ_W is the character of the starting representation for H then $\mathrm{Ind}_H^G[\chi_W]$ is the character of the induced representation of G. (Hint, only fixed points of the G-action on G/H contribute to the character, and $gx = x$ means that $ga_x = a_x h$. Thus $D^W(h) = D^W(a_x^{-1} g a_x)$.)

iii) Given a representation $D^V(g) : V \to V$ of G we can trivially obtain a (generally reducible) representation $\mathrm{Res}_H^G(V)$ of $H \subset G$ by restricting G to H. Define the usual inner product on the group functions by

$$\langle \phi_1, \phi_2 \rangle_G = \frac{1}{|G|} \sum_{g \in G} \phi_1(g^{-1}) \phi_2(g),$$

and show that if ψ is a class function on H and ϕ a class function on G then

$$\langle \psi, \mathrm{Res}_H^G[\phi] \rangle_H = \langle \mathrm{Ind}_H^G[\psi], \phi \rangle_G.$$

Thus, Ind_H^G and Res_H^G are, in some sense, adjoint operations. Mathematicians would call them a pair of mutually *adjoint functors*.

iv) By applying the result from part (iii) to the characters of the irreducible representations of G and H, deduce *Frobenius' reciprocity theorem*: The number of times an irrep $D^J(g)$ of G occurs in the representation induced from an irrep $D^K(h)$ of H is equal to the number of times that D^K occurs in the decomposition of D^J into irreps of H.

The representation of the Poincaré group (= the SO(1, 3) Lorentz group together with space-time translations) that classifies the states of a spin-J elementary particle are those induced from the spin-J representation of its SO(3) rotation subgroup. The quantum state of a mass m elementary particle is therefore of the form $|k, \sigma\rangle$ where k is the particle's four-momentum, which lies is the coset SO(1, 3)/SO(3), and σ is the label from the $|J, \sigma\rangle$ spin state.

Chapter 6

Lie Groups

Lie groups are named after the Norwegian mathematician Sophus Lie. They consist of a manifold G equipped with a group multiplication rule $(g_1, g_2) \mapsto g_3$ which is a smooth function of the g's, as is the operation of taking the inverse of a group element. The most commonly met examples in physics are the infinite families of *matrix groups* $\mathrm{GL}(n)$, $\mathrm{SL}(n)$, $\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$, $\mathrm{SU}(n)$, and $\mathrm{Sp}(n)$, togther with the family of five *exceptional* Lie groups: G_2, F_4, E_6, E_7, and E_8, which have applications in string theory.

One of the properties of a Lie group is that, considered as a manifold, the neighbourhood of any point looks exactly like that of any other. The group's dimension and most of its structure can be understood by examining the immediate vicinity any chosen point, which we may as well take to be the identity element. The vectors lying in the tangent space at the identity element make up the *Lie algebra* of the group. Computations in the Lie algebra are often easier than those in the group, and provide much of the same information. This chapter will be devoted to studying the interplay between the Lie group itself and this Lie algebra of infinitesimal elements.

6.1 Matrix Groups

The *Classical Groups* are described in a book with this title by Hermann Weyl. They are subgroups of the *general linear group*, $\mathrm{GL}(n, \mathbb{F})$, which consists of invertible n-by-n matrices over the field $\mathbb{F}$. We will mostly consider the cases $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$.

A near-identity matrix in $\mathrm{GL}(n, \mathbb{R})$ can be written $g = I + \epsilon A$ where A

is an arbitrary n-by-n real matrix. This matrix contains n^2 real entries, so we can move away from the identity in n^2 distinct directions. The tangent space at the identity, and hence the group manifold itself, is therefore n^2 dimensional. The manifold of $\mathrm{GL}(n,\mathbb{C})$ has n^2 *complex* dimensions, and this corresponds to $2n^2$ real dimensions.

If we restrict the determinant of a $\mathrm{GL}(n,\mathbb{F})$ matrix to be unity, we get the *special linear group*, $\mathrm{SL}(n,\mathbb{F})$. An element near the identity in this group can still be written as $g = I + \epsilon A$, but since

$$\det(I + \epsilon A) = 1 + \epsilon \operatorname{tr}(A) + O(\epsilon^2) \tag{6.1}$$

this requires $\operatorname{tr}(A) = 0$. The restriction on the trace means that $\mathrm{SL}(n,\mathbb{R})$ has dimension $n^2 - 1$.

6.1.1 The Unitary and Orthogonal Groups

Perhaps the most important of the matrix groups are the unitary and orthogonal groups.

The Unitary group

The unitary group $\mathrm{U}(n)$ comprises the set of n-by-n complex matrices U such that $U^\dagger = U^{-1}$. If we consider matrices near the identity

$$U = I + \epsilon A, \tag{6.2}$$

with ϵ real, then unitarity requires

$$\begin{aligned} I + O(\epsilon^2) &= (I + \epsilon A)(I + \epsilon A^\dagger) \\ &= I + \epsilon(A + A^\dagger) + O(\epsilon^2), \end{aligned} \tag{6.3}$$

so $A_{ij} = -A^*_{ji}$ and A is skew hermitian. A complex skew-hermitian matrix contains

$$n + 2 \times \frac{1}{2} n(n-1) = n^2$$

real parameters. In this counting the first "n" is the number of entries on the diagonal, each of which must be of the form i times a real number. The $n(n-1)/2$ is the number of entries above the main diagonal, each of which can be an arbitrary complex number. The number of real dimensions in the

group manifold is therefore n^2. The rows or columns in the matrix U form an orthonormal set of vectors. Their entries are therefore bounded, $|U_{ij}| \leq 1$, and this property leads to the n^2 dimensional group manifold of $\mathrm{U}(n)$ being a compact set.

When a group manifold is compact, we say that the group itself is a *compact group*. There is a natural notion of volume on a group manifold and compact Lie groups have finite total volume. Because of this, they have many properties in common with the finite groups we studied in the last chapter.

Recall that a group is *simple* if it possesses no invariant subgroups. $\mathrm{U}(n)$ is not simple. Its centre is an invariant $\mathrm{U}(1)$ subgroup consisting of matrices of the form $U = e^{i\theta} I$. The *special unitary group* $\mathrm{SU}(n)$, consists of n-by-n unimodular (having determinant $+1$) unitary matrices. It is not strictly simple because its center Z consists of the discrete subgroup of matrices $U_m = \omega^m I$ with ω an n-th root of unity, and this is an invariant subgroup. Because Z, its only invariant subgroup, is not a continuous group, $\mathrm{SU}(n)$ is counted as being simple in Lie theory. With $U = I + \epsilon A$, as above, the unimodularity imposes the additional constraint on A that $\mathrm{tr}\, A = 0$, so the $\mathrm{SU}(n)$ group manifold is $n^2 - 1$ dimensional.

The Orthogonal Group

The orthogonal group $\mathrm{O}(n)$, consists of the the set of real matrices O with the property that $O^T = O^{-1}$. For a matrix in the neighbourhood of the identity, $O = I + \epsilon A$, this condition requires that A be skew symmetric: $A_{ij} = -A_{ij}$. Skew symmetric real matrices have $n(n-1)/2$ independent entries, and so the group manifold of $\mathrm{O}(n)$ is $n(n-1)/2$ dimensional. The condition $O^T O = I$ means that the rows or columns of O, considered as row or column vectors, are orthonormal. All entries are bounded $|O_{ij}| \leq 1$, and again this leads to $\mathrm{O}(n)$ being a compact group.

The identity

$$1 = \det\left(O^T O\right) = \det O^T \det O = (\det O)^2 \tag{6.4}$$

tells us that $\det O = \pm 1$. The subset of orthogonal matrices with $\det O = +1$ constitute a subgroup of $\mathrm{O}(n)$ called the *special orthogonal group*, $\mathrm{SO}(n)$. The unimodularity condition discards a disconnected part of the group manifold and does not reduce its dimension, which remains $n(n-1)/2$.

6.1.2 Symplectic Groups

The symplectic groups (named from Greek meaning to "fold together") are probably less familiar than the other matrix groups.

We start with a non-degenerate skew-symmetric matrix ω. The symplectic group $\mathrm{Sp}(2n, \mathbb{F})$ is then defined by

$$\mathrm{Sp}(2n, \mathbb{F}) = \{S \in \mathrm{GL}(2n, \mathbb{F}) : S^T \omega S = \omega\}. \tag{6.5}$$

Here $\mathbb{F}$ can be $\mathbb{R}$ or $\mathbb{C}$. When $\mathbb{F} = \mathbb{C}$, we still use the transpose "T," not †, in this definition. Setting $S = I_{2n} + \epsilon A$ and demanding that $S^T \omega S = \omega$ shows that $A^T \omega + \omega A = 0$.

It does not matter what skew matrix ω we start from, because we can always find a basis in which ω takes its canonical form:

$$\omega = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}. \tag{6.6}$$

In this basis we find, after a short computation, that the most general form for A is

$$A = \begin{pmatrix} a & b \\ c & -a^T \end{pmatrix}. \tag{6.7}$$

Here a is any n-by-n matrix, and b and c are symmetric ($b^T = b$ and $c^T = c$) n-by-n matrices. If the matrices are real, then counting the degrees of freedom gives the dimension of the *real symplectic group* as

$$\dim \mathrm{Sp}(2n, \mathbb{R}) = n^2 + 2 \times \frac{n}{2}(n+1) = n(2n+1). \tag{6.8}$$

The entries in a, b, c can be arbitrarily large. $\mathrm{Sp}(2n, \mathbb{R})$ is not compact.

The determinant of any symplectic matrix is $+1$. To see this take the elements of ω to be ω_{ij}, and let

$$\omega(x, y) = \omega_{ij} x^i y^j \tag{6.9}$$

be the associated skew bilinear (*not* sesquilinear) form . Then Weyl's identity from exercise **??.??** shows that

$$\begin{aligned} &\mathrm{Pf}\,(\omega)\,(\det M)\det|x_1, \ldots x_{2n}| \\ &\quad = \frac{1}{2^n n!} \sum_{\pi \in S_{2n}} \mathrm{sgn}\,(\pi) \omega(Mx_{\pi(1)}, Mx_{\pi(2)}) \cdots \omega(Mx_{\pi(2n-1)}, Mx_{\pi(2n)}), \end{aligned}$$

for any linear map M. If $\omega(x,y) = \omega(Mx, My)$, we conclude that $\det M = 1$ — but preserving ω is exactly the condition that M be an element of the symplectic group. Since the matrices in $\mathrm{Sp}(2n, \mathbb{F})$ are automatically unimodular there is no "special symplectic" group.

Unitary Symplectic Group

The intersection of two groups is also a group. We therefore define the *unitary symplectic group* as

$$\mathrm{Sp}(n) = \mathrm{Sp}(2n, \mathbb{C}) \cap \mathrm{U}(2n). \tag{6.10}$$

This group is compact. We will see that its dimension is $n(2n+1)$, the same as the non-compact $\mathrm{Sp}(2n, \mathbb{R})$. $\mathrm{Sp}(n)$ may also be defined as $\mathrm{U}(n, \mathbb{H})$ where $\mathbb{H}$ denotes the skew field of quaternions.

Warning: Physics papers often make no distinction between $\mathrm{Sp}(n)$, which is a compact group, and $\mathrm{Sp}(2n, \mathbb{R})$ which is non-compact. To add to the confusion the compact $\mathrm{Sp}(n)$ is also sometimes called $\mathrm{Sp}(2n)$. You have to judge from the context what group the author has in mind.

Physics Application: Kramers' degeneracy. Let $C = i\hat{\sigma}_2$. Therefore

$$C^{-1}\hat{\sigma}_n C = -\hat{\sigma}_n^*. \tag{6.11}$$

A time-reversal invariant Hamiltonian containing $\mathbf{L}\cdot\mathbf{S}$ spin-orbit interactions obeys

$$C^{-1}HC = H^*. \tag{6.12}$$

If we regard the $2n$-by-$2n$ matrix H as being an n-by-n matrix whose entries H_{ij} are themselves 2-by-2 matrices, which we expand as

$$H_{ij} = h_{ij}^0 + i\sum_{n=1}^{3} h_{ij}^n \hat{\sigma}_n,$$

then the condition (6.12) implies that the h_{ij}^a are real numbers. We say that H is *real quaternionic*. This is because the Pauli sigma matrices are algebraically isomorphic to Hamilton's quaternions under the identification

$$\begin{aligned} i\hat{\sigma}_1 &\leftrightarrow \mathbf{i}, \\ i\hat{\sigma}_2 &\leftrightarrow \mathbf{j}, \\ i\hat{\sigma}_3 &\leftrightarrow \mathbf{k}. \end{aligned} \tag{6.13}$$

The hermiticity of H requires that $H_{ji} = \overline{H}_{ij}$ where the overbar denotes quaternionic conjugation

$$q^0 + iq^1\hat{\sigma}_1 + iq^2\hat{\sigma}_2 + iq^3\hat{\sigma}_3 \to q^0 - iq^1\hat{\sigma}_1 - iq^2\hat{\sigma}_2 - iq^3\hat{\sigma}_3. \tag{6.14}$$

If $H\psi = E\psi$, then $HC\psi^* = E\psi^*$. Since C is skew, ψ and $C\psi^*$ are necessarily orthogonal. Therefore all states are doubly degenerate. This is *Kramers'* degeneracy.

H may be diagonalized by a matrix in $\mathrm{U}(n, \mathbb{H})$, where $\mathrm{U}(n, \mathbb{H})$ consists of those elements of $\mathrm{U}(2n)$ that satisfy $C^{-1}UC = U^*$. We may rewrite this condition as

$$C^{-1}UC = U^* \Rightarrow UCU^T = C,$$

so $\mathrm{U}(n, \mathbb{H})$ consists of the unitary matrices that preserve the skew matrix C. Thus $\mathrm{U}(n, \mathbb{H}) \subseteq \mathrm{Sp}(n)$. Further investigation shows that $\mathrm{U}(n, \mathbb{H}) = \mathrm{Sp}(n)$.

We can exploit the quaternionic viewpoint to count the dimensions. Let $U = I + \epsilon B$ be in $\mathrm{U}(n, \mathbb{H})$, then $B_{ij} + \overline{B}_{ji} = 0$. The diagonal elements of B are thus pure "imaginary" quaternions having no part proportional to I. There are therefore 3 parameters for each diagonal element. The upper triangle has $n(n-1)/2$ independent elements, each with 4 parameters. Counting up, we find

$$\dim \mathrm{U}(n, \mathbb{H}) = \dim \mathrm{Sp}(n) = 3n + 4 \times \frac{n}{2}(n-1) = n(2n+1). \tag{6.15}$$

Thus, as promised, we see that the compact group $\mathrm{Sp}(n)$ and the non-compact group $\mathrm{Sp}(2n, \mathbb{R})$ have the same dimension.

We can also count the dimension of $\mathrm{Sp}(n)$ by looking at our previous matrices

$$A = \begin{pmatrix} a & b \\ c & -a^T \end{pmatrix}$$

where a b and c are now allowed to be complex, but with the restriction that $S = I + \epsilon A$ be unitary. This requires A to be skew-hermitian, so $a = -a^\dagger$, and $c = -b^\dagger$, while b (and hence c) remains symmetric. There are n^2 free real parameters in a, and $n(n+1)$ in b, so

$$\dim \mathrm{Sp}(n) = (n^2) + n(n+1) = n(2n+1)$$

as before.

Exercise 6.1: Show that

$$\mathrm{SO}(2N) \cap \mathrm{Sp}(2N, \mathbb{R}) \cong \mathrm{U}(N).$$

Hint: Group the $2N$ basis vectors on which $\mathrm{O}(2N)$ acts into pairs $\mathbf{x}_n$ and $\mathbf{y}_n$, $n = 1, \ldots, N$. Assemble these pairs into $\mathbf{z}_n = \mathbf{x}_n + i\mathbf{y}_n$ and $\bar{\mathbf{z}} = \mathbf{x}_n - i\mathbf{y}_n$. Let ω be the linear map that takes $\mathbf{x}_n \to \mathbf{y}_n$ and $\mathbf{y}_n \to -\mathbf{x}_n$. Show that the subset of $\mathrm{SO}(2N)$ that commutes with ω mixes $\mathbf{z}_i$'s only with $\mathbf{z}_i$'s and $\bar{\mathbf{z}}_i$'s only with $\bar{\mathbf{z}}_i$'s.

6.2 Geometry of SU(2)

To get a sense of Lie groups as geometric objects, we will study the simplest non-trivial case of SU(2) in some detail.

A general 2-by-2 complex matrix can be parametrized as

$$U = \begin{pmatrix} x^0 + ix^3 & ix^1 + x^2 \\ ix^1 - x^2 & x^0 - ix^3 \end{pmatrix}. \tag{6.16}$$

The determinant of this matrix is unity provided

$$(x^0)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 = 1. \tag{6.17}$$

When this condition is met, and if in addition the x^i are real, the matrix is unitary: $U^\dagger = U^{-1}$. The group manifold of SU(2) can therefore be identified with the three-sphere S^3. We will take as local co-ordinates x^1, x^2, x^3. When we desire to know x^0 we will find it from $x^0 = \sqrt{1 - (x^1)^2 - (x^2)^2 - (x^3)^2}$. This co-ordinate chart only labels the points in the half of the three-sphere with $x^0 > 0$, but this is typical of any non-trivial manifold. A complete atlas of charts can be constructed if needed.

We can simplify our notation by using the Pauli sigma matrices

$$\hat{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \hat{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \hat{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{6.18}$$

These obey

$$[\hat{\sigma}_i, \hat{\sigma}_j] = 2i\epsilon_{ijk}\hat{\sigma}_k, \quad \text{and} \quad \sigma_i, \hat{\sigma}_j + \hat{\sigma}_j\hat{\sigma}_i = 2\delta_{ij}I. \tag{6.19}$$

In terms of them, we can write

$$g = U = x^0 I + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3. \tag{6.20}$$

Elements of the group in the neighbourhood of the identity differ from $e \equiv I$ by real linear combinations of the $i\hat{\sigma}_i$. The three-dimensional vector space spanned by these matrices is therefore the tangent space TG_e at the identity element. For any Lie group this tangent space is called the *Lie algebra*, $\mathfrak{g} = \mathrm{Lie}\, G$ of the group. There will be a similar set of matrices $i\hat{\lambda}_i$ for any matrix group. They are called the *generators* of the Lie algebra, and satisfy commutation relations of the form

$$[i\hat{\lambda}_i, i\hat{\lambda}_j] = -f_{ij}{}^k (i\hat{\lambda}_k), \tag{6.21}$$

or equivalently

$$[\hat{\lambda}_i, \hat{\lambda}_j] = i f_{ij}{}^k \hat{\lambda}_k \tag{6.22}$$

The $f_{ij}{}^k$ are called the *structure constants* of the algebra. The "i"'s associated with the $\hat{\lambda}$'s in this expression are conventional in physics texts because for quantum mechanics application we usually desire the $\hat{\lambda}_i$ to be hermitian. They are usually absent in books written for mathematicians.

Exercise 6.2: Let $\hat{\lambda}_1$ and $\hat{\lambda}_2$ be hermitian matrices. Show that if we define $\hat{\lambda}_3$ by the relation $[\hat{\lambda}_1, \hat{\lambda}_2] = i\hat{\lambda}_3$, then $\hat{\lambda}_3$ is also a hermitian matrix.

Exercise 6.3: For the group $\mathrm{O}(n)$ the matrices "$i\hat{\lambda}$" are real n-by-n skew symmetric matrices A. Show that if A_1 and A_2 are real skew symmetric matrices, then so is $[A_1, A_2]$.

Exercise 6.4: For the group $\mathrm{Sp}(2n, \mathbb{R})$ the $i\hat{\lambda}$ matrices are of the form

$$A = \begin{pmatrix} a & b \\ c & -a^T \end{pmatrix}$$

where a is any real n-by-n matrix and b and c are symmetric ($a^T = a$ and $b^T = b$) real n-by-n matrices. Show that the commutator of any two matrices of this form is also of this form.

6.2.1 Invariant vector fields

Consider a matrix group, and in it a group element $I + i\epsilon\hat{\lambda}_i$ lying close to the identity $e \equiv I$. Draw an arrow connecting I to $I + i\epsilon\hat{\lambda}_i$, and regard this arrow as a vector L_i lying in TG_e. Next map the infinitesimal element $I + i\epsilon\hat{\lambda}_i$ to the neighbourhood an arbitrary group element g by multiplying

on the *left* to get $g(I+i\epsilon\hat{\lambda}_i)$. By drawing an arrow from g to $g(I+i\epsilon\hat{\lambda}_i)$, we obtain a vector $L_i(g)$ lying in TG_g. This vector at g is the push forward of the vector at e by left multiplication by g. For example, consider SU(2) with infinitesimal element $I+i\epsilon\hat{\sigma}_3$. We find

$$\begin{aligned} g(I+i\epsilon\hat{\sigma}_3) &= (x^0+ix^1\hat{\sigma}_1+ix^2\hat{\sigma}_2+ix^3\hat{\sigma}_3)(I+i\epsilon\hat{\sigma}_3) \\ &= (x^0-\epsilon x^3)+i\hat{\sigma}_1(x^1-\epsilon x^2)+i\hat{\sigma}_2(x^2+\epsilon x^1)+i\hat{\sigma}_3(x^3+\epsilon x^0). \end{aligned} \tag{6.23}$$

This computation can also be interpreted as showing that the multiplication of $g\in \mathrm{SU}(2)$ on the *right* by $(I+i\epsilon\hat{\sigma}_3)$ displaces the point g, changing its x^i parameters by an amount

$$\delta\begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \epsilon\begin{pmatrix} -x^3 \\ -x^2 \\ x^1 \\ x^0 \end{pmatrix}. \tag{6.24}$$

Knowing how the displacement looks in terms of the x^1, x^2, x^3 co-ordinate system lets us read off the $\partial/\partial x^\mu$ components of the vector L_3 lying in TG_g

$$L_3 = -x^2\partial_1+x^1\partial_2+x^0\partial_3. \tag{6.25}$$

Since g can be any point in the group, we have constructed a globally defined vector field L_3 that acts on a function $F(g)$ on the group manifold as

$$L_3F(g) = \lim_{\epsilon\to 0}\left\{\frac{1}{\epsilon}\left[F\left(g(I+i\epsilon\hat{\sigma}_3)\right)-F(g)\right]\right\}. \tag{6.26}$$

Similarly we obtain

$$\begin{aligned} L_1 &= x^0\partial_1-x^3\partial_2+x^2\partial_3 \\ L_2 &= x^3\partial_1+x^0\partial_2-x^1\partial_3. \end{aligned} \tag{6.27}$$

The vector fields L_i are said to be *left invariant* because the push-forward of the vector $L_i(g)$ lying in the tangent space at g by multiplication on the left by any g' produces a vector $g'_*[L_i(g)]$ lying in the tangent space at $g'g$, and this pushed-forward vector coincides with the $L_i(g'g)$ already there. We can express this statement tersely as $g_*L_i = L_i$.

Using $\partial_i x^0 = -x^i/x_0$, $i = 1, 2, 3$, we can compute the Lie brackets and find

$$[L_1, L_2] = -2L_3. \tag{6.28}$$

In general

$$[L_i, L_j] = -2\epsilon_{ijk}L_k, \tag{6.29}$$

which coincides with the matrix commutator of the $i\hat{\sigma}_i$.

This construction works for all Lie groups. For each basis vector L_i in the tangent space at the identity e, we push it forward to the tangent space at g by left multiplication by g, and so construct the global left-invariant vector field L_i. The Lie bracket of these vector fields will be

$$[L_i, L_j] = -f_{ij}{}^k L_k, \tag{6.30}$$

where the coefficients $f_{ij}{}^k$ are guaranteed to be position independent because (see exercise 3.5) the operation of taking the Lie bracket of two vector fields commutes with the operation of pushing-forward the vector fields. Consequently the Lie bracket at any point is just the image of the Lie bracket calculated at the identity. When the group is a matrix group, this Lie bracket will coincide with the commutator of the $i\hat{\lambda}_i$, that group's analogue of the $i\hat{\sigma}_i$ matrices.

The Exponential Map

Recall that given a vector field $X \equiv X^\mu \partial_\mu$ we define associated *flow* by solving the equation

$$\frac{dx^\mu}{dt} = X^\mu(x(t)). \tag{6.31}$$

If we do this for the left-invariant vector field L, with initial condition $x(0) = e$, we obtain a t-dependent group element $g(x(t))$, which we denote by $\mathrm{Exp}\,(tL)$. The symbol "Exp" stands for the *exponential map* which takes elements of the Lie algebra to elements of the Lie group. The reason for the name and notation is that for matrix groups this operation corresponds to the usual exponentiation of matrices. Elements of the matrix Lie group are therefore exponentials of matrices in the the Lie algebra. To see this suppose that L_i is the left invariant vector field derived from $i\hat{\lambda}_i$. Then the matrix

$$g(t) = \exp(it\hat{\lambda}_i) \equiv I + it\hat{\lambda}_i - \frac{1}{2}t^2\hat{\lambda}^2 - i\frac{1}{3!}t^3\hat{\lambda}^3 + \cdots \tag{6.32}$$

is an element of the group, and

$$g(t+\epsilon) = \exp(it\hat{\lambda})\exp(i\epsilon\hat{\lambda}_i) = g(t)\left(I + i\epsilon\hat{\lambda}_i + O(\epsilon^2)\right). \tag{6.33}$$

From this we deduce that

$$\frac{d}{dt}g(t) = \lim_{\epsilon\to 0}\left\{\frac{1}{\epsilon}[g(t)(I+i\epsilon\hat{\lambda}_i) - g(t)]\right\} = L_i g(t). \tag{6.34}$$

Since $\exp(it\hat{\lambda}) = I$ when $t = 0$, we deduce that $\mathrm{Exp}\,(tL_i) = \exp(it\hat{\lambda}_i)$.

Right-invariant vector fields

We can use multiplication on the *right* to push forward an infinitesimal group element. For example:

$$\begin{aligned}(I+i\epsilon\hat{\sigma}_3)g &= (I+i\epsilon\hat{\sigma}_3)(x^0 + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3)\\ &= (x^0 - \epsilon x^3) + i\hat{\sigma}_1(x^1 + \epsilon x^2) + i\hat{\sigma}_2(x^2 - \epsilon x^1) + i\hat{\sigma}_3(x^3 + \epsilon x^0).\end{aligned} \tag{6.35}$$

This motion corresponds to the *right-invariant vector field*

$$R_3 = x^2\partial_1 - x^1\partial_2 + x^0\partial_3. \tag{6.36}$$

Similarly, we obtain

$$\begin{aligned}R_1 &= x^3\partial_1 - x^0\partial_2 + x^1\partial_3\\ R_2 &= x^0\partial_1 + x^3\partial_2 - x^2\partial_3,\end{aligned} \tag{6.37}$$

and find that

$$[R_1, R_2] = +2R_3. \tag{6.38}$$

In general,

$$[R_i, R_j] = +2\epsilon_{ijk}R_k. \tag{6.39}$$

For any Lie group, the Lie brackets of the right-invariant fields will be

$$[R_i, R_j] = +f_{ij}{}^k R_k. \tag{6.40}$$

whenever

$$[L_i, L_j] = -f_{ij}{}^k L_k, \tag{6.41}$$

are the Lie brackets of the left-invariant fields. The relative minus sign between the bracket algebra of the left and right invariant vector fields has the same origin as the relative sign between the commutators of space- and body-fixed rotations in classical mechanics. Because multiplication from the left does not interfere with multiplication from the right, the left and right invariant fields commute:

$$[L_i, R_j] = 0. \tag{6.42}$$

6.2.2 Maurer-Cartan Forms

If $g \in G$, then $dg\, g^{-1} \in \mathrm{Lie}\, G$. For example, starting from

$$\begin{aligned} g &= x^0 + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3 \\ g^{-1} &= x^0 - ix^1\hat{\sigma}_1 - ix^2\hat{\sigma}_2 - ix^3\hat{\sigma}_3 \end{aligned} \tag{6.43}$$

we have

$$\begin{aligned} dg &= dx^0 + idx^1\hat{\sigma}_1 + idx^2\hat{\sigma}_2 + idx^3\hat{\sigma}_3 \\ &= (x^0)^{-1}(-x^1dx^1 - x^2dx^2 - x^3dx^3) + idx^1\hat{\sigma}_1 + idx^2\hat{\sigma}_2 + idx^3\hat{\sigma}_3. \end{aligned} \tag{6.44}$$

From this we find

$$\begin{aligned} dgg^{-1} &= i\hat{\sigma}_1\left((x^0 + (x^1)^2/x^0)dx^1 + (x^3 + (x^1x^2)/x^0)dx^2 + (-x^2 + (x^1x^3)/x^0)dx^3\right) \\ &\quad + i\hat{\sigma}_2\left((-x^3 + (x^2x^1)/x^0)dx^1 + (x^0 + (x^2)^2/x^0)dx^2 + (x^1 + (x^2x^3)/x^0)dx^3\right) \\ &\quad + i\hat{\sigma}_3\left((x^2 + (x^3x^1)/x^0)dx^1 + (-x^1 + (x^3x^2)/x^0)dx^2 + (x^0 + (x^3)^2/x^0)dx^3\right). \end{aligned} \tag{6.45}$$

The part proportional to the identity matrix has cancelled. The result is therefore a Lie algebra-valued 1-form. We define the (right invariant) Maurer-Cartan forms ω_R^i by

$$dgg^{-1} = \omega_R = (i\hat{\sigma}_i)\omega_R^i. \tag{6.46}$$

If we evaluate one-form ω_R^1 on the right invariant vector field R_1, we find

$$\begin{aligned} \omega_R^1(R_1) &= (x^0 + (x^1)^2/x^0)x^0 + (x^3 + (x^1x^2)/x^0)x^3 + (-x^2 + (x^1x^3)/x^0)(-x^2) \\ &= (x^0)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 \\ &= 1. \end{aligned} \tag{6.47}$$

Working similarly, we find

$$\begin{aligned}\omega^1_R(R_2) &= (x^0 + (x^1)^2/x^0)(-x^3) + (x^3 + (x^1x^2)/x^0)x^0 + (-x^2 + (x^1x^3)/x^0)x^1 \\ &= 0. \end{aligned} \tag{6.48}$$

In general we discover that $\omega^i_R(R_j) = \delta^i_j$. These Maurer-Cartan forms therefore constitute the dual basis to the right-invariant vector fields.

We may also define the left invariant Maurer-Cartan forms

$$g^{-1}dg = \omega_L = (i\hat{\sigma}_i)\omega^i_L. \tag{6.49}$$

These obey $\omega^i_L(L_j) = \delta^i_j$, showing that the ω^i_L are the dual basis to the left-invariant vector fields.

Acting with the exterior derivative d on $gg^{-1} = I$ tells us that $d(g^{-1}) = -g^{-1}dgg^{-1}$. By exploiting this fact, together with the anti-derivation property

$$d(a \wedge b) = da \wedge b + (-1)^p a \wedge db,$$

we may compute the exterior derivative of ω_R. We find that

$$d\omega_R = d(dgg^{-1}) = (dgg^{-1}) \wedge (dgg^{-1}) = \omega_R \wedge \omega_R. \tag{6.50}$$

A matrix product is implicit here. If it were not, the product of the two identical 1-forms on the right would automatically be zero. If we make this matrix structure explicit we find that

$$\begin{aligned}\omega_R \wedge \omega_R &= \omega^i_R \wedge \omega^j_R (i\hat{\sigma}_i)(i\hat{\sigma}_j) \\ &= \frac{1}{2}\omega^i_R \wedge \omega^j_R \, [i\hat{\sigma}_i, i\hat{\sigma}_j] \\ &= -\frac{1}{2} f_{ij}{}^k (i\hat{\sigma}_k)\, \omega^i_R \wedge \omega^j_R, \end{aligned} \tag{6.51}$$

so

$$d\omega^k_R = -\frac{1}{2} f_{ij}{}^k \omega^i_R \wedge \omega^j_R. \tag{6.52}$$

These equations are known as the *Maurer-Cartan relations* for the right-invariant forms.

For the left-invariant forms we have

$$d\omega_L = d(g^{-1}dg) = -(g^{-1}dg) \wedge (g^{-1}dg) = -\omega_L \wedge \omega_L, \tag{6.53}$$

or

$$d\omega_L^k = +\frac{1}{2} f_{ij}{}^k \omega_L^i \wedge \omega_L^j. \tag{6.54}$$

The Maurer-Cartan relations appear when we quantize gauge theories. They are one part of the BRST transformations of the Fadeev-Popov ghost fields.

6.2.3 Euler Angles

In physics it is common to use *Euler angles* to parameterize SU(2). We can write an arbitrary SU(2) matrix U as a product

$$\begin{aligned} U &= \exp\{-i\phi\hat{\sigma}_3/2\}\exp\{-i\theta\hat{\sigma}_2/2\}\exp\{-i\psi\hat{\sigma}_3/2\}, \\ &= \begin{pmatrix} e^{-i\phi/2} & 0 \\ 0 & e^{i\phi/2} \end{pmatrix} \begin{pmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{pmatrix} \begin{pmatrix} e^{-i\psi/2} & 0 \\ 0 & e^{i\psi/2} \end{pmatrix}, \\ &= \begin{pmatrix} e^{-i(\phi+\psi)/2}\cos\theta/2 & -e^{i(\psi-\phi)/2}\sin\theta/2 \\ e^{i(\phi-\psi)/2}\sin\theta/2 & e^{+i(\psi+\phi)/2}\cos\theta/2 \end{pmatrix}. \end{aligned} \tag{6.55}$$

Comparing with the earlier expression for U in terms of the x^μ, we obtain the Euler-angle parameterization of the three-sphere

$$\begin{aligned} x^0 &= \cos\theta/2\cos(\psi+\phi)/2, \\ x^1 &= \sin\theta/2\sin(\phi-\psi)/2, \\ x^2 &= -\sin\theta/2\cos(\phi-\psi)/2, \\ x^3 &= -\cos\theta/2\sin(\psi+\phi)/2. \end{aligned} \tag{6.56}$$

If the angles are taken in the range $0 \le \phi < 2\pi$, $0 \le \theta < \pi$, $0 \le \psi < 4\pi$ we cover the entire three-sphere once.

Exercise 6.5: Show that the Hopf map, defined in chapter 3, Hopf : $S^3 \to S^2$ is the "forgetful" map $(\theta, \phi, \psi) \to (\theta, \phi)$, where θ and ϕ are spherical polar co-ordinates on the two-sphere.

Exercise 6.6: Show that

$$U^{-1}dU = -\frac{i}{2}\hat{\sigma}_i\,\Omega_{\rm L}^i,$$

where

$$\begin{aligned} \Omega_{\rm L}^1 &= \sin\psi\,d\theta - \sin\theta\cos\psi\,d\phi, \\ \Omega_{\rm L}^2 &= \cos\psi\,d\theta - \sin\theta\sin\psi\,d\phi, \\ \Omega_{\rm L}^3 &= d\psi + \cos\theta\,d\phi. \end{aligned}$$

Compare these 1-forms with the components

$$\begin{aligned}\omega_X &= \sin\psi\,\dot\theta - \sin\theta\cos\psi\,\dot\phi,\\ \omega_Y &= \cos\psi\,\dot\theta - \sin\theta\sin\psi\,\dot\phi,\\ \omega_Z &= \dot\psi + \cos\theta\,\dot\phi.\end{aligned}$$

of the angular velocity $\boldsymbol{\omega}$ of a body with respect to the *body-fixed* XYZ axes in the Euler-angle conventions of exercise 2.17.

Similarly show that

$$dUU^{-1} = -\frac{i}{2}\hat\sigma_i\,\Omega^i_{\rm R},$$

where

$$\begin{aligned}\Omega^1_{\rm R} &= -\sin\phi\,d\theta + \sin\theta\cos\psi\,d\psi,\\ \Omega^2_{\rm R} &= \cos\phi\,d\theta + \sin\theta\sin\psi\,d\psi,\\ \Omega^3_{\rm R} &= d\phi + \cos\theta\,d\psi,\end{aligned}$$

Compare these 1-forms with components ω_x, ω_y, ω_z of the same angular velocity vector $\boldsymbol{\omega}$, but now with respect to the *space-fixed* xyz frame.

6.2.4 Volume and Metric

The manifold of any Lie group has a natural metric which is obtained by transporting the Killing form (see section 6.3.2) from the tangent space at the identity to any other point g by either left or right multiplication by g. In the case of a compact group, the resultant left and right invariant metrics coincide. In the case of SU(2) this metric is the usual metric on the three-sphere.

Using the Euler angle expression for the x^μ to compute the dx^μ, we can express the metric on the sphere as

$$\begin{aligned}\text{“}ds^{2\prime\prime} &= (dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2,\\ &= \frac{1}{4}\left(d\theta^2 + \cos^2\theta/2(d\psi + d\phi)^2 + \sin^2\theta/2(d\psi - d\phi)^2\right),\\ &= \frac{1}{4}\left(d\theta^2 + d\psi^2 + d\phi^2 + 2\cos\theta d\phi d\psi\right).\end{aligned} \tag{6.57}$$

Here, to save space, we have used the traditional physics way of writing a metric. In the more formal notation, where we think of the metric as being

a bilinear function, we would write the last line as

$$g(\ \ ,\ \ ) = \frac{1}{4}\left(d\theta \otimes d\theta + d\psi \otimes d\psi + d\phi \otimes d\phi + \cos\theta(d\phi \otimes d\psi + d\psi \otimes d\phi)\right) \tag{6.58}$$

From (6.58) we find

$$\begin{aligned} g = \det\left(g_{\mu\nu}\right) &= \frac{1}{4^3}\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & \cos\theta \\ 0 & \cos\theta & 1 \end{vmatrix} \\ &= \frac{1}{64}(1-\cos^2\theta) = \frac{1}{64}\sin^2\theta. \end{aligned} \tag{6.59}$$

The volume element, $\sqrt{g}\, d\theta d\phi d\psi$, is therefore

$$d(\text{Volume}) = \frac{1}{8}\sin\theta d\theta d\phi d\psi, \tag{6.60}$$

and the total volume of the sphere is

$$\text{Vol}(S^3) = \frac{1}{8}\int_0^{\pi} \sin\theta d\theta \int_0^{2\pi} d\phi \int_0^{4\pi} d\psi = 2\pi^2. \tag{6.61}$$

This coincides with the standard expression for the volume of S^{d-1}, the surface of the d-dimensional unit ball,

$$\text{Vol}(S^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}, \tag{6.62}$$

when $d = 4$.

Exercise 6.7: Evaluate the Maurer-Cartan form ω_L^3 in terms of the Euler angle parameterization and show that

$$i\omega_L^3 = \frac{1}{2}\text{tr}\left(\hat{\sigma}_3 U^{-1} dU\right) = -\frac{i}{2}(d\psi + \cos\theta\, d\phi).$$

Now recall that the Hopf map takes the point on the three-sphere with Euler angle co-ordinates (θ, ϕ, ψ) to the point on the two-sphere with spherical polar co-ordinates (θ, ϕ). Thus, if we set $A = -d\psi - \cos\theta\, d\phi$, then we find

$$F \equiv dA = \sin\theta\, d\theta\, d\phi = \text{Hopf}^*(d[\text{Area}\, S^2]).$$

Also observe that

$$A \wedge F = -\sin\theta \, d\theta \, d\phi \, d\psi.$$

From this show that Hopf index of the Hopf map itself is equal to

$$\frac{1}{16\pi^2} \int_{S^3} A \wedge F = -1.$$

Exercise 6.8: Show that for U the defining two-by-two matrices of SU(2), we have

$$\int_{\mathrm{SU}(2)} \mathrm{tr}\,[(U^{-1}dU)^3] = 24\pi^2.$$

Suppose we have a map $g : \mathbb{R}^3 \to \mathrm{SU}(2)$ such that $g(x)$ goes to the identity element at infinity. Consider the integral

$$S[g] = \frac{1}{24\pi^2} \int_{\mathbb{R}^3} \mathrm{tr}\,(g^{-1}dg)^3,$$

where the 3-form $\mathrm{tr}\,(g^{-1}dg)^3$ is the pull-back to $\mathbb{R}^3$ of the form $\mathrm{tr}\,[(U^{-1}dU)^3]$ on SU(2). Show that if we vary $g \to g + \delta g$, then

$$\delta S[g] = \frac{1}{24\pi^2} \int_{\mathbb{R}^3} d\left\{ 3\,\mathrm{tr}\left((g^{-1}\delta g)(g^{-1}dg)^2 \right) \right\} = 0,$$

and so $S[g]$ is topological invariant of the map g. Conclude that the functional $S[g]$ is an integer, that integer being the Brouwer degree, or winding number, of the map $g : S^3 \to S^3$.

Exercise 6.9: Generalize the result of the previous problem to show, for any mapping $x \mapsto g(x)$ into a Lie group G, and for n an odd integer, that the n-form $\mathrm{tr}\,(g^{-1}dg)^n$ constructed from the Maurer-Cartan form is closed, and that

$$\delta\,\mathrm{tr}\,(g^{-1}dg)^n = d\left\{ n\,\mathrm{tr}\left((g^{-1}\delta g)(g^{-1}dg)^{n-1} \right) \right\}.$$

(Note that for even n the trace of $(g^{-1}dg)^n$ vanishes identically.)

6.2.5 $\mathrm{SO}(3) \simeq \mathrm{SU}(2)/\mathbb{Z}_2$

The groups SU(2) and SO(3) are *locally isomorphic*. They have the same Lie algebra, but differ in their global topology. Although rotations in space are elements of SO(3), electrons respond to these rotations by transforming under the two-dimensional defining representation of SU(2). As we shall see,

this means that after a rotation through 2π the electron wavefunction comes back to minus itself. The resulting topological entanglement is characteristic of the *spinor* representation of rotations and is intimately connected with the Fermi statistics of the electron. The spin representations were discovered by Élie Cartan in 1913, long before they were needed in physics.

The simplest way to motivate the spin/rotation connection is via the Pauli sigma matrices. These matrices are hermitian, traceless, and obey

$$\hat{\sigma}_i\hat{\sigma}_j + \hat{\sigma}_j\hat{\sigma}_i = 2\delta_{ij}I, \tag{6.63}$$

If, for any $U \in \mathrm{SU}(2)$, we define

$$\hat{\sigma}'_i = U\hat{\sigma}_i U^{-1}, \tag{6.64}$$

then the $\hat{\sigma}'_i$ are also hermitian, traceless, and obey (6.63). Since the original $\hat{\sigma}_i$ form a basis for the space of hermitian traceless matrices, we must have

$$\hat{\sigma}'_i = \hat{\sigma}_j R_{ji} \tag{6.65}$$

for some real 3-by-3 matrix having entries R_{ij}. From (6.63) we find that

$$\begin{aligned} 2\delta_{ij} &= \hat{\sigma}'_i\hat{\sigma}'_j + \hat{\sigma}'_j\hat{\sigma}'_i \\ &= (\hat{\sigma}_l R_{li})(\hat{\sigma}_m R_{mj}) + (\hat{\sigma}_m R_{mj})(\hat{\sigma}_l R_{li}) \\ &= (\hat{\sigma}_l\hat{\sigma}_m + \hat{\sigma}_m\hat{\sigma}_l)R_{li}R_{mj} \\ &= 2\delta_{lm}R_{li}R_{mj}. \end{aligned}$$

Thus

$$R_{mi}R_{mk} = \delta_{ik}. \tag{6.66}$$

In other words, $R^T R = I$, and R is an element of $\mathrm{O}(3)$. Now the determinant of any orthogonal matrix is ± 1, but the manifold of $\mathrm{SU}(2)$ is a connected set and $R = I$ when $U = I$. Since a continuous map from a connected set to the integers must be a constant, we conclude that $\det R = 1$ for all U. The R matrices are therefore in $\mathrm{SO}(3)$.

We now exploit the principle of the sextant to show that the correspondance goes both ways, *i.e.* we can find a $U(R)$ for any element $R \in \mathrm{SO}(3)$.

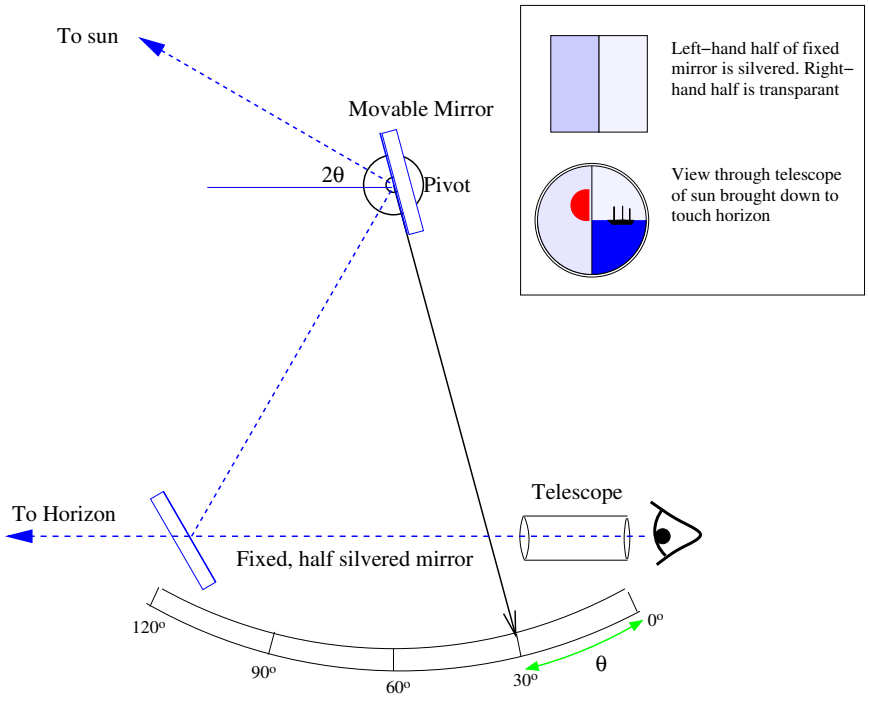

Figure 6.1: *The sextant.*

This familiar instrument is used to measure the altitude of the sun above the horizon while standing on the pitching deck of a ship at sea. A theodolite or similar device would be rendered useless by the ship's motion. The sextant exploits the fact that successive reflection in two mirrors inclined at an angle θ to one another serves to rotate the image through an angle 2θ about the line of intersection of the mirror planes. This rotation is used to superimpose the image of the sun onto the image of the horizon, where it stays even if the instrument is rocked back and forth. Exactly the same trick is used in constructing the spinor representations of the rotation group.

To do this, consider a vector $\mathbf{x}$ with components x^i and form the matrix $\widehat{\mathbf{x}} = x^i\hat{\sigma}_i$. Now, if $\mathbf{n}$ is a unit vector with components n^i, then

$$(-\hat{\sigma}_i n^i)\widehat{\mathbf{x}}(\hat{\sigma}_k n^k) = \left(x^j - 2(\mathbf{n}\cdot\mathbf{x})(n^j)\right)\hat{\sigma}_j = \widehat{\mathbf{x}} - 2(\mathbf{n}\cdot\mathbf{x})\widehat{\mathbf{n}} \tag{6.67}$$

The vector $\mathbf{x}-2(\mathbf{n}\cdot\mathbf{x})\mathbf{n}$ is the result of reflecting $\mathbf{x}$ in the plane perpendicular to $\mathbf{n}$. Consequently

$$-(\hat{\sigma}_1\cos\theta/2 + \hat{\sigma}_2\sin\theta/2)(-\hat{\sigma}_1)\,\widehat{\mathbf{x}}\,(\hat{\sigma}_1)(\hat{\sigma}_1\cos\theta/2 + \hat{\sigma}_2\sin\theta/2) \tag{6.68}$$

performs two successive reflections on $\mathbf{x}$, first in the "1" plane, and then in a plane at an angle $\theta/2$ to it. Multiplying out the factors, and using the $\hat{\sigma}_i$ algebra, we find

$$\begin{aligned}(\cos\theta/2 - \hat{\sigma}_1\hat{\sigma}_2 \sin\theta/2)\hat{\mathbf{x}}(\cos\theta/2 + \hat{\sigma}_1\hat{\sigma}_2 \sin\theta/2) \\ = \hat{\sigma}_1(\cos\theta\, x^1 - \sin\theta\, x^2) + \hat{\sigma}_2(\sin\theta\, x^1 + \cos\theta\, x^2) + \hat{\sigma}_3 x^3. \quad (6.69)\end{aligned}$$

The effect on $\mathbf{x}$ is a rotation through θ, as claimed. We can drop the x^i and re-express (6.69) as

$$U\hat{\sigma}_i U^{-1} = \hat{\sigma}_j R_{ji}, \tag{6.70}$$

where R_{ij} is the 3-by-3 rotation matrix for a rotation through angle θ in the 1-2 plane, and

$$U = \exp\left\{-\frac{i}{2}\hat{\sigma}_3\theta\right\} = \exp\left\{-i\frac{1}{4i}[\hat{\sigma}_1, \hat{\sigma}_2]\theta\right\} \tag{6.71}$$

is an element of SU(2). We have exhibited two ways of writing the exponents in (6.71) because the subscript 3 on $\hat{\sigma}_3$ indicates the axis about which we are rotating, while the $1, 2$ in $[\hat{\sigma}_1, \hat{\sigma}_2]$ indicates the plane in which the rotation occurs. It is the second language that generalizes to higher dimensions. More on the use of mirrors for creating and combining rotations can be found in the the appendix to Misner, Thorn, and Wheeler's *Gravitation*.

The mirror construction shows that for any $R \in \mathrm{SO}(3)$ there is a two-dimensional unitary matrix $U(R)$ such that

$$U(R)\hat{\sigma}_i U^{-1}(R) = \hat{\sigma}_j R_{ji}. \tag{6.72}$$

This $U(R)$ is not unique however. If $U \in \mathrm{SU}(2)$ then so is $-U$. Furthermore

$$U(R)\hat{\sigma}_i U^{-1}(R) = (-U(R))\hat{\sigma}_i(-U(R))^{-1}, \tag{6.73}$$

and so $U(R)$ and $-U(R)$ implement exactly the same rotation R. Conversely, if two SU(2) matrices U, V obey

$$U\sigma_i U^{-1} = V\sigma_i V^{-1} \tag{6.74}$$

then $V^{-1}U$ commutes with all 2-by-2 matrices and, by Schur's lemma, must be a multiple of the identity. But if $\lambda I \in \mathrm{SU}(2)$ then $\lambda = \pm 1$. Thus $U = \pm V$. The mapping between SU(2) and SO(3) is therefore two-to-one. Since U and

$-U$ correspond to the same R, the group manifold of SO(3) is the three-sphere *with antipodal points identified*. Unlike the two-sphere, where the identification of antipodal points gives the non-orientable projective plane, this three-manifold is is orientable. It is not, however, simply connected: a path on the three-sphere from a point to its antipode forms a closed loop in SO(3), but one not contractable to a point. If we continue on from the antipode back to the original point, the combined path *is* contractable. This means that the first *Homotopy group*, the group of based paths with composition given by concatenation, is $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$. This is the topology behind the Phillipine (or Balinese) Candle Dance, and is how the electron knows whether a sequence of rotations that eventually bring it back to its original orientation should be counted as a 360° rotation ($U = -I$) or a 720° $\sim 0°$ rotation ($U = +I$).

Exercise 6.10: Verify that

$$U(R)\hat{\sigma}_i U^{-1}(R) = \hat{\sigma}_j R_{ji}$$

is consistent with $U(R_2)U(R_1) = \pm U(R_2 R_1)$.

Spinor representations of SO(N)

The mirror trick can be extended to perform rotations in N dimensions. We replace the three $\hat{\sigma}_i$ matrices by a set of N *Dirac gamma matrices*, which obey the defining relations of a *Clifford algebra*

$$\hat{\gamma}_\mu \hat{\gamma}_\nu + \hat{\gamma}_\nu \hat{\gamma}_\mu = 2\delta_{\mu\nu} I. \tag{6.75}$$

These relations are a generalization of the key algebraic property of the Pauli sigma matrices.

If $N\,(=2n)$ is even, then we can find 2^n-by-2^n hermitian matrices, $\hat{\gamma}_\mu$, satisfying this algebra. If $N\,(=2n+1)$ is odd, we append to the matrices for $N = 2n$ the hermitian matrix $\hat{\gamma}_{2n+1} = -(i)^n \hat{\gamma}_1 \hat{\gamma}_2 \cdots \hat{\gamma}_{2n}$ which obeys $\hat{\gamma}_{2n+1}^2 = 1$ and anti-commutes with all the other $\hat{\gamma}_\mu$. The $\hat{\gamma}$ matrices therefore act on a $2^{[N/2]}$ dimensional space, where the square brackets denote the *integer part* of $N/2$.

The $\hat{\gamma}$'s do not form a Lie algebra as they stand, but a rotation through θ in the mn-plane is obtained from

$$e^{-i\frac{1}{4i}[\hat{\gamma}_m,\hat{\gamma}_n]\theta}\,\hat{\gamma}_i\,e^{i\frac{1}{4i}[\hat{\gamma}_m,\hat{\gamma}_n]\theta} = \hat{\gamma}_j R_{ji}, \tag{6.76}$$

and we find that the hermitian matrices $\hat{\Gamma}_{mn} = \frac{1}{4i}[\hat{\gamma}_m, \hat{\gamma}_n]$ form a basis for the Lie algebra of SO(N). The $2^{[N/2]}$ dimensional space on which they act is the Dirac spinor representation of SO(N). Although the matrices $\exp\{i\hat{\Gamma}_{\mu\nu}\theta_{\mu\nu}\}$ are unitary, they are not, in general, the entirety of U($2^{[N/2]}$), but instead constitute a subgroup called Spin(N).

If N is even then we can still construct the matrix $\hat{\gamma}_{2n+1}$ that anti-commutes with all the other $\hat{\gamma}_\mu$'s. It cannot be the identity matrix, therefore, but it commutes with all the Γ_{mn}. By Schur's lemma, this means that the SO($2n$) Dirac spinor representation space V is *reducible*. Now $\hat{\gamma}_{2n+1}^2 = I$, and so $\hat{\gamma}_{2n+1}$ has eigenvalues ± 1. The two eigenspaces are invariant under the action of the group, and thus the Dirac spinor space decomposes into two irreducible *Weyl spinor* representations

$$V = V_{\text{odd}} \oplus V_{\text{even}}. \tag{6.77}$$

Here V_{even} and V_{odd}, the plus and minus eigenspaces of $\hat{\gamma}_{2n+1}$, are called the spaces of right and left *chirality*. When N is odd the spinor representation is irreducible.

Exercise 6.11: Starting from the defining relations of the Clifford algebra (6.75) show that, for $N = 2n$,

$$\begin{aligned}
\text{tr}\,(\hat{\gamma}_\mu) &= 0, \\
\text{tr}\,(\hat{\gamma}_{2n+1}) &= 0, \\
\text{tr}\,(\hat{\gamma}_\mu\hat{\gamma}_\nu) &= \text{tr}\,(I)\,\delta_{\mu\nu}, \\
\text{tr}\,(\hat{\gamma}_\mu\hat{\gamma}_\nu\hat{\gamma}_\sigma) &= 0, \\
\text{tr}\,(\hat{\gamma}_\mu\hat{\gamma}_\nu\hat{\gamma}_\sigma\hat{\gamma}_\tau) &= \text{tr}\,(I)\,(\delta_{\mu\nu}\delta_{\sigma\tau} - \delta_{\mu\sigma}\delta_{\nu\tau} + \delta_{\mu\tau}\delta_{\nu\sigma}).
\end{aligned}$$

Exercise 6.12: Consider the space $\Omega(\mathbb{C}) = \bigoplus_p \Omega^p(\mathbb{C})$ of complex-valued skew symmetric tensors $A_{\mu_1\ldots\mu_p}$ for $0 \le p \le N = 2n$. Let

$$\psi_{\alpha\beta} = \sum_{p=0}^{N} \frac{1}{p!} \left(\hat{\gamma}_{\mu_1} \cdots \hat{\gamma}_{\mu_p}\right)_{\alpha\beta} A_{\mu_1\ldots\mu_p}$$

define a mapping from $\Omega(\mathbb{C})$ into the space of complex matrices of the same size as the $\hat{\gamma}_\mu$. Show that this mapping is invertible — *i.e.* given $\psi_{\alpha\beta}$ you can recover the $A_{\mu_1\ldots\mu_p}$. By showing that the dimension of $\Omega(\mathbb{C})$ is 2^N, deduce that the $\hat{\gamma}_\mu$ must be at least 2^n-by-2^n matrices.

Exercise 6.13: Show that the $\mathbb{R}^{2n}$ *Dirac operator* $D = \hat{\gamma}_\mu \partial_\mu$ obeys $D^2 = \nabla^2$. Recall that Hodge operator $d - \delta$ from section 4.7.1 is also a "square root" of the Laplacian:

$$(d - \delta)^2 = -(d\delta + \delta d) = \nabla^2.$$

Show that

$$\psi_{\alpha\beta} \to (D\psi)_{\alpha\beta} = (\hat{\gamma}_\mu)_{\alpha\alpha'} \partial_\mu \psi_{\alpha'\beta}$$

corresponds to the action of $d - \delta$ on the space $\Omega(\mathbb{R}^{2n}, \mathbb{C})$ of differential forms

$$A = \frac{1}{p!} A_{\mu_1 \ldots \mu_p}(x) dx^{\mu_1} \cdots dx^{\mu_p}$$

The space of complex-valued differential forms has thus been made to look like a collection of 2^n Dirac spinor fields, one for each value of the "flavour index" β. These $\psi_{\alpha\beta}$ are called *Kähler-Dirac* fields. They are not really flavoured spinors because a rotation transforms both the α and β indices.

Exercise 6.14: That a set of $2n$ Dirac γ's have a 2^n-by-2^n matrix representation is most naturally established by using the tools of second quantization. To this end, let a_i, $a_i^\dagger$ $i = 1, \ldots, n$ be set of anti-commuting annihilation and creation operators obeying

$$a_i a_j + a_j a_i = 0, \quad a_i a_j^\dagger + a_j^\dagger a_i = \delta_{ij} I,$$

and let $|0\rangle$ be the "no particle" state such that $a_i|0\rangle = 0$, $i = 1, \ldots, n$. Then the 2^n states

$$|m_1, \ldots, m_n\rangle = (a_1^\dagger)^{m_1} \cdots (a_n^\dagger)^{m_n} |0\rangle,$$

where the m_i take the value 0 or 1, constitute a basis for a space on which the a_i and $a_i^\dagger$ act irreducibly. Show that the $2n$ operators

$$\begin{aligned} \gamma_i &= a_i + a_i^\dagger \\ \gamma_{i+n} &= i(a_i - a_i^\dagger) \end{aligned}$$

obey

$$\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\delta_{\mu\nu} I,$$

and hence can be represented by 2^n-by-2^n matrices. Deduce further that spaces specs of left and right chirality are the spaces of odd or even "particle number."

The Adjoint Representation

The spin/rotation correspondence involves conjugation: $\hat{\sigma}_i \to U\hat{\sigma}_i U^{-1}$. The idea of obtaining a representation by conjugation works for an arbitrary Lie group. It is easiest, however, to describe in the case of a matrix group where we consider an infinitesimal element $I + i\epsilon\hat{\lambda}_i$. The conjugate element $g(I + i\epsilon\hat{\lambda}_i)g^{-1}$ will also be an infinitesimal element. Since $gIg^{-1} = I$, this means that $g(i\hat{\lambda}_i)g^{-1}$ must be expressible as a linear combination of the $i\hat{\lambda}_i$ matrices. Consequently we can define a linear map acting on the element $X = \xi^i\hat{\lambda}_i$ of the Lie algebra by setting

$$\mathrm{Ad}(g)\hat{\lambda}_i \equiv g\hat{\lambda}_i g^{-1} = \hat{\lambda}_j[\mathrm{Ad}\,(g)]^j{}_i. \tag{6.78}$$

The matrices with entries $[\mathrm{Ad}\,(g)]^j{}_i$ form the *adjoint* representation of the group. The dimension of the adjoint representation coincides with that of the group manifold. The spinor construction shows that the defining representation of SO(3) is the adjoint representation of SU(2).

For a general Lie group, we make $\mathrm{Ad}(g)$ act on a vector in the tangent space at the identity by pushing the vector forward to TG_g by left multiplication by g, and then pushing it back from TG_g to TG_e by right multiplication by g^{-1}.

Exercise 6.15: Show that

$$[\mathrm{Ad}\,(g_1g_2)]^j{}_i = [\mathrm{Ad}\,(g_1)]^j{}_k[\mathrm{Ad}\,(g_2)]^k{}_i,$$

thus confirming that $\mathrm{Ad}(g)$ is a representation.

6.2.6 Peter-Weyl Theorem

The volume element constructed in section 6.2.4 has the feature that it is *invariant*. In other words if we have a subset Ω of the group manifold with volume V, then the image set $g\Omega$ under left multiplication has the exactly the same volume. We can also construct a volume element that is invariant under right multiplication by g, and in general these will be different. For a group whose manifold is a compact set, however, both left- and right-invariant volume elements coincide. The resulting measure on the group manifold is called the *Haar* measure.

For a *compact* group, therefore, we can replace the sums over the group elements that occur in the representation theory of finite groups, by convergent integrals over the group elements using the invariant Haar measure,

which is usually denoted by $d[g]$. The invariance property is expressed by $d[g_1 g] = d[g]$ for any constant element g_1. This allows us to make a change-of-variables transformation, $g \to g_1 g$, identical to that which played such an important role in deriving the finite group theorems. Consequently, all the results from finite groups, such as the existence of an invariant inner product and the orthogonality theorems, can be taken over by the simple replacement of a sum by an integral. In particular, if we normalize the measure so that the volume of the group manifold is unity, we have the orthogonality relation

$$\int d[g] \left(D^J_{ij}(g)\right)^* D^K_{lm}(g) = \frac{1}{\dim J}\delta^{JK}\delta_{il}\delta_{jm}. \tag{6.79}$$

The Peter-Weyl theorem asserts that the representation matrices, $D^J_{mn}(g)$, form a complete set of orthogonal functions on the group manifold. In the case of SU(2) this tells us that the spin J representation matrices

$$\begin{aligned} D^J_{mn}(\theta,\phi,\psi) &= \langle J,m|e^{-iJ_3\phi}e^{-iJ_2\theta}e^{-iJ_3\psi}|J,n\rangle, \\ &= e^{-im\phi}d^J_{mn}(\theta)e^{-in\psi}, \end{aligned} \tag{6.80}$$

which you will know from quantum mechanics courses,[1] are a complete set of functions on the three-sphere with

$$\begin{aligned} &\frac{1}{16\pi^2}\int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \int_0^{4\pi} d\psi \left(D^J_{mn}(\theta,\phi,\psi)\right)^* D^{J'}_{m'n'}(\theta,\phi,\psi) \\ &= \frac{1}{2J+1}\delta^{JJ'}\delta_{mm'}\delta_{nn'}. \end{aligned} \tag{6.81}$$

Since the D^L_{m0} (where L has to be an integer for $n = 0$ to be possible) are independent of the third Euler angle, ψ, we can do the trivial integral over ψ to get

$$\frac{1}{4\pi}\int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\phi \left(D^L_{m0}(\theta,\phi)\right)^* D^{L'}_{m'0}(\theta,\phi) = \frac{1}{2L+1}\delta^{LL'}\delta_{mm'}. \tag{6.82}$$

Comparing with the definition of the spherical harmonics, we see that we can identify

$$Y^L_m(\theta,\phi) = \sqrt{\frac{2L+1}{4\pi}}\left(D^L_{m0}(\theta,\phi,\psi)\right)^*. \tag{6.83}$$

[1] See, for example, G. Baym *Lectures on Quantum Mechanics*, Ch 17.

The complex conjugation is necessary here because $D^J_{mn}(\theta,\phi,\psi) \propto e^{-im\phi}$, while $Y^L_m(\theta,\phi) \propto e^{im\phi}$.

The character, $\chi^J(g) = \sum_n D^J_{nn}(g)$ will be a function only of the angle θ we have rotated through, not the axis of rotation — all rotations through a common angle being conjugate to one another. Because of this $\chi^J(\theta)$ can be found most simply by looking at rotations about the z axis, since these give rise to easily computed diagonal matrices. We find

$$\begin{aligned}\chi(\theta) &= e^{iJ\theta} + e^{i(J-1)\theta} + \cdots + e^{-i(J-1)\theta} + e^{-iJ\theta}, \\ &= \frac{\sin(2J+1)\theta/2}{\sin\theta/2}.\end{aligned} \tag{6.84}$$

Warning: The angle θ in this formula and the next is the not the Euler angle.

For integer J, corresponding to non-spinor rotations, a rotation through an angle θ about an axis $\mathbf{n}$ and a rotation though an angle $2\pi - \theta$ about $-\mathbf{n}$ are the same operation. The maximum rotation angle is therefore π. For spinor rotations this equivalence does not hold, and the rotation angle θ runs from 0 to 2π. The character orthogonality must therefore be

$$\frac{1}{\pi}\int_0^{2\pi} \chi^J(\theta)\chi^{J'}(\theta)\sin^2\left(\frac{\theta}{2}\right)d\theta = \delta^{JJ'}, \tag{6.85}$$

implying that the volume fraction of the rotation group containing rotations through angles between θ and $\theta + d\theta$ is $\sin^2(\theta/2)d\theta/\pi$.

Exercise 6.16: Prove this last statement about the volume of the equivalence classes by showing that the volume of the unit three-sphere that lies between a rotation angle of θ and $\theta + d\theta$ is $2\pi\sin^2(\theta/2)d\theta$.

6.2.7 Lie Brackets *vs.* Commutators

There is an irritating minus sign problem that needs to be acknowledged. The Lie bracket $[X, Y]$ of of two vector fields is defined by first running along X, then Y and then back in the reverse order. If we do this for the action of matrices, $\hat{X}$ and $\hat{Y}$, on a vector space, however, then, reading from right to left as we always do for matrix operations, we have

$$e^{-t_2\hat{Y}}e^{-t_1\hat{X}}e^{t_2\hat{Y}}e^{t_1\hat{X}} = I - t_1t_2[\hat{X},\hat{Y}] + \cdots, \tag{6.86}$$

which has the other sign. Consider for example rotations about the x, y, z axes, and look at effect these have on the co-ordinates of a point:

$$L_x: \quad \left\{ \begin{array}{lcl} \delta y &=& -z\,\delta\theta_x \\ \delta z &=& +y\,\delta\theta_x \end{array} \right\} \Longrightarrow L_x = y\partial_z - z\partial_y, \quad \hat{L}_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$L_y: \quad \left\{ \begin{array}{lcl} \delta z &=& -x\,\delta\theta_y \\ \delta x &=& +z\,\delta\theta_y \end{array} \right\} \Longrightarrow L_y = z\partial_x - x\partial_z, \quad \hat{L}_y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

$$L_z: \quad \left\{ \begin{array}{lcl} \delta x &=& -y\,\delta\theta_z \\ \delta y &=& +x\,\delta\theta_z \end{array} \right\} \Longrightarrow L_z = x\partial_y - y\partial_x, \quad \hat{L}_y = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

From this we find

$$[L_x, L_y] = -L_z, \tag{6.87}$$

as a Lie bracket of vector fields, but

$$[\hat{L}_x, \hat{L}_y] = +\hat{L}_z, \tag{6.88}$$

as a commutator of matrices. This is the reason why it is the *left* invariant vector fields whose Lie bracket coincides with the commutator of the $i\hat{\lambda}_i$ matrices.

Some insight into all this can be had by considering the action of the left invariant fields on the representation matrices, $D^J_{mn}(g)$. For example

$$\begin{aligned} L_i D^J_{mn}(g) &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn}(g(1+i\epsilon\hat{\lambda}_i)) - D^J_{mn}(g)\right)\right] \\ &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn'}(g)D^J_{n'n}(1+i\epsilon\hat{\lambda}_i) - D^J_{mn}(g)\right)\right] \\ &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn'}(g)(\delta_{n'n} + i\epsilon(\hat{\Lambda}^J_i)_{n'n}) - D^J_{mn}(g)\right)\right] \\ &= D^J_{mn'}(g)(i\hat{\Lambda}^J_i)_{n'n} \end{aligned} \tag{6.89}$$

where $\hat{\Lambda}^J_i$ is the matrix representing $\hat{\lambda}_i$ in the representation J. Repeating this exercise we find that

$$L_i\left(L_j D^J_{mn}(g)\right) = D^J_{mn''}(g)(i\hat{\Lambda}^J_i)_{n''n'}(i\hat{\Lambda}^J_j)_{n'n}, \tag{6.90}$$

Thus

$$[L_i, L_j]D^J_{mn}(g) = D^J_{mn'}(g)[i\hat{\Lambda}^J_i, i\hat{\Lambda}^J_j]_{n'n}, \tag{6.91}$$

and we get the commutator of the representation matrices in the "correct" order only if we multiply the infinitesimal elements in successively from the right.

There appears to be no escape from this sign problem. Many texts simply ignore it, a few define the Lie bracket of vector fields with the opposite sign, and a few simply point out the inconvenience and get on the with the job. We will follow the last route.

6.3 Lie Algebras

A Lie algebra $\mathfrak{g}$ is a (real or complex) finite-dimensional vector space with a non-associative binary operation $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ that assigns to each ordered pair of elements, X_1, X_2, a third element called the Lie bracket, $[X_1, X_2]$. The bracket is:

a) Skew symmetric: $[X, Y] = -[Y, X]$,
b) Linear: $[\lambda X + \mu Y, Z] = \lambda[X, Z] + \mu[Y, Z]$,

and in place of associativity, obeys

c) The Jacobi identity: $[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$.

Example: Let $M(n)$ denote the algebra of real n-by-n matrices. As a vector space over $\mathbb{R}$, this algebra is n^2 dimensional. Setting $[A, B] = AB - BA$, makes $M(n)$ into a Lie Algebra.

Example: Let $\mathfrak{b}^+$ denote the subset of $M(n)$ consisting of upper triangular matrices with any number (including zero) allowed on the diagonal. Then $\mathfrak{b}^+$ with the above bracket is a Lie algebra. (The "b" stands for the French mathematician and statesman Émile *Borel*).

Example: Let $\mathfrak{n}^+$ denote the subset of $\mathfrak{b}^+$ consisting of strictly upper triangular matrices — those with zero on the diagonal. Then $\mathfrak{n}^+$ with the above bracket is a Lie algebra. (The "n" stands for *nilpotent.*)

Example: Let G be a Lie group, and L_i the left invariant vector fields. We know that

$$[L_i, L_j] = f_{ij}{}^k L_k \tag{6.92}$$

where $[\ \ ,\ \ ]$ is the Lie bracket of vector fields. The resulting Lie algebra, $\mathfrak{g} = \mathrm{Lie}\, G$ is the Lie algebra of the group.

Example: The set N^+ of upper triangular matrices with 1's on the diagonal forms a Lie group and has $\mathfrak{n}^+$ as its Lie algebra. Similarly, the set B^+ consisting of upper triangular matrices, with any non-zero number allowed on the diagonal, is also a Lie group, and has $\mathfrak{b}^+$ as its Lie algebra.

Ideals and Quotient algebras

As we saw in the examples, we can define subalgebras of a Lie algebra. If we want to define quotient algebras by analogy to quotient groups, we need a concept analogous to that of invariant subgroups. This is provided by the notion of an *ideal*. A ideal is a subalgebra $\mathfrak{i} \subseteq \mathfrak{g}$ with the property that

$$[\mathfrak{i}, \mathfrak{g}] \subseteq \mathfrak{i}. \tag{6.93}$$

In other words, taking the bracket of any element of $\mathfrak{g}$ with any element of $\mathfrak{i}$ gives an element in $\mathfrak{i}$. With this definition we can form $\mathfrak{g} - \mathfrak{i}$ by identifying $X \sim X + I$ for any $I \in \mathfrak{i}$. Then

$$[X + \mathfrak{i}, Y + \mathfrak{i}] = [X, Y] + \mathfrak{i}, \tag{6.94}$$

and the bracket of two equivalence classes is insensitive to the choice of representatives.

If a Lie group G has an invariant subgroup H which is also a Lie group, then the Lie algebra $\mathfrak{h}$ of the subgroup is an ideal in $\mathfrak{g} = \mathrm{Lie}\, G$ and the Lie algebra of the quotient group G/H is the quotient algebra $\mathfrak{g} - \mathfrak{h}$.

If the Lie algebra has no non-trivial ideals, then it is said to be *simple*. The Lie algebra of a simple Lie group will be simple.

Exercise 6.17: Let $\mathfrak{i}_1$ and $\mathfrak{i}_2$ be ideals in $\mathfrak{g}$. Show that $\mathfrak{i}_1 \cap \mathfrak{i}_2$ is also an ideal in $\mathfrak{g}$.

6.3.1 Adjoint Representation

Given an element $X \in \mathfrak{g}$ let it act on the Lie algebra considered as a vector space by a linear map $\mathrm{ad}\,(x)$ defined by

$$\mathrm{ad}\,(X)Y = [X, Y]. \tag{6.95}$$

The Jacobi identity is then equivalent to the statement

$$(\mathrm{ad}\,(X)\mathrm{ad}\,(Y) - \mathrm{ad}\,(Y)\mathrm{ad}\,(X))\, Z = \mathrm{ad}\,([X, Y])Z. \tag{6.96}$$

Thus

$$(\mathrm{ad}\,(X)\mathrm{ad}\,(Y) - \mathrm{ad}\,(Y)\mathrm{ad}\,(X)) = \mathrm{ad}\,([X,Y]), \tag{6.97}$$

or

$$[\mathrm{ad}\,(X), \mathrm{ad}\,(Y)] = \mathrm{ad}\,([X,Y]), \tag{6.98}$$

and the map $X \to \mathrm{ad}\,(X)$ is a representation of the algebra called the *adjoint representation*.

The linear map "ad (X)" exponentiates to give a map $\exp[\mathrm{ad}\,(tX)]$ defined by

$$\exp[\mathrm{ad}\,(tX)]Y = Y + t[X,Y] + \frac{1}{2}t^2[X,[X,Y]] + \cdots. \tag{6.99}$$

You probably know the matrix identity[2]

$$e^{tA}Be^{-tA} = B + t[A,B] + \frac{1}{2}t^2[A,[A,B]] + \cdots. \tag{6.100}$$

Now, earlier in the chapter, we defined the adjoint representation "Ad" of the *group* on the vector space of the Lie algebra. We did this setting $gXg^{-1} = \mathrm{Ad}\,(g)X$. Comparing the two previous equations we see that

$$\mathrm{Ad}\,(\mathrm{Exp}\,Y) = \exp(\mathrm{ad}\,(Y)). \tag{6.101}$$

6.3.2 The Killing form

Using "ad" we can define an inner product $\langle \quad , \quad \rangle$ on a real Lie algebra by setting

$$\langle X, Y\rangle = \mathrm{tr}\,(\mathrm{ad}\,(X)\mathrm{ad}\,(Y)). \tag{6.102}$$

This inner product is called the *Killing form*, after Wilhelm Killing. Using the Jacobi identity, and the cyclic property of the trace, we find that

$$\langle \mathrm{ad}\,(X)Y, Z\rangle + \langle Y, \mathrm{ad}\,(X)Z\rangle = 0, \tag{6.103}$$

or, equivalently,

$$\langle [X,Y], Z\rangle + \langle Y, [X,Z]\rangle = 0. \tag{6.104}$$

From this we deduce (by differentiating with respect to t) that

$$\langle \exp(\mathrm{ad}\,(tX))Y, \exp(\mathrm{ad}\,(tX))Z\rangle = \langle Y, Z\rangle, \tag{6.105}$$

[2]In case you do not, it is easily proved by setting $F(t) = e^{tA}Be^{-tA}$, noting that $\frac{d}{dt}F(t) = [A, F(t)]$, and observing that the RHS is the unique series solution to this equation satisfying the boundary condition $F(0) = B$.

so the Killing form is invariant under the action of the adjoint representation of the *group* on the algebra. When our group is simple, any other invariant inner product will be proportional to this Killing-form product.

Exercise 6.18: Let $\mathfrak{i}$ be an ideal in $\mathfrak{g}$. Show that for $I_1, I_2 \in \mathfrak{i}$

$$\langle I_1, I_2 \rangle_{\mathfrak{g}} = \langle I_1, I_2 \rangle_{\mathfrak{i}}$$

where $\langle \quad , \quad \rangle_{\mathfrak{i}}$ is the Killing form on $\mathfrak{i}$ considered as a Lie algebra in its own right. (This equality of inner products is not true for subalgebras that are not ideals.)

Semi-simplicity

Recall that a Lie algebra containing no non-trivial ideals is said to be *simple*. When the Killing form is non degenerate, the Lie Algebra is said to be *semi-simple*. The reason for this name is that a semi-simple algebra is *almost* simple, in that it can be decomposed into a direct sum of decoupled simple algebras

$$\mathfrak{g} = \mathfrak{s}_1 \oplus \mathfrak{s}_2 \oplus \cdots \oplus \mathfrak{s}_n. \tag{6.106}$$

Here the direct sum symbol "$\oplus$" implies not only a direct sum of vector spaces but also that $[\mathfrak{s}_i, \mathfrak{s}_j] = 0$ for $i \neq j$.

The Lie algebra of all the matrix groups $\mathrm{O}(n)$, $\mathrm{Sp}(n)$, $\mathrm{SU}(n)$, *etc.* are semi-simple (indeed they are usually simple) but this is not true of the algebras $\mathfrak{n}^+$ and $\mathfrak{b}^+$.

Cartan showed that our Killing-form definition of semi-simplicity is equivalent his original definition of a Lie algebra being semi-simple if it contains no *abelian* ideal — *i.e.* no ideal with $[I_i, I_j] = 0$ for all $I_i \in \mathfrak{i}$. The following exercises establish the direct sum decomposition, and, *en passant*, the easy half of Cartan's result.

Exercise 6.19: Use the identity (6.104) to show that if $\mathfrak{i} \subset \mathfrak{g}$ is an ideal, then $\mathfrak{i}^\perp$, the set of elements orthogonal to $\mathfrak{i}$ with respect to the Killing form, is also an ideal.

Exercise 6.20: Show that if $\mathfrak{a}$ is an abelian ideal, then every element of $\mathfrak{a}$ is Killing perpendicular to the entire Lie algebra. (Thus non-degeneracy $\Rightarrow$ no non-trivial abelian ideal. The null space of the Killing form is not necessarily an abelian ideal, though, so establishing the converse is harder.)

Exercise 6.21: Let $\mathfrak{g}$ be semi-simple and $\mathfrak{i} \subset \mathfrak{g}$ an ideal. We know from exercise 6.17 that $\mathfrak{i} \cap \mathfrak{i}^\perp$ is an ideal. Use (6.104) coupled with the non-degeneracy of the Killing form to show that it is an *abelian* ideal. Use the previous exercise to conclude that $\mathfrak{i} \cap \mathfrak{i}^\perp = \{0\}$, and from this that $[\mathfrak{i}, \mathfrak{i}^\perp] = 0$.

Exercise 6.22: Let $\langle\ ,\ \rangle$ be a non-degenerate inner product on a vector space V. Let $W \subseteq V$ be a subspace. Show that

$$\dim W + \dim W^\perp = \dim V.$$

(This is not as obvious as it looks. For a non-positive-definite inner product W and $W^\perp$ can have a non-trivial intersection. Consider two-dimensional Minkowski space. If W is the space of right-going, light-like, vectors then $W \equiv W^\perp$, but $\dim W + \dim W^\perp$ still equals two.)

Exercise 6.23: Put the two preceding exercises together to show that

$$\mathfrak{g} = \mathfrak{i} \oplus \mathfrak{i}^\perp.$$

Show that $\mathfrak{i}$ and $\mathfrak{i}^\perp$ are semi-simple in their own right as Lie algebras. We can therefore continue to break up $\mathfrak{i}$ and $\mathfrak{i}^\perp$ until we end with $\mathfrak{g}$ decomposed into a direct sum of simple algebras.

Compactness

If the Killing form is negative definite, a real Lie Algebra is said to be *compact*, and is the Lie algebra of a compact group. With the physicist's habit of writing iX_i for the generators of the Lie algebra, a compact group has Killing metric tensor

$$g_{ij} = \mathrm{tr}\,\{\mathrm{ad}\,(X_i)\mathrm{ad}\,(X_j)\} \tag{6.107}$$

that is a *positive definite* matrix. In a basis where $g_{ij} = \delta_{ij}$, the $\exp(\mathrm{ad}\,X)$ matrices of the adjoint representations of a compact group G form a subgroup of the orthogonal group $\mathrm{O}(N)$, where N is the dimension of G.

Totally anti-symmetric structure constants

Given a basis iX_i for the Lie-algebra vector space, we define the structure constants ${f_{ij}}^k$ by

$$[X_i, X_j] = i{f_{ij}}^k X_k. \tag{6.108}$$

In terms of the ${f_{ij}}^k$, the skew symmetry of $\mathrm{ad}\,(X_i)$, as expressed by equation (6.103), becomes

$$\begin{aligned} 0 &= \langle \mathrm{ad}\,(X_k)X_i, X_j\rangle + \langle X_i, \mathrm{ad}\,(X_k)X_j\rangle \\ &\equiv \langle [X_k, X_i], X_j\rangle + \langle X_i, [X_k, X_j]\rangle \\ &= i({f_{ki}}^l g_{lj} + g_{il}{f_{kj}}^l) \\ &= i(f_{kij} + f_{kji}). \end{aligned} \tag{6.109}$$

In the last line we have used the Killing metric to "lower" the index l and so define the symbol f_{ijk}. Thus f_{ijk} is skew symmetric under the interchange of its second pair of indices. Since the skew symmetry of the Lie bracket ensures that f_{ijk} is skew symmetric under the interchange of the first pair of indices, it follows that f_{ijk} is skew symmetric under the interchange of *any* pair of its indices.

By comparing the definition of the structure constants with

$$[X_i, X_j] = \mathrm{ad}\,(X_i)X_j = X_k[\mathrm{ad}\,(X_i)]^k{}_j, \tag{6.110}$$

we read-off that the matrix representing $\mathrm{ad}\,(X_i)$ has entries

$$[(\mathrm{ad}\,(X_i)]^k{}_j = i{f_{ij}}^k. \tag{6.111}$$

Consequently

$$g_{ij} = \mathrm{tr}\,\{\mathrm{ad}\,(X_i)\mathrm{ad}\,(X_j)\} = -{f_{ik}}^l{f_{jl}}^k. \tag{6.112}$$

The quadratic Casimir

The only "product" that is defined in the abstract Lie algebra $\mathfrak{g}$ is the Lie bracket $[X, Y]$. Once we have found matrices forming a representation of the Lie algebra, however, we can form the ordinary matrix product of these. Suppose that we have a Lie algebra $\mathfrak{g}$ with basis X_i and have found matrices $\hat{X}_i$ with the same commutation relations as the X_i. Suppose further that the algebra is semisimple and so g^{ij}, the inverse of the Killing metric, exists. We can use g^{ij} to construct the matrix

$$\hat{C}_2 = g^{ij}\hat{X}_i\hat{X}_j. \tag{6.113}$$

This matrix is called the *quadratic Casimir* operator, after Hendrik Casimir. Its chief property is that it commutes with all the $\hat{X}_i$:

$$[\hat{C}_2, \hat{X}_i] = 0. \tag{6.114}$$

If our representation is irreducible then Shur's lemma tells us that

$$\hat{C}_2 = c_2 I \tag{6.115}$$

where the number c_2 is referred to as the "value" of the quadratic Casimir in that irrep.[3]

Exercise 6.24: Show that $[\hat{C}_2, X_i] = 0$ is another consequence of the complete skew symmetry of the f_{ijk}.

6.3.3 Roots and Weights

We now want to study the representation theory of Lie groups. It is, in fact, easier to study the representations of the Lie algebra, and then exponentiate these to find the representations of the group. In other words given an abstract Lie algebra with bracket

$$[X_i, X_j] = i f_{ij}{}^k X_k, \tag{6.116}$$

we seek to find all matrices $\hat{X}_i^J$ such that

$$[\hat{X}_i^J, \hat{X}_j^J] = i f_{ij}{}^k \hat{X}_k^J. \tag{6.117}$$

(Here, as with the representations of finite groups, we use the superscript J to distinguish one representation from another.) Then, given a representation $\hat{X}_i^J$ of the Lie algebra, the matrices

$$D^J(g(\xi)) = \exp\left\{ i\xi^i \hat{X}_i^J \right\}, \tag{6.118}$$

where $g(\xi) = \mathrm{Exp}\left\{ i\xi^i X_i \right\}$, will form a representation of the Lie *group*. To be more precise, they will form a representation of that part of the group which is connected to the identity element. The numbers ξ^i will serve as co-ordinates for some neighbourhood of the identity. For compact groups there will be a restriction on the range of the ξ^i because there must be ξ^i for which $\exp\left\{ i\xi^i \hat{X}_i^J \right\} = I$.

[3] Mathematicians do sometimes consider formal products of Lie algebra elements $X, Y \in \mathfrak{g}$. When they do, they equip them with the rule that $XY - YX - [X, Y] = 0$, where XY and YX are formal products, and $[X, Y]$ is the Lie algebra product. These formal products are not elements of the Lie algebra, but instead live in an extended mathematical structure called the *Universal enveloping algebra* of $\mathfrak{g}$, and denoted by $U(\mathfrak{g})$. The quadratic Casimir can then be considered to be an element of this larger algebra.

SU(2)

The quantum-mechanical angular momentum algebra consists of the commutation relation

$$[J_1, J_2] = i\hbar J_3, \tag{6.119}$$

together with two similar equations related by cyclic permutations. This, once we set $\hbar = 1$, is the Lie algebra $\mathfrak{su}(2)$ of the group SU(2). The goal of representation theory is to find all possible sets of matrices which have the same commutation relations as these operators. Since the group SU(2) is compact, we can use the group-averaging trick from section 5.2.2 to define an inner product with respect to which these representations are unitary, and the matrices J_i hermitian.

Remember how this problem is solved in quantum mechanics courses, where we find a representation for each spin $j = \frac{1}{2}, 1, \frac{3}{2}$, *etc.* We begin by constructing "ladder" operators

$$J_+ = J_1 + iJ_2, \quad J_- = J_+^\dagger = J_1 - iJ_2, \tag{6.120}$$

which are eigenvectors of $\mathrm{ad}\,(J_3)$

$$\mathrm{ad}\,(J_3)J_\pm = [J_3, J_\pm] = \pm J_\pm. \tag{6.121}$$

From (6.121) we see that if $|j, m\rangle$ is an eigenstate of J_3 with eigenvalue m, then $J_\pm|j, m\rangle$ is an eigenstate of J_3 with eigenvalue $m \pm 1$.

Now in any finite-dimensional representation there must be a *highest weight* state, $|j, j\rangle$, such that $J_3|j, j\rangle = j|j, j\rangle$ for some real number j, and such that $J_+|j, j\rangle = 0$. From $|j, j\rangle$ we work down by successive applications of J_- to find $|j, j-1\rangle$, $|j, j-2\rangle$... We can find the normalization factors of the states $|j, m\rangle \propto (J_-)^{j-m}|j, j\rangle$ by repeated use of the identities

$$\begin{aligned} J_+J_- &= (J_1^2 + J_2^2 + J_3^2) - (J_3^2 - J_3), \\ J_-J_+ &= (J_1^2 + J_2^2 + J_3^2) - (J_3^2 + J_3). \end{aligned} \tag{6.122}$$

The combination $J^2 \equiv J_1^2 + J_2^2 + J_3^2$ is the quadratic Casimir of $\mathfrak{su}(2)$, and hence in any irrep is proportional to the identity matrix: $J^2 = c_2 I$. Because

$$\begin{aligned} 0 &= \|J_+|j, j\rangle\|^2 \\ &= \langle j, j|J_+^\dagger J_+|j, j\rangle \\ &= \langle j, j|J_-J_+|j, j\rangle \\ &= \langle j, j|\left(J^2 - J_3(J_3 + 1)\right)|j, j\rangle \\ &= [c_2 - j(j+1)]\langle j, j|j, j\rangle, \end{aligned} \tag{6.123}$$

and $\langle j,j|j,j\rangle \equiv \||j,j\rangle\|^2$ is not zero, we must have $c_2 = j(j+1)$.

We now compute

$$\begin{aligned}
\|J_-|j,m\rangle\|^2 &= \langle j,m|J_-^\dagger J_-|j,m\rangle \\
&= \langle j,m|J_+ J_-|j,m\rangle \\
&= \langle j,m|\left(J^2 - J_3(J_3-1)\right)|j,m\rangle \\
&= [j(j+1) - m(m-1)]\langle j,m|j,m\rangle, \qquad (6.124)
\end{aligned}$$

and deduce that the resulting set of normalized states $|j,m\rangle$ can be chosen to obey

$$\begin{aligned}
J_3|j,m\rangle &= m|j,m\rangle, \\
J_-|j,m\rangle &= \sqrt{j(j+1)-m(m-1)}|j,m-1\rangle, \\
J_+|j,m\rangle &= \sqrt{j(j+1)-m(m+1)}|j,m+1\rangle. \qquad (6.125)
\end{aligned}$$

If we take j to be an integer or a half-integer, we will find that $J_-|j,-j\rangle = 0$. In this case we are able to construct a total of $2j+1$ states, one for each integer-spaced m in the range $-j \le m \le j$. If we select some other fractional value for j, then the set of states will not terminate gracefully, and we will find an infinity of states with $m < -j$. These will have $\|J_-|j,m\rangle\|^2 < 0$, so the resultant representation cannot be unitary.

SU(3)

The strategy of finding ladder operators works for any semi-simple Lie algebra. Consider, for example, $\mathfrak{su}(3) = \mathrm{Lie}(\mathrm{SU}(3))$. The matrix Lie algebra $\mathfrak{su}(3)$ is spanned by the Gell-Mann λ-matrices

$$\begin{aligned}
\hat\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat\lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat\lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\hat\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \hat\lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \hat\lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\
\hat\lambda_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \hat\lambda_8 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \qquad (6.126)
\end{aligned}$$

which form a basis for the real vector space of 3-by-3 traceless, hermitian matrices. They have been chosen and normalized so that

$$\mathrm{tr}\,(\hat{\lambda}_i\hat{\lambda}_j) = 2\delta_{ij}, \tag{6.127}$$

by analogy with the properties of the Pauli matrices. Notice that $\hat{\lambda}_3$ and $\hat{\lambda}_8$ commute with each other, and that this will be true in any representation.

The matrices

$$\begin{aligned} t_\pm &= \frac{1}{2}(\hat{\lambda}_1 \pm i\hat{\lambda}_2), \\ v_\pm &= \frac{1}{2}(\hat{\lambda}_4 \pm i\hat{\lambda}_5), \\ u_\pm &= \frac{1}{2}(\hat{\lambda}_6 \pm i\hat{\lambda}_7). \end{aligned} \tag{6.128}$$

have unit entries, rather like the step up and step down matrices $\sigma_\pm = \frac{1}{2}(\hat{\sigma}_1 \pm i\hat{\sigma}_2)$.

Let us define Λ_i to be abstract operators with the same commutation relations as $\hat{\lambda}_i$, and define

$$\begin{aligned} T_\pm &= \frac{1}{2}(\Lambda_1 \pm i\Lambda_2), \\ V_\pm &= \frac{1}{2}(\Lambda_4 \pm i\Lambda_5), \\ U_\pm &= \frac{1}{2}(\Lambda_6 \pm i\Lambda_7). \end{aligned} \tag{6.129}$$

These are simultaneous eigenvectors of the commuting pair of operators $\mathrm{ad}\,(\Lambda_3)$ and $\mathrm{ad}\,(\Lambda_8)$:

$$\begin{aligned} \mathrm{ad}\,(\Lambda_3)T_\pm &= [\Lambda_3, T_\pm] = \pm 2T_\pm, \\ \mathrm{ad}\,(\Lambda_3)V_\pm &= [\Lambda_3, V_\pm] = \pm V_\pm, \\ \mathrm{ad}\,(\Lambda_3)U_\pm &= [\Lambda_3, U_\pm] = \mp U_\pm, \\ \mathrm{ad}\,(\Lambda_8)T_\pm &= [\Lambda_8, T_\pm] = 0 \\ \mathrm{ad}\,(\Lambda_8)V_\pm &= [\Lambda_8, V_\pm] = \pm\sqrt{3}V_\pm, \\ \mathrm{ad}\,(\Lambda_8)U_\pm &= [\Lambda_8, U_\pm] = \pm\sqrt{3}U_\pm, \end{aligned} \tag{6.130}$$

Thus, in any representation, the $T_\pm$, $U_\pm$, $V_\pm$, act as ladder operators, changing the simultaneous eigenvalues of the commuting pair Λ_3, Λ_8. Their eigenvalues, λ_3, λ_8, are called the *weights*, and there will be a set of such weights

for each possible representation. By using the ladder operators one can go from any weight in a representation to any other, but you cannot get outside this set. The amount by which the ladder operators change the weights are called the *roots* or *root vectors*, and the root diagram characterizes the Lie algebra.

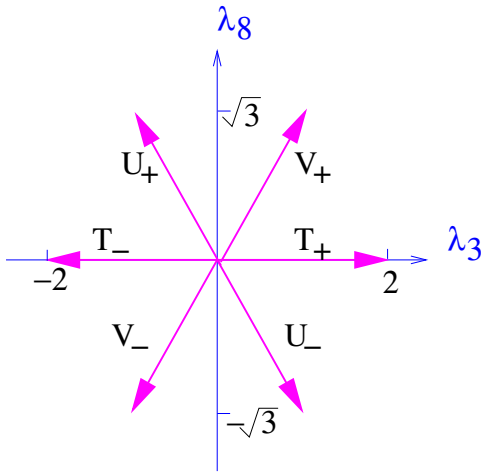

Figure 6.2: *The root vectors of* $su(3)$.

In a finite-dimensional representation there must be a highest weight state $|\lambda_3, \lambda_8\rangle$ that is killed by all three of U_+, T_+ and V_+. We can then obtain all other states in the representation by repeatedly acting on the highest weight state with U_-, T_- or V_- and their products. Since there is usually more than one route by which we can step down from the highest weight to another weight, the weight spaces may be *degenerate* —*i.e* there may be more than one linearly independent state with the same eigenvalues of Λ_3 and Λ_8. Exactly what states are obtained, and with what multiplicity, is not immediately obvious. We will therefore restrict ourselves to describing the outcome of this procedure without giving proofs.

What we find is that the weights in a finite-dimensional representation of $\mathfrak{su}(3)$ form a hexagonally symmetric "crystal" lying on a triangular lattice, and the representations may be labelled by pairs of integers (zero allowed) p, q which give the length of the sides of the crystal. These representations have dimension $d = \frac{1}{2}(p+1)(q+1)(p+q+2)$.

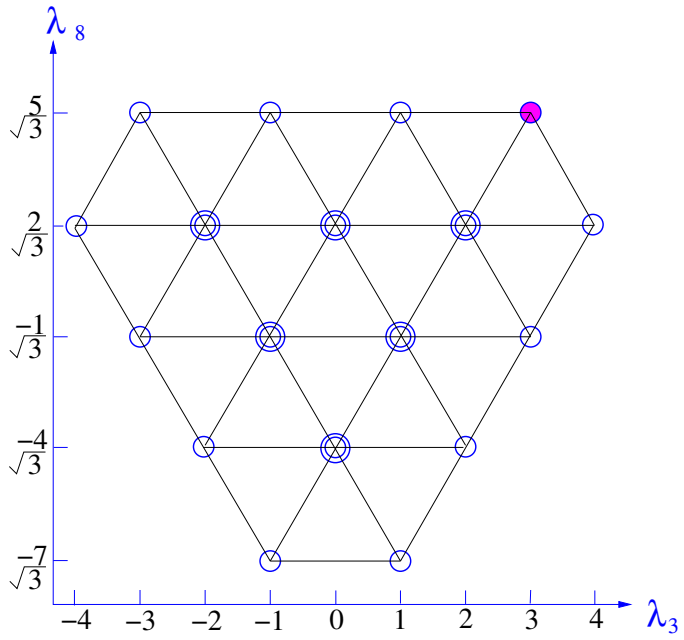

Figure 6.3: *The weight diagram of the 24 dimensional irrep with* $p = 3$, $q = 1$. *The highest weight is shaded.*

Figure 6.3 shows the set of weights occurring in the representation of SU(3) with $p = 3$ and $q = 1$. Each circle represents a state, whose weight (λ_3, λ_8) may be read off from the displayed axes. A double circle indicates that there are two linearly independent vectors with the same weight. A count confirms that the number of independent weights, and hence the dimension of the representation, is 24. For SU(3) representations the degeneracy—*i.e.* the number of states with a given weight—increases by unity at each "layer" until we reach a triangular inner core, all of whose weights have the same degeneracy.

In particle physics applications representations are often labelled by their dimension. The defining representation of SU(3) and its complex conjugate are denoted by 3 and $\bar{3}$,

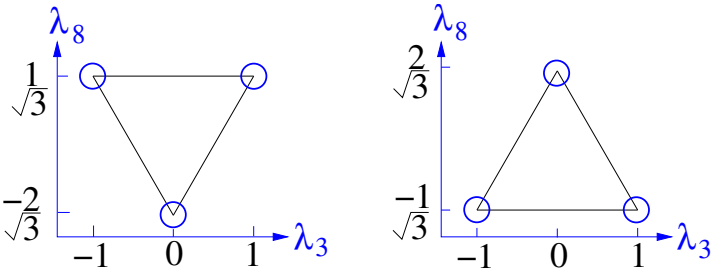

Figure 6.4: *The weight diagrams of the irreps with* $p = 1$, $q = 0$, *and* $p = 0$, $q = 1$, *also known, respectively, as the* 3 *and the* $\overline{3}$.

while the weight diagrams of the eight dimensional adjoint represention and the 10 have shape shown in figure 6.5.

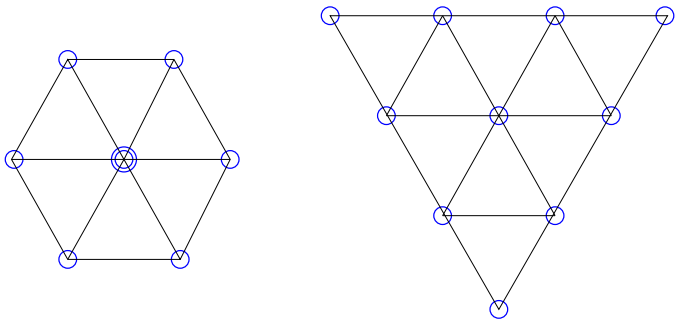

Figure 6.5: *The irreps* 8 *(the adjoint) and* 10.

Cartan algebras: roots and co-roots

For a general simple Lie algebra we may play the same game. We first find a maximal linearly independent set of commuting generators, h_i. The h_i form a basis for the *Cartan algebra*, $\mathfrak{h}$, whose dimension is the *rank* of the Lie algbera. We next find ladder operators by diagonalizing the "ad" action of the h_i on the rest of the algebra.

$$\text{ad}\,(h_i)e_\alpha = [h_i, e_\alpha] = \alpha_i e_\alpha. \tag{6.131}$$

The simultaneous eigenvectors e_α are the ladder operators that change the eigenvalues of the h_i. The corresponding eigenvalues α, thought of as vectors with components α_i, are the *roots*, or root vectors. The roots are therefore the weights of the adjoint representation. It is possible to put factors of "i"

in the appropriate places so that the α_i are real, and we will assume that this has been done. For example in $\mathfrak{su}(3)$ we have already seen that $\alpha_T = (2, 0)$, $\alpha_V = (1, \sqrt{3})$, $\alpha_U = (-1, \sqrt{3})$.

Here are the basic properties and ideas that emerge from this process:

i) Since $\alpha_i \langle e_\alpha, h_j \rangle = \langle \mathrm{ad}\,(h_i) e_\alpha, h_j \rangle = -\langle e_\alpha, [h_i, h_j] \rangle = 0$ we see that $\langle h_i, e_\alpha \rangle = 0$.

ii) Similarly, we see that $(\alpha_i + \beta_i)\langle e_\alpha, e_\beta \rangle = 0$, so the e_α are orthogonal to one another unless $\alpha + \beta = 0$. Since our Lie algebra is semisimple, and consequently the Killing form non-degenerate, we deduce that if α is a root, so is $-\alpha$.

iii) Since the Killing form is non-degenerate, yet the h_i are orthogonal to all the e_α, it must also be non-degenerate when restricted to the Cartan algebra. Thus the metric tensor, $g_{ij} = \langle h_i, h_j \rangle$, must be invertible with inverse g^{ij}. We will use the notation $\alpha \cdot \beta$ to represent $\alpha_i \beta_j g^{ij}$.

iv) If α, β are roots, then the Jacobi identity shows that

$$[h_i, [e_\alpha, e_\beta]] = (\alpha_i + \beta_i)[e_\alpha, e_\beta],$$

so if $[e_\alpha, e_\beta]$ is non-zero then $\alpha + \beta$ is also a root, and $[e_\alpha, e_\beta] \propto e_{\alpha+\beta}$.

v) It follows from iv), that $[e_\alpha, e_{-\alpha}]$ commutes with all the h_i, and since $\mathfrak{h}$ was assumed maximal, it must either be zero or a linear combination of the h_i. A short calculation shows that

$$\langle h_i, [e_\alpha, e_{-\alpha}] \rangle = \alpha_i \langle e_\alpha, e_{-\alpha} \rangle,$$

and, since $\langle e_\alpha, e_{-\alpha} \rangle$ does not vanish, $[e_\alpha, e_{-\alpha}]$ is non-zero. Thus

$$[e_\alpha, e_{-\alpha}] \propto \frac{2\alpha^i}{\alpha^2} h_i \equiv h_\alpha$$

where $\alpha^i = g^{ij}\alpha_j$, and h_α obeys

$$[h_\alpha, e_{\pm\alpha}] = \pm 2 e_{\pm\alpha}.$$

The h_α are called the *co-roots*.

vi) The importance of the co-roots stems from the observation that the triad h_α, $e_{\pm\alpha}$ obey the same commutation relations as $\hat{\sigma}_3$ and $\sigma_\pm$, and so form an $\mathfrak{su}(2)$ subalgebra of $\mathfrak{g}$. In particular h_α (being the analogue of $2J_3$) has only *integer* eigenvalues. For example in $\mathfrak{su}(3)$

$$[T_+, T_-] \;=\; h_T = \Lambda_3,$$

$$\begin{aligned}[V_+, V_-] &= h_V = \frac{1}{2}\Lambda_3 + \frac{\sqrt{3}}{2}\Lambda_8, \\ [U_+, U_-] &= h_U = -\frac{1}{2}\Lambda_3 + \frac{\sqrt{3}}{2}\Lambda_8,\end{aligned}$$

and in the defining representation

$$\begin{aligned}h_T &= \begin{pmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & 0\end{pmatrix} \\ h_V &= \begin{pmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & -1\end{pmatrix} \\ h_U &= \begin{pmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{pmatrix},\end{aligned}$$

have eigenvalues ± 1.

vii) Since

$$\mathrm{ad}\,(h_\alpha)e_\beta = [h_\alpha, e_\beta] = \frac{2\alpha\cdot\beta}{\alpha^2}e_\beta,$$

we conclude that $2\alpha\cdot\beta/\alpha^2$ must be an integer for any pair of roots α, β.

viii) Finally, there can only be one e_α for each root α. If not, and there were an independent e'_α, we could take linear combinations so that $e_{-\alpha}$ and e'_α are Killing orthogonal, and hence $[e_{-\alpha}, e'_\alpha] = \alpha^i h_i \langle e_{-\alpha}, e'_\alpha\rangle = 0$. Thus $\mathrm{ad}\,(e_{-\alpha})e'_\alpha = 0$, and e'_α is killed by the step-down operator. It would therefore be the lowest weight in some $\mathfrak{su}(2)$ representation. At the same time, however, $\mathrm{ad}\,(h_\alpha)e'_\alpha = 2e'_\alpha$, and we know that the lowest weight in any spin J representation cannot have positive eigenvalue.

The conditions that

$$\frac{2\alpha\cdot\beta}{\alpha^2} \in \mathbb{Z}$$

for any pair of roots tightly constrains the possible root systems, and is the key to Cartan and Killing's classification of the semisimple Lie algebras. For example the angle θ between any pair of roots obeys $\cos^2\theta = n/4$ so θ can take only the values $0°, 30°, 45°, 60°, 90°, 120°, 135°, 150°$, or $180°$.

These constraints lead to a complete classification of possible root systems into the infinite families

$$\begin{array}{lll} A_n, & n = 1, 2, \cdots. & \mathfrak{sl}(n+1, \mathbb{C}), \\ B_n, & n = 2, 3, \cdots. & \mathfrak{so}(2n+1, \mathbb{C}), \\ C_n, & n = 3, 3, \cdots. & \mathfrak{sp}(2n, \mathbb{C}), \\ D_n, & n = 4, 5, \cdots. & \mathfrak{so}(2n, \mathbb{C}), \end{array}$$

together with the root systems G_2, F_4, E_6, E_7, and E_8 of the exceptional algebras. The latter do not correspond to any of the classical matrix groups. For example G_2 is the root system of $\mathfrak{g}_2$, the Lie algebra of the group G_2 of automorphisms of the *octonions*. This group is also the subgroup of SL(7) preserving the general totally antisymmetric trilinear form.

The restrictions on n's are to avoid repeats arising from "accidental" isomorphisms. If we allow $n = 1, 2, 3$, in each series, then $C_1 = D_1 = A_1$. This corresponds to $\mathfrak{sp}(2, \mathbb{C}) \cong \mathfrak{so}(3, \mathbb{C}) \cong \mathfrak{sl}(2, \mathbb{C})$. Similarly $D_2 = A_1 + A_1$, corresponding to isomorphism $\mathrm{SO}(4) \cong \mathrm{SU}(2) \times \mathrm{SU}(2)/\mathbb{Z}_2$, while $C_2 = B_2$ implies that, locally, the compact $\mathrm{Sp}(2) \cong \mathrm{SO}(5)$. Finally $D_3 = A_3$ implies that $\mathrm{SU}(4)/\mathbb{Z}_2 \cong \mathrm{SO}(6)$.

6.3.4 Product Representations

Given two representations $\Lambda_i^{(1)}$ and $\Lambda_i^{(2)}$ of $\mathfrak{g}$, we can form a new representation that exponentiates to the tensor product of the corresponding representations of the group G. Motivated by the result of exercise 5.13:

$$\exp(A \otimes I_n + I_m \otimes B) = \exp(A) \otimes \exp(B) \tag{6.132}$$

we set

$$\Lambda_i^{(1\otimes 2)} = \Lambda_i^{(1)} \otimes I^{(2)} + I^{(1)} \otimes \Lambda_i^{(2)}. \tag{6.133}$$

Then

$$\begin{aligned} [\Lambda_i^{(1\otimes 2)}, \Lambda_j^{(1\otimes 2)}] &= ([\Lambda_i^{(1)} \otimes I^{(2)} + I^{(1)} \otimes \Lambda_i^{(2)}), (\Lambda_j^{(1)} \otimes I^{(2)} + I^{(1)} \otimes \Lambda_j^{(2)})] \\ &= [\Lambda_i^{(1)}, \Lambda_j^{(1)}] \otimes I^{(2)} + [\Lambda_i^{(1)}, I^{(1)}] \otimes \Lambda_j^{(2)} \\ &\qquad + \Lambda_i^{(1)} \otimes [I^{(2)}, \Lambda_j^{(2)}] + I^{(1)} \otimes [\Lambda_i^{(2)}, \Lambda_j^{(2)}] \\ &= [\Lambda_i^{(1)}, \Lambda_j^{(1)}] \otimes I^{(2)} + I^{(1)} \otimes [\Lambda_i^{(2)}, \Lambda_j^{(2)}], \end{aligned} \tag{6.134}$$

showing that the $\Lambda_i^{(1\otimes 2)}$ also obey the Lie algebra.

This process of combining representations is analogous to the addition of angular momentum in quantum mechanics. Perhaps more precisely, the addition of angular momentum is an example of this general construction. If representation $\Lambda_i^{(1)}$ has weights $m_i^{(1)}$, *i.e.* $h_i^{(1)}|m^{(1)}\rangle = m_i^{(1)}|m^{(1)}\rangle$, and $\Lambda_i^{(2)}$ has weights $m_i^{(2)}$, then, writing $|m^{(1)}, m^{(2)}\rangle$ for $|m^{(1)}\rangle \otimes |m^{(2)}\rangle$, we have

$$\begin{aligned} h_i^{(1\otimes 2)}|m^{(1)}, m^{(2)}\rangle &= (h_i^{(1)} \otimes 1 + 1 \otimes h_i^{(2)})|m^{(1)}, m^{(2)}\rangle \\ &= (m_i^{(1)} + m_i^{(2)})|m^{(1)}, m^{(2)}\rangle \end{aligned} \tag{6.135}$$

so the weights appearing in the representation $\Lambda_i^{(1\otimes 2)}$ are $m_i^{(1)} + m_i^{(2)}$.

The new representation is usually decomposible. We are familiar with this decomposition for angular momentum where, if $j > j'$,

$$j \otimes j' = (j + j') \oplus (j + j' - 1) \oplus \cdots (j - j'). \tag{6.136}$$

This can be understood from adding weights. For example consider adding the weights of $j = 1/2$, which are $m = \pm 1/2$ to those of $j = 1$, which are $m = -1, 0, 1$. We get $m = -3/2$, $-1/2$ (twice) $+1/2$ (twice) and $m = 3/2$. These decompose as shown in figure 6.6.

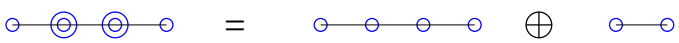

Figure 6.6: *The weights for* $1/2 \otimes 1 = 3/2 \oplus 1/2$.

The rules for decomposing products in other groups are more complicated than for SU(2), but can be obtained from weight diagrams in the same manner. In SU(3), we have, for example

$$\begin{aligned} 3 \otimes \bar{3} &= 1 \oplus 8, \\ 3 \otimes 8 &= 3 \oplus \bar{6} \oplus 15, \\ 8 \otimes 8 &= 1 \oplus 8 \oplus 8 \oplus 10 \oplus \overline{10} \oplus 27. \end{aligned} \tag{6.137}$$

To illustrate the first of these we show, in figure 6.7 the addition of the weights in $\bar{3}$) to each of the weights in the 3.

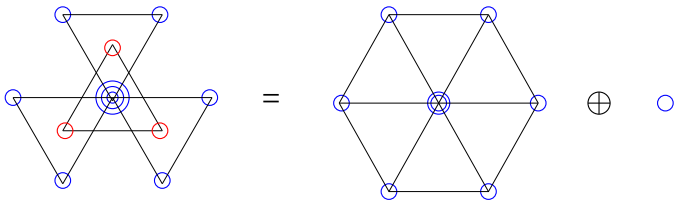

Figure 6.7: *Adding the weights of* 3 *and* $\bar{3}$.

The resultant weights decompose (uniquely) into the weight diagrams for the 8 together with a singlet.

6.3.5 Sub-algebras and branching rules

As with finite groups, a representation that is irreducible under the full Lie group or algebra will in general become reducible when restricted to a sub-group or sub-algebra. The pattern of the decomposition is again called a *branching rule*. Here we provide some examples to illustrate the ideas.

The three operators $V_\pm$ and $h_V = \frac{1}{2}\Lambda_3 + \frac{\sqrt{3}}{2}\Lambda_8$ of $\mathfrak{su}(3)$ form a Lie sub-algebra that is isomorphic to $\mathfrak{su}(2)$ under the map that takes them to $\sigma_\pm$ and σ_3 respectively. When restricted to this sub-algebra, the 8 dimensional representation of $\mathfrak{su}(3)$ becomes reducible, decomposing as

$$8 = 3 \oplus 2 \oplus 2 \oplus 1, \tag{6.138}$$

where the 3, 2 and 1 are the $j = 1$, $\frac{1}{2}$ and 0 representations of $\mathfrak{su}(2)$.

We can visualize this decomposition coming about by first projecting the (λ_3, λ_8) weights to the "m" of the $|j, m\rangle$ labelling of $\mathfrak{su}(2)$ as

$$m = \frac{1}{4}\lambda_3 + \frac{\sqrt{3}}{4}\lambda_8 \tag{6.139}$$

and then stripping off the $\mathfrak{su}(2)$ irreps as we did when decomposing product representions.

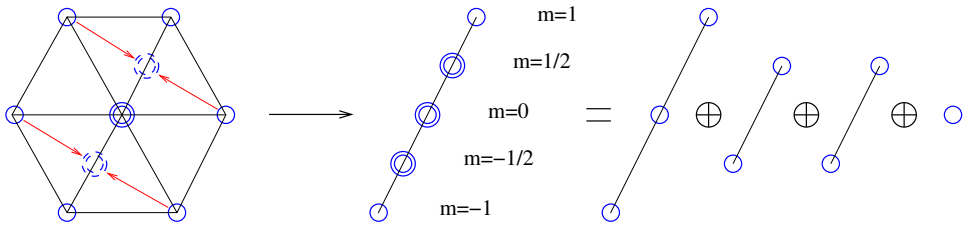

Figure 6.8: *Projection of the* $\mathfrak{su}(3)$ *weights onto* $\mathfrak{su}(2)$*, and the decomposition* $8 = 3 \oplus 2 \oplus 2 \oplus 1$.

This branching pattern occurs in the strong interactions where the mass of the strange quark s being much larger than that of the light quarks u and d causes the octet of pseudo-scalar mesons, which would all have the same mass if SU(3) flavour symmetry was exact, to decompose into the triplet of pions π^+ ,π^0 and π^-, the pair K^+ and K^0, their antiparticles K^- and $\bar{K}^0$, and the singlet η.

There are obviously other $\mathfrak{su}(2)$ sub-algebras consisting of $\{T_\pm, h_T\}$ and $\{U_\pm, h_U\}$, each giving rise to similar decompositions. These sub-algebras, and a continuous infinity of related ones, are obtained from the $\{V_\pm, h_V\}$ algebra by conjugation by elements of SU(3).

Another, unrelated, $\mathfrak{su}(2)$ sub-algebra consists of

$$\begin{aligned} \sigma_+ &\simeq \sqrt{2}(U_+ + T_+), \\ \sigma_- &\simeq \sqrt{2}(U_- + T_-), \\ \sigma_3 &\simeq 2h_V = (\Lambda_3 + \sqrt{3}\Lambda_8). \end{aligned} \tag{6.140}$$

The factor of two between the assignment $\sigma_3 \simeq h_V$ of our previous example and the present assignment $\sigma_3 \simeq 2h_V$ has a non-trivial effect on the branching rules. Under restriction to this new subalgebra, the 8 of $\mathfrak{su}(3)$ decomposes as

$$8 = 5 \oplus 3 \tag{6.141}$$

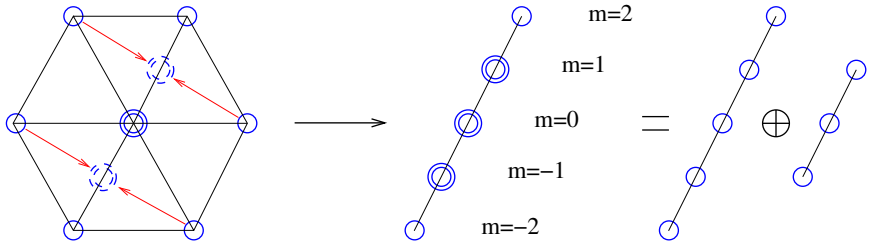

Figure 6.9: *The projection and decomposition for* $8 = 5 \oplus 3$.

where the 5 and 3 are the $j = 2$ and $j = 1$ representations of $\mathfrak{su}(2)$. A clue to the origin and significance of this sub-algebra is found by noting that the 3 and $\bar{3}$ representations of $\mathfrak{su}(3)$ both remain irreducible, but project to the same $j = 1$ representation of $\mathfrak{su}(2)$. Interpreting this $j = 1$ representation as the defining vector representation of $\mathfrak{so}(3)$ suggests (correctly) that our new $\mathfrak{su}(2)$ sub-algebra is the Lie algebra of the SO(3) subgroup of SU(3) consisting of SU(3) matrices with real entries.

6.4 Further Exercises and Problems

Exercise 6.25: *Campbell-Baker-Hausdorff Formulae*. Here are some useful formula for working with exponentials of matrices that do not commute with each other.

a) Let X and X be matrices. Show that

$$e^{tX} Y e^{-tX} = Y + t[X, Y] + \frac{1}{2} t^2 [X, [X, Y]] + \cdots,$$

the terms on the right being the series expansion of $\exp[\mathrm{ad}(tX)]Y$.

b) Let X and δX be matrices. Show that

$$\begin{aligned} e^{-X} e^{X+\delta X} &= 1 + \int_0^1 e^{-tX} \delta X e^{tX} dt + O\left[(\delta X)^2\right] \\ &= 1 + \delta X - \frac{1}{2}[X, \delta X] + \frac{1}{3!}[X, [X, \delta X]] + \cdots + O\left[(\delta X)^2\right] \\ &= 1 + \left(\frac{1 - e^{\mathrm{ad}(X)}}{\mathrm{ad}(X)}\right) \delta X + O\left[(\delta X)^2\right] \end{aligned} \tag{6.142}$$

c) By expanding out the exponentials, show that

$$e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]+\text{higher}},$$

where "higher" means terms higher order in X, Y. The next two terms are, in fact, $\frac{1}{12}[X,[X,Y]]+\frac{1}{12}[Y,[Y,X]]$. You will find the general formula in part d).

d) By using the formula from part b), show that that $e^X e^Y$ can be written as e^Z, where

$$Z = X + \int_0^1 g(e^{\mathrm{ad}(X)} e^{\mathrm{ad}(tY)}) Y \, dt.$$

Here

$$g(z) \equiv \frac{\ln z}{1 - 1/z}$$

has a power series expansion

$$g(z) = 1 + \frac{1}{2}(z-1) + \frac{1}{6}(z-1)^2 + \frac{1}{12}(z-1)^3 + \cdots,$$

which is convergent for $|z| < 1$. Show that $g(e^{\mathrm{ad}(X)} e^{\mathrm{ad}(tY)})$ can be expanded as a double power series in $\mathrm{ad}(X)$ and $\mathrm{ad}(tY)$, provided X and Y are small enough. This $\mathrm{ad}(X)$, $\mathrm{ad}(tY)$ expansion allows us to evaluate the product of two matrix exponentials as a third matrix exponential provided we know their commutator algebra.

Exercise 6.26: SU(2) *Disentangling theorems*: Almost any 2×2 matrix can be factored (Gaussian decomposition) as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix}.$$

Use this trick to work the following problems:

a) Show that

$$\exp\left\{\frac{\theta}{2}(e^{i\phi}\hat{\sigma}_+ - e^{-i\phi}\hat{\sigma}_-)\right\} = \exp(\alpha\hat{\sigma}_+)\exp(\lambda\hat{\sigma}_3)\exp(\beta\hat{\sigma}_-),$$

where $\hat{\sigma}_\pm = (\hat{\sigma}_1 \pm i\hat{\sigma}_2)/2$, and

$$\begin{aligned} \alpha &= e^{i\phi}\tan\theta/2, \\ \lambda &= -\ln\cos\theta/2, \\ \beta &= -e^{-i\phi}\tan\theta/2. \end{aligned}$$

b) Use the fact that the spin-$\frac{1}{2}$ representation of $SU(2)$ is faithful, to show that

$$\exp\left\{\frac{\theta}{2}(e^{i\phi}\hat{J}_+ - e^{-i\phi}\hat{J}_-)\right\} = \exp(\alpha\hat{J}_+)\exp(2\lambda\hat{J}_3)\exp(\beta\hat{J}_-),$$

where $\hat{J}_\pm = \hat{J}_1 \pm i\hat{J}_2$. Take care, the reasoning here is subtle! Notice that the series expansion of exponentials of $\hat{\sigma}_\pm$ truncates after the second term, but the same is **not** true of the expansion of exponentials of the $\hat{J}_\pm$. You need to explain why the formula continues to hold in the absence of this truncation.

Exercise 6.27: *Invariant tensors for* SU(3). Let λ_i be the Gell-Mann lambda matrices. The totally antisymmetric structure constants, f_{ijk}, and a set of totally symmetric constants d_{ijk} are defined by

$$f_{ijk} = \frac{1}{2}\mathrm{tr}\,(\lambda_i[\lambda_j, \lambda_k]), \qquad d_{ijk} = \frac{1}{2}\mathrm{tr}\,(\lambda_i\{\lambda_j, \lambda_k\}).$$

Let $D^8_{ij}(g)$ be the matrices representing SU(3) in "8" — the eight-dimensional adjoint representation.

a) Show that

$$\begin{aligned} f_{ijk} &= D^8_{il}(g)D^8_{jm}(g)D^8_{kn}(g)f_{lmn}, \\ d_{ijk} &= D^8_{il}(g)D^8_{jm}(g)D^8_{kn}(g)d_{lmn}, \end{aligned}$$

and so f_{ijk} and d_{ijk} are *invariant tensors* in the same sense that δ_{ij} and $\epsilon_{i_1\ldots i_n}$ are invariant tensors for SO(n).

b) Let $w_i = f_{ijk}u_jv_k$. Show that if $u_i \to D^8_{ij}(g)u_k$ and $v_i \to D^8_{ij}(g)v_k$, then $w_i \to D^8_{ij}(g)w_k$. Similarly for $w_i = d_{ijk}u_jv_k$. (Hint: show first that the D^8 matrices are real and orthogonal.) Deduce that f_{ijk} and d_{ijk} are *Clebsh-Gordan coefficients* for the $8 \oplus 8$ part of the decomposition

$$8 \otimes 8 = 1 \oplus 8 \oplus 8 \oplus 10 \oplus \overline{10} \oplus 27.$$

c) Similarly show that $\delta_{\alpha\beta}$ and the lambda matrices $(\lambda_i)_{\alpha\beta}$ can be regarded as Clebsch-Gordan coefficients for the decomposition

$$\bar{3} \otimes 3 = 1 \oplus 8.$$

d) Use the graphical method of plotting weights and peeling off irreps to obtain the tensor product decomposition in part b).

Chapter 7

The Geometry of Fibre Bundles

In earlier chapters we have used the language of bundles and connections, but in a relatively casual manner. We deferred proper mathematical definitions until now, because, for the applications we meet in physics, it helps to first have acquired an understanding of the geometry of Lie groups.

7.1 Fibre Bundles

We begin with a formal definition of a bundle and then illustrate the definition with examples from quantum mechanics. These allow us to appreciate the physics that the definition is designed to capture.

7.1.1 Definitions

A smooth *bundle* is a triple (E, π, M) where E and M are manifolds, and $\pi : E \to M$ is a smooth map. The manifold E is called the *total space*, M is the *base space* and π the projection map. The inverse image $\pi^{-1}(x)$ of a point in M (*i.e.* the set of points in E that map to x in M), is the *fibre* over x.

We usually require that all fibres be diffeomorphic to some fixed manifold F. The bundle is then a *fibre bundle*, and F is "the fibre" of the bundle. In a similar vein, we sometimes also refer to the total space E as "the bundle." Examples of possible fibres are vector spaces (in which case we have a *vector bundle*), spheres (in which case we have a sphere bundle), and Lie groups. When the fibre is a Lie group we speak of a *principal bundle*. A principal

bundle can be thought of the parent of various *associated bundles*, which are constructed by allowing the Lie group to act on a fibre. A bundle whose fibre is a one dimensional vector space is called a *line bundle*.

The simplest example of a fibre bundle consists of setting E equal to the Cartesian product $M \times F$ of the base space and the fibre. In this case the projection just "forgets" the point $f \in F$, and so $\pi : (x, f) \mapsto x$.

A more interesting example can be constructed by taking M to be the circle S^1, and F as the one-dimensionsional interval $I = [-1, 1]$. We can assemble these ingredients to make E into a *Möbius strip*. We do this by gluing the copy of I over $\theta = 2\pi$ to that over $\theta = 0$ with a half twist so that the end $-1 \in [-1, 1]$ is attached to $+1$, and vice versa.

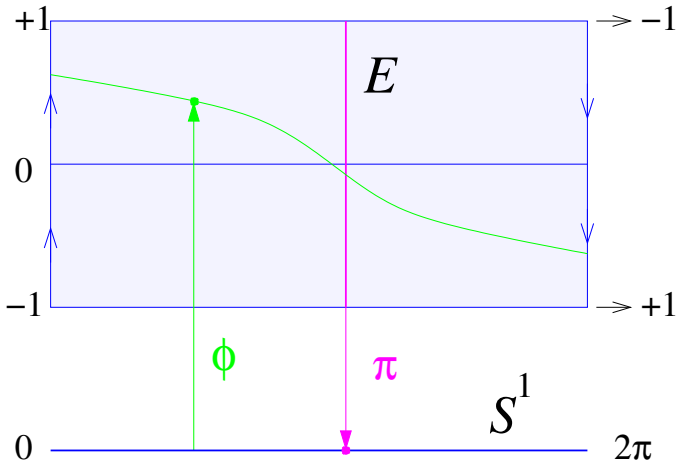

Figure 7.1: *Möbius strip bundle, together with a section ϕ.*

A bundle that is a product $E = M \times F$, is said to be *trivial*. The Möbius strip is not a Cartesian product, and is said to be a *twisted* bundle. The Möbius strip is, however, *locally trivial* in that for each $x \in M$ there is an open retractable neighbourhood $U \subset M$ of x in which E looks like a product $U \times F$. We will assume that all our bundles are locally trivial in this sense. If $\{U_i\}$ is a cover of M (*i.e.* if $M = \bigcup U_i$) by such retractable neighbourhoods, and F is a fixed fibre, then a bundle can be assembled out of the collection of $U_i \times F$ product bundles by giving gluing rules that identify points on the fibre over $x \in U_i$ in the product $U_i \times F$ with points in the fibre over $x \in U_j$ in $U_j \times F$ for each $x \in U_i \cap U_j$. These identifications are made by means of invertible maps $\varphi_{U_i U_j}(x) : F \to F$ that are defined for each x in the overlap

$U_i \cap U_j$. The $\varphi_{U_iU_j}$ are known as *transition functions*. They must satisfy the consistency conditions

$$\begin{aligned} \varphi_{U_iU_i}(x) &= \text{Identity}, \\ \varphi_{U_iU_j}(x) &= \phi^{-1}_{U_jU_i}(x) \\ \varphi_{U_iU_j}(x)\varphi_{U_jU_k}(x) &= \varphi_{U_iU_k}(x), \quad x \in U_i \cap U_j \cap U_k \neq \emptyset. \end{aligned} \tag{7.1}$$

A *section* of a fibre bundle (E, π, M) is a smooth map $\phi : M \to E$ such that $\phi(x)$ lies in the fibre $\pi^{-1}(x)$ over x. Thus $\pi \circ \phi = \text{Identity}$. When the total space E is a product $M \times F$ this ϕ is simply a function $\phi : M \to F$. When the bundle is twisted, as is the Möbius strip, then the section is no longer a function as it takes no unique value at the points x above which the fibres are being glued together. Observe that in the Möbius strip the half-twist forces the section $\phi(x)$ to pass through $0 \in [-1, 1]$. The Möbius bundle therefore has no nowhere-zero globally defined sections. Many twisted bundles have no globally defined sections at all.

7.2 Physics Examples

We now provide three applications where the bundle concept appears in quantum mechanics. The first two illustrations are re-expressions of well-known physics. The third, the geometric approach to quantization, is perhaps less familiar.

7.2.1 Landau levels

Consider the Schrödinger eigenvalue problem

$$-\frac{1}{2m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right) = E\psi \tag{7.2}$$

for a particle moving on a flat two-dimensional torus. We think of the torus as a $L_x \times L_y$ rectangle with the understanding that as a particle disappears through the right-hand boundary it immediately re-appears at the point with the same y co-ordinate on the left-hand boundary; similarly for the upper and lower boundaries. In quantum mechanics we implement these rules by imposing periodic boundary conditions on the wave function:

$$\psi(0, y) = \psi(L_x, y) \qquad \psi(x, 0) = \psi(x, L_y). \tag{7.3}$$

These conditions make the wavefunction a well-defined and continuous function on the torus, in the sense that after pasting the edges of the rectangle together to make a real toroidal surface the function has no jumps, and each point on the surface assigns a unique value to ψ. The wavefunction is a section of an untwisted line bundle with the torus as its base-space, the fibre over (x, y) being the one-dimensional complex vector space $\mathbb{C}$ in which $\psi(x, y)$ takes its value.

Now try to carry out the same program for a particle of charge e moving in a uniform magnetic field B perpendicular to the $x-y$ plane. The Schrödinger equation becomes

$$-\frac{1}{2m}\left(\frac{\partial}{\partial x} - ieA_x\right)^2\psi - \frac{1}{2m}\left(\frac{\partial}{\partial y} - ieA_y\right)^2\psi = E\psi, \tag{7.4}$$

where (A_x, A_y) is the vector potential. We at once meet a problem. Although the magnetic field is constant, the vector potential cannot be chosen to be constant — or even periodic. In the *Landau gauge*, for example, where we set $A_x = 0$, the remaining component becomes $A_y = Bx$. This means that as the particle moves out of the right-hand edge of the rectangle representing the torus we must perform a gauge transformation that prepares it for motion in the (A_x, A_y) field it will encounter when it reappears at the left. If (7.4) holds, then it continues to hold after the simultaneous change

$$\begin{aligned} \psi(x,y) &\rightarrow e^{-ieBL_x y}\psi(x,y) \\ -ieA_y &\rightarrow -ieA_y + e^{-iBL_x y}\frac{\partial}{\partial y}e^{+ieBL_x y} = -ie(A_y - BL_x). \end{aligned} \tag{7.5}$$

At the right-hand boundary $x = L_x$ this gauge transformation resets the vector potential A_y back to its value at the left-hand boundary. Accordingly, we modify the boundary conditions to

$$\psi(0, y) = e^{-ieBL_x y}\psi(L_x, y), \qquad \psi(x, 0) = \psi(x, L_y). \tag{7.6}$$

The new boundary conditions make the wavefunction into a section[1] of a —it twisted line bundle over the torus. The fibre is again the one-dimensional complex vector space $\mathbb{C}$.

[1]That the wave "function" is no longer a function should not be disturbing. Schrödinger's ψ is never really a *function* of space-time. Seen from a frame moving at velocity v, $\psi(x, t)$ acquires factor of $\exp(-imvx - mv^2t/2)$, and this is no way for a self-respecting function of x and t to behave.

We have already met the language in which the gauge field $-ieA_\mu$ is a called *connection* on the bundle, and the associated ieB field is the *curvature*. We will explain how connections fit into the formal bundle language in section 7.3.

The twisting of the boundary conditions by the gauge transformation seems innocent, but within it lurks an important constraint related to the consistency conditions in (7.1). We can find the value of $\psi(L_x, L_y)$ from that of $\psi(0,0)$ by using the relations in (7.6) in the order $\psi(0,0) \to \psi(0, L_y) \to \psi(L_x, L_y)$, or in the order $\psi(0,0) \to \psi(L_x, 0) \to \psi(L_x, L_y)$. Since we must obtain the same $\psi(L_x, L_y)$ whichever route we use, we need to satisfy the condition

$$e^{ieBL_xL_y} = 1. \tag{7.7}$$

This tells us that the Schrödinger problem makes sense only when the magnetic flux BL_xL_y through the torus obeys

$$eBL_xL_y = 2\pi N \tag{7.8}$$

for some integer N. We cannot continuously vary the flux through a finite torus. This means that if we introduce torus boundary conditions as a mathematical convenience in a calculation, then physical effects may depend discontinuously on the field.

The integer N counts the number of times the phase of the wavefunction is twisted as we travel from $x = L_x, y = 0$ to $x = L_x, y = L_y$ gluing the right-hand edge wavefunction to back to the left-hand edge wavefunction. This twisting number is a topological invariant. We have met this invariant before, in section 4.6. It is the first *Chern number* of the wavefunction bundle. If we permit B to become position without altering the total twist N, then quantities such as energies and expectation values can change smoothly with B. If N is allowed to change, however, the these quantities may jump discontinuously.

The energy $E = E_n$ solutions to (7.4) with boundary conditions (7.6) are given by

$$\Psi_{n,k}(x,y) = \sum_{p=-\infty}^{\infty} \psi_n\left(x - \frac{k}{B} - pL_x\right) e^{i(eBpL_x+k)y}. \tag{7.9}$$

Here $\psi_n(x)$ is a harmonic-oscillator wavefunction obeying

$$-\frac{1}{2m}\frac{d^2\psi_n}{dx^2} + \frac{1}{2}m\omega^2\psi_n = E_n\psi_n, \tag{7.10}$$

with $\omega = eB/m$ the classical cyclotron frequency, and $E_n = \omega(n+1/2)$. The parameter k takes the values $2\pi q/L_y$ for q an integer. At each energy E_n we obtain N independent eigenfunctions as q runs from 1 to $eBL_xL_y/2\pi$. These N-fold degenerate states are the *Landau levels*. The degeneracy, being of necessity an integer, provides yet another explanation for why the flux must be quantized.

7.2.2 The Berry connection

Suppose we are in possession of a quantum-mechanical hamiltonian $\hat{H}(\xi)$ depending on some parameters $\xi = (\xi^1, \xi^2, \ldots) \in M$, and know the eigenstates $|n;\xi\rangle$ that obey

$$\hat{H}(\xi)|n;\xi\rangle = E_n(\xi)|n;\xi\rangle. \tag{7.11}$$

If, for fixed n, we can find a smooth family of eigenstates $|n;\xi\rangle$, one for every ξ in the parameter space M, we have a vector bundle over the space M. The fibre above ξ is the one-dimensional vector space spanned by $|n;\xi\rangle$. This bundle is a *sub-bundle* of the product bundle $M \times \mathcal{H}$ where $\mathcal{H}$ is the Hilbert space on which $\hat{H}$ acts. Although the larger bundle is not twisted, the sub-bundle may be. It may also not exist: if the state $|n;\xi\rangle$ become degenerate with another state $|m;\xi\rangle$ at some value of ξ, then both states can vary discontinuously with the parameters, and we wish to exclude this possibility.

In the previous paragraph we considered the evolution of the eigenstates of a time-independent Hamiltonian as we varied its parameters. Another, more physical, evolution is given by solving the *time-dependent* Schrödinger equation

$$i\partial_t|\psi(t)\rangle = \hat{H}(\xi(t))|\psi(t)\rangle \tag{7.12}$$

so as to follow the evolution of a state $|\psi(t)\rangle$ as the parameters are slowly varied. If the initial state $|\psi(0)\rangle$ coincides with with the eigenstate $|0,\xi(0)\rangle$, and if the time evolution of the parameters is slow enough, then $|\psi\rangle$ is expected to remain close to the corresponding eigenstate $|0;\xi(t)\rangle$ of the time-independent Schrödinger equation for the hamiltonian $\hat{H}(\xi(t))$. To determine exactly how "close" it stays, insert the expansion

$$|\psi(t)\rangle = \sum_n a_n(t)|n;\xi(t)\rangle \exp\left\{-i\int_0^t E_0(\xi(t))\,dt\right\}. \tag{7.13}$$

into (7.12) and take the inner-product with $|m;\xi\rangle$. For $m \neq 0$, we expect that the overlap $\langle m;\xi|\psi(t)\rangle$ will be small and of order $O(\partial\xi/\partial t)$. Assuming that this is so, we read off that

$$\dot{a}_0 + a_0\langle 0;\xi|\partial_\mu|0;\xi\rangle\frac{\partial\xi^\mu}{\partial t} = 0, \quad (m=0) \tag{7.14}$$

$$a_m = ia_0\frac{\langle m;\xi|\partial_\mu|0;\xi\rangle}{E_m - E_0}\frac{\partial\xi^\mu}{\partial t}, \quad (m\neq 0) \tag{7.15}$$

up to first-order accuracy in time derivatives of the $|n;\xi(t)\rangle$. Hence

$$|\psi(t)\rangle = e^{i\gamma_{\rm Berry}(t)}\left\{|0;\xi\rangle + i\sum_{m\neq 0}\frac{|m;\xi\rangle\langle m;\xi|\partial_\mu|0;\xi\rangle}{E_m - E_0}\frac{\partial\xi^\mu}{\partial t} + \ldots\right\}e^{-i\int_0^t E_0(t)dt}, \tag{7.16}$$

where the dots refer to terms of higher order in time derivatives.

Equation (7.16) constitutes the first two terms in a systematic *adiabatic series expansion*. The factor $a_0(t) = \exp\{i\gamma_{\rm Berry}(t)\}$ is the solution of the differential equation (7.14). The angle $\gamma_{\rm Berry}$ is known as *Berry's phase* after the British mathematical physicist Michael Berry. It is needed to take up the slack between the arbitrary ξ-dependent phase choice at our disposal when defining the $|0;\xi\rangle$, and the specific phase selected by the Schrödinger equation as it evolves the state $|\psi(t)\rangle$. Berry's phase is also called the *geometric phase* because it depends only on the Hillbert-space geometry of the family of states $|0;\xi\rangle$, and not on their energies. We can write

$$\gamma_{\rm Berry}(t) = i\int_0^t \langle 0;\xi|\partial_\mu|0;\xi\rangle\frac{\partial\xi^\mu}{\partial t}\,dt \tag{7.17}$$

and regard the one-form

$$A_{\rm Berry} \stackrel{\rm def}{=} \langle 0;\xi|\partial_\mu|0;\xi\rangle d\xi^\mu = \langle 0;\xi|d|0;\xi\rangle \tag{7.18}$$

as a connection on the bundle of states over the space of parameters. The equation

$$\dot{\xi}^\mu\left(\frac{\partial}{\partial\xi^\mu} + A_{{\rm Berry},\mu}\right)\psi = 0 \tag{7.19}$$

then identifies the Schrödinger time evolution with parallel transport. It seems reasonable to refer to this particular parallel transport as "Berry transport."

In order for corrections to the approximation $|\psi(t)\rangle \approx (\text{phase})|0;\xi(t)\rangle$ to remain small, we need the denominator $(E_m - E_0)$ to remain large when compared to its numerator. The state that we are following must therefore never become degenerate with any other state.

Monople bundle

Consider, for example a spin-1/2 particle in a magnetic field. If the field points in direction $\mathbf{n}$, the Hamiltonian is

$$\hat{H}(\mathbf{n}) = \mu|B|\,\hat{\boldsymbol{\sigma}}\cdot\mathbf{n} \tag{7.20}$$

There are are two eigenstates with energy $E_\pm = \pm\mu|B|$. Let is focus on the eigenstate $|\psi_+\rangle$ corresponding to E_+. For each $\mathbf{n}$ we can obtain an E_+ eigenstate by applying the projection operator

$$\hat{P} = \frac{1}{2}(\mathbf{I} + \mathbf{n}\cdot\hat{\boldsymbol{\sigma}}) = \frac{1}{2}\begin{pmatrix} 1+n_z & n_x - in_y \\ n_x + in_y & 1-n_z \end{pmatrix} \tag{7.21}$$

to almost any vector, and then multiplying by a real normalization constant $\mathcal{N}$. Applying $\hat{P}$ to a "spin-up" state, for example gives

$$= \mathcal{N}\frac{1}{2}(\mathbf{I} + \mathbf{n}\cdot\hat{\boldsymbol{\sigma}})\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\cos\theta/2\\ e^{i\phi}\sin\theta/2\end{pmatrix}. \tag{7.22}$$

Here θ and ϕ are spherical polar angles on S^2 that specify the direction of $\mathbf{n}$.

Although the bundle of $E = E_+$ eigenstates is globally defined, the family of states $|\psi_+^{(1)}(\mathbf{n})\rangle$ that we have obtained, and would like to use as base for the fibre over $\mathbf{n}$, becomes singular when $\mathbf{n}$ is in the vicinity of the south pole $\theta = \pi$. This is because the factor $e^{i\phi}$ is multivalued at the south pole. There is no problem at the north pole because the ambiguous phase $e^{i\phi}$ multiples $\sin\theta/2$, which is zero there.

Near the south pole, however, we can project from a "spin-down" state to find.

$$|\psi_+^{(2)}(\mathbf{n})\rangle = \mathcal{N}\frac{1}{2}(\mathbf{I} + \mathbf{n}\cdot\hat{\boldsymbol{\sigma}})\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}e^{-i\phi}\cos\theta/2\\ \sin\theta/2\end{pmatrix}. \tag{7.23}$$

This family of eigenstates is smooth near the south pole, but is ill-defined at the north pole. As in section 4.6, we are compelled to cover the sphere S^2

by two caps D_+ and D_-, and use $|\psi_+^{(1)}\rangle$ in D_+ and $|\psi_+^{(2)}\rangle$ in D_-. The two families are related by

$$|\psi_+^{(1)}(\mathbf{n})\rangle = e^{i\phi}|\psi_+^{(2)}(\mathbf{n})\rangle \tag{7.24}$$

in the cingular overlap region $D_+ \cap D_-$. Here $e^{i\phi}$ is the transition function that glues the two families of eigenstates together.

The Berry connections are

$$\begin{aligned} A_+^{(1)} &= \langle\psi_+^{(1)}|d|\psi_+^{(1)}\rangle = \frac{i}{2}(\cos\theta - 1)d\phi \\ A_+^{(2)} &= \langle\psi_+^{(2)}|d|\psi_+^{(2)}\rangle = \frac{i}{2}(\cos\theta + 1)d\phi. \end{aligned} \tag{7.25}$$

In their common domain of definition, they are related by a gauge transformation

$$A_+^{(2)} = A_+^{(1)} + id\phi. \tag{7.26}$$

The curvature of either connection is

$$dA = -\frac{i}{2}\sin\theta d\theta d\phi = -\frac{i}{2}d(\text{Area}). \tag{7.27}$$

The curvature being the area two-form tells us that when we slowly change the direction of B and bring it back to its original orientation the spin state will, in addition to the *dynamical phase* $\exp\{-iE_+t\}$, have accumulated a phase equal to (minus) one-half of the area enclosed by the trajectory of $\mathbf{n}$ on S^2. The two-form field dA can be though of as the flux of a magnetic monople residing at the centre of the sphere. The bundle of one-dimensional vector spaces span$[|\psi_+(\mathbf{n})\rangle]$ over S^2 is therefore called the *monople bundle*.

7.2.3 Quantization

In this section we provide a short introduction to *geometric quantization*. This idea, due largely to Kirilov, Kostant and Souriau, extends the familiar technique of canonical quantization to phase spaces with more structure than that of the harmonic oscillator. We illustrate the formalism by quantizing spin, and show how the resulting Hilbert space provides an example of the Borel-Weil-Bott construction of the representations of a semi-simple Lie group as spaces of sections of holomorphic line bundles.

Prequantization

The passage from classical mechanics to quantum mechanics involves replacing the classical variables by operators in such a way that the classical Poisson-bracket algebra is mirrored by the operator commutator algebra. In general, this process of *quantization* is not possible without making some compromises. It is, however, usually possible to *pre-quantize* a phase-space with its associated Poisson algebra.

Let M be a $2n$-dimensional classical phase-space with its closed symplectic form ω. Classically a function $f : M \to \mathbb{R}$ give rise to a Hamiltonian vector field v_f *via* Hamilton's equations

$$df = -i_{v_f}\omega. \tag{7.28}$$

We saw in section 2.4.2 that the closure condition $d\omega = 0$ ensures that that the Poisson bracket

$$\{f, g\}{=}v_f g = \omega(v_f, v_g) \tag{7.29}$$

obeys

$$[v_f, v_g] = v_{\{f,g,\}}. \tag{7.30}$$

Now suppose that the cohomology class of $(2\pi\hbar)^{-1}\omega$ in $H^2(M, \mathbb{R})$ has the property that its integrals over cycles in $H_2(M, \mathbb{Z})$ are integers. Then (it can be shown) there exists a line bundle L over M with curvature $F = -i\hbar^{-1}\omega$. If we locally write $\omega = d\eta$, where $\eta = \eta_\mu dx^\mu$, then the connection one-form is $A = -i\hbar^{-1}\eta$ and the covariant derivative

$$\nabla_v \equiv v^\mu(\partial_\mu - i\hbar^{-1}\eta_\mu), \tag{7.31}$$

acts on sections of the Line bundle. The corresponding curvature is

$$F(u, v){=}[\nabla_u, \nabla_v] - \nabla_{[u,v]} = -i\hbar^{-1}\omega(u, v). \tag{7.32}$$

We define a pre-quantized operator $\widehat{\rho}(f)$ that acting on sections $\Psi(x)$ of the line bundle corresponds to the classical function f:

$$\widehat{\rho}(f) \stackrel{\text{def}}{=} -i\hbar\nabla_{v_f} + f. \tag{7.33}$$

For hamiltonian vector fields v_f and v_g we have

$$\begin{aligned}
[\hbar\nabla_{v_f} + if, \nabla_{v_g}] &= \hbar\nabla_{[v_f,v_g]} - i\omega(v_f, v_g) + i[f, \nabla_{v_g}] \\
&= \hbar\nabla_{[v_f,v_g]} - i(i_{v_f}\omega + df)(v_g) \\
&= \hbar\nabla_{[v_f,v_g]},
\end{aligned} \tag{7.34}$$

and so

$$\begin{aligned}[-i\hbar\nabla_{v_f}+f,-i\hbar\nabla_{v_g}+g] &= -\hbar^2\nabla_{[v_f,v_g]}-i\hbar\, v_f g \\ &= -i\hbar(-i\hbar\nabla_{[v_f,v_g]}+\{f,g\}) \\ &= -i\hbar(-i\hbar\nabla_{v_{\{f,g\}}}+\{f,g\}). \qquad (7.35)\end{aligned}$$

Equation (7.35) is Dirac's quantization rule:

$$i\,[\widehat{\rho}(f),\widehat{\rho}(g)] = \hbar\,\widehat{\rho}(\{f,g\}). \qquad (7.36)$$

The process of quantization is completed, when possible, by defining a *polarization*. This is a restriction on the variables that we allow the wavefunctions to depend on. For example, if there is a global set of Darboux co-ordinates p, q we may demand that the wavefunction depend only on q, or only on the combination $p + iq$. Such a restriction is necessary so that the representation $f \mapsto \widehat{\rho}(f)$ is *irreducible.* Since globally defined Darboux co-ordinates do not usually exist, this step is the hard part of quantization.

The precise definition of a polarized section is rather complicated. We can only sketch it here, but give a concrete example in the next section. At each point $x \in M$ the symplectic form defines a skew bilinear form. We seek a Lagrangian subspace of $V_x \subset TM_p$ for this form. A Lagrangian subspace is one such that $V_x = V_x^{\perp}$. For example, if

$$\omega = dp_1 \wedge dq_1 + dp_2 \wedge dq_2, \qquad (7.37)$$

then the space spanned by the ∂_q's is Lagrangian, as is the space spanned by the ∂_p's. We allow the coefficients of the vectors in V_x to be complex numbers. The vectors fields spanning the V_x's form a distribution. We require it to be integrable, so that the V_x are the tangent spaces to a global foliation of M. A section Ψ of the Line bundle is *polarized* if $\nabla_{\bar{\xi}}\Psi = 0$ for all $\bar{\xi} \in V_x$.

We define an inner product on the space of polarized sections by using the Liouville measure $\omega^n/n!$ on the phase space. The quantum Hilbert space then consists of finite-norm polarized sections of L. Only classical functions that give rise to polarization-compatible vector fields will have their Poisson-bracket algebra coincide with the quantum commutator algebra.

Quantizing spin

To illustrate these ideas, we quantize spin. The classical mechanics of spin was discussed in section 2.4.2. There we showed that the appropriate phase

space is the 2-sphere equipped with a symplectic form proportional to the area form. Here we must be specific about the constant of proportionality. We choose units in which $\hbar \to 1$, and take $\omega = j\, d(\text{Area})$. The integrality of $\omega/2\pi$ requires that j be an integer or half integer. We will assume that j is positive.

We parametrize the 2-sphere with complex sterographic co-ordinates z, $\overline{z}$ which are constructed similarly to those in section 3.4.3. This choice will allow us to impose a natural complex polarization on the wavefunctions. In contrast to section 3.4.3, however, it is here convenient to make the point $z = 0$ correspond to the *south* pole, so the polar co-ordinates θ, ϕ, on the sphere are related to z, $\overline{z}$ *via*

$$\begin{aligned} \cos\theta &= \frac{|z|^2 - 1}{|z|^2 + 1}, \\ e^{i\phi}\sin\theta &= \frac{2z}{|z|^2 + 1}, \\ e^{-i\phi}\sin\theta &= \frac{2\overline{z}}{|z|^2 + 1}. \end{aligned} \tag{7.38}$$

In terms of the z, $\overline{z}$ co-ordinates

$$\omega = \frac{2ij}{(1 + |z|^2)^2}\, dz \wedge d\overline{z}. \tag{7.39}$$

As long as we avoid the north pole where $z = \infty$, we can write

$$\omega = d\left\{ ij \frac{z\, d\overline{z} - \overline{z}\, dz}{1 + |z|^2} \right\} = d\eta, \tag{7.40}$$

and so the local connection form has components proportional to

$$\eta_z = -ij\frac{\overline{z}}{|z|^2 + 1}, \quad \eta_{\overline{z}} = ij\frac{z}{|z|^2 + 1}. \tag{7.41}$$

The covariant derivatives are therefore

$$\nabla_z = \frac{\partial}{\partial z} - j\frac{\overline{z}}{|z|^2 + 1}, \quad \nabla_{\overline{z}} = \frac{\partial}{\partial \overline{z}} + j\frac{z}{|z|^2 + 1}. \tag{7.42}$$

We impose the polarization condition that $\nabla_{\overline{z}}\Psi = 0$. This condition requires the allowed sections to be of the form

$$\Psi(z, \overline{z}) = (1 + |z|^2)^{-j}\psi(z), \tag{7.43}$$

where ψ depends only on z. It is natural to combine the $(1+|z|^2)^{-j}$ prefactor with the Liouville measure so that the inner product becomes

$$\langle\psi|\chi\rangle = \frac{2j+1}{2\pi i}\int_{\mathbb{C}} \frac{d\overline{z}\wedge dz}{(1+|z|^2)^{2j+2}}\,\overline{\psi(z)}\chi(z). \tag{7.44}$$

The normalizable wavefunctions are then polynomials in z of degree less than or equal to $2j$, and a complete orthonormal set is given by

$$\psi_m(z) = \sqrt{\frac{2j!}{(j-m)!(j+m)!}}\,z^{j+m}, \quad -j\le m\le j. \tag{7.45}$$

We desire to find the quantum operators $\widehat{\rho}(J_i)$ corresponding to the components

$$J_1 = j\sin\theta\cos\phi, \quad J_2 = j\sin\theta\sin\phi, \quad J_3 = j\cos\theta, \tag{7.46}$$

of a classical spin $\mathbf{J}$ of magnitude j, and also to the ladder-operator components $J_\pm = J_1 \pm iJ_2$. In our complex co-ordinates these functions become

$$\begin{aligned} J_3 &= j\frac{|z|^2-1}{|z|^2+1}, \\ J_+ &= j\frac{2z}{|z|^2+1}, \\ J_- &= j\frac{2\overline{z}}{|z|^2+1}. \end{aligned} \tag{7.47}$$

Hamilton's equations read

$$\begin{aligned} \dot{z} &= i\frac{(1+|z|^2)^2}{2j}\frac{\partial H}{\partial\overline{z}}, \\ \dot{\overline{z}} &= -i\frac{(1+|z|^2)^2}{2j}\frac{\partial H}{\partial z}, \end{aligned} \tag{7.48}$$

and the Hamiltonian vector fields corresponding to the classical phase space functions J_3, J_+ and J_- are

$$\begin{aligned} v_{J_3} &= iz\partial_z - i\overline{z}\partial_{\overline{z}}, \\ v_{J_+} &= -iz^2\partial_z - i\partial_{\overline{z}}, \\ v_{J_-} &= i\partial_z + i\overline{z}^2\partial_{\overline{z}}. \end{aligned} \tag{7.49}$$

Using the recipe (7.33) for $\widehat{\rho}(H)$ from the previous section, and the fact that $\nabla_{\overline{z}}\Psi = 0$, we find, for example, that

$$\begin{aligned}\widehat{\rho}(J_+)(1+|z|^2)^{-j}\psi(z) &= \left[-z^2\left(\frac{\partial}{\partial z} - \frac{j\overline{z}}{(1+|z|^2)}\right) + \frac{2jz}{(1+|z|^2)}\right](1+|z|^2)^{-j}\psi(z), \\ &= (1+|z|^2)^{-j}\left[-z^2\frac{\partial}{\partial z} + 2jz\right]\psi \qquad (7.50)\end{aligned}$$

It is natural to define operators

$$\widehat{J}_i = (1+|z|^2)^j\widehat{\rho}(J_i)(1+|z|^2)^{-j} \qquad (7.51)$$

that act only on the z-polynomial part $\psi(z)$ of the section $\Psi(z,\overline{z})$. We then have

$$\widehat{J}_+ = -z^2\frac{\partial}{\partial z} + 2jz. \qquad (7.52)$$

Similarly, we find that

$$\widehat{J}_- = \frac{\partial}{\partial z}, \qquad (7.53)$$

$$\widehat{J}_3 = z\frac{\partial}{\partial z} - j. \qquad (7.54)$$

These operators obey the $\mathfrak{s}u(2)$ Lie algebra relations

$$\begin{aligned}[\widehat{J}_3, \widehat{J}_\pm] &= \pm\widehat{J}_\pm, \\ [\widehat{J}_+, \widehat{J}_-] &= 2\widehat{J}_3, \qquad (7.55)\end{aligned}$$

and act on the $\psi_m(z)$ monomials as

$$\begin{aligned}\widehat{J}_3\psi_m(z) &= m\,\psi_m(z) \\ \widehat{J}_\pm\psi_m(z) &= \sqrt{j(j+1) - m(m\pm 1)}\,\psi_{m\pm 1}(z). \qquad (7.56)\end{aligned}$$

This is the familiar action of the $\mathfrak{s}u(2)$ generators on $|j,m\rangle$ basis states.

Exercise 7.1: Show that with respect to the inner product (7.44) we have

$$\widehat{J}_3^\dagger = \widehat{J}_3, \quad \widehat{J}_+^\dagger = \widehat{J}_-.$$

Coherent states and the Borel-Weil-Bott theorem

We now explain how the spin wavefunctions $\psi_m(z)$ can be understood as sections of a *holomorphic* line bundle.

Suppose that we have a compact Lie group G and a unitary irreducible representation $g \in G \mapsto D^J(g)$. Let $|0\rangle$ be the normalized highest (or lowest) weight state in the representation space. Consider the states

$$|g\rangle = D^J(g)|0\rangle, \quad \langle g| = \langle 0| \left[D^J(g)\right]^\dagger . \tag{7.57}$$

The $|g\rangle$ compose a family of *generalized coherent states*.[2] There is a continuous infinity of the $|g\rangle$, and so they cannot constitute an orthonormal set on the finite dimensional representation space. The matrix-element orthogonality property (6.79), however, provides us us with a useful *over-completeness relation*

$$\mathbf{I} = \frac{\dim(J)}{\mathrm{Vol}\, G} \int_G |g\rangle\langle g|. \tag{7.58}$$

The integral is over all of G, but many points in G give the same contribution. The *maximal torus* T is the abelian subgroup of G obtained by exponentiating elements of the Cartan algebra. Because any weight vector is a common eigenvector of the Cartan algebra, elements of T leave $|0\rangle$ fixed up to a phase. The set of distinct $|g\rangle$ in the integral can therefore be identified with G/T. This coset space is always an even dimensional manifold, and thus a candidate phase space.

Consider in particular the spin-j representation of SU(2). The coset space G/T is then $\mathrm{SU}(2)/U(1) \simeq S^2$. We can write a general element of SU(2) as

$$U = \exp(\overline{z}J_+)\exp(\theta J_3)\exp(\gamma J_-) \tag{7.59}$$

for some complex parameters $\overline{z}$, θ and γ which are functions of the three real co-ordinates that parameterize SU(2). We let U act on the lowest-weight state $|j,-j\rangle$. The rightmost factor has no effect on the lowest weight state, and the middle factor only multiplies it by a constant. We therefore restrict our attention to the states

$$|\overline{z}\rangle = \exp(\overline{z}J_+)|j,-j\rangle, \quad \langle z| = \langle j,-j|\exp(zJ_-) = (|\overline{z}\rangle)^\dagger. \tag{7.60}$$

[2] A. Perelomov, *Generalized Coherent States and their Applications*, (Springer-Verlag, Berlin 1986).

These states are not normalized, but have the advantage that the $\langle z|$ are holomorphic in the parameter z — *i.e.* they depend on z, but not on $\overline{z}$.

The set of distinct $|\overline{z}\rangle$ can still be identified with the 2-sphere, and z, $\overline{z}$ are its complex sterographic co-ordinates. This identification is an example of a general property of compact Lie groups:

$$G/T \cong G_{\mathbb{C}}/B_+. \tag{7.61}$$

Here $G_{\mathbb{C}}$ is the *complexification* of G — the group G, but with its parameters allowed to be complex — and B_+ is the Borel group whose Lie algebra consists of the Cartan algebra together with the step-up ladder operators.

The inner product of two $|\overline{z}\rangle$ states is

$$\langle z'|\overline{z}\rangle = (1+\overline{z}z')^{2j}, \tag{7.62}$$

and the eigenstates $|j,m\rangle$ of J^2 and J_3 possess *coherent state wavefunctions*

$$\psi^{(1)}_m(z) \equiv \langle z|j,m\rangle = \sqrt{\frac{2j!}{(j-m)!(j+m)!}}\, z^{j+m}. \tag{7.63}$$

We recognize these as our spin wavefunctions from the previous section.

The over-completeness relation can be written as

$$\mathbf{I} = \frac{2j+1}{2\pi i}\int \frac{d\overline{z}\wedge dz}{(1+\overline{z}z)^{2j+2}}\,|\overline{z}\rangle\langle z|, \tag{7.64}$$

and provides the inner product for the coherent-state wavefunctions. If $\psi(z)=\langle z|\psi\rangle$ and $\chi(z)=\langle z|\chi\rangle$ then

$$\begin{aligned}\langle\psi|\chi\rangle &= \frac{2j+1}{2\pi i}\int \frac{d\overline{z}\wedge dz}{(1+\overline{z}z)^{2j+2}}\,\langle\psi|\overline{z}\rangle\langle z|\chi\rangle \\ &= \frac{2j+1}{2\pi i}\int \frac{d\overline{z}\wedge dz}{(1+\overline{z}z)^{2j+2}}\,\overline{\psi(z)}\chi(z),\end{aligned} \tag{7.65}$$

which coincides with (7.44).

The wavefunctions $\psi^{(1)}_m(z)$ are singular at the north pole where $z=\infty$. Indeed there is no actual state $\langle\infty|$ because the phase of this putative limiting state would depend on the direction from which we approach the point at infinity. We may, however, define a second family of coherent states

$$|\overline{\zeta}\rangle_2 = \exp(\overline{\zeta}J_-)|j,j\rangle, \quad {}_2\langle\zeta| = \langle j,j|\exp(\zeta J_+), \tag{7.66}$$

and form the wavefunctions

$$\psi^{(2)}_m(\zeta) = {}_2\langle\zeta|j,m\rangle. \tag{7.67}$$

These new states and wavefunctions are well defined in the vicinity of the north pole, but singular near the south pole.

To find the relation between $\psi^{(2)}(\zeta)$ and $\psi^{(1)}(z)$ we note that the matrix identity

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -z^{-1} & 1 \end{bmatrix} \begin{bmatrix} -z & 0 \\ 0 & -z^{-1} \end{bmatrix} \begin{bmatrix} 1 & z^{-1} \\ 0 & 1 \end{bmatrix}, \tag{7.68}$$

coupled with the faithfulness of the spin-$\frac{1}{2}$ representation of SU(2), implies the relation

$$\hat{w}\exp(zJ_+) = \exp\left(-z^{-1}J_-\right)(-z)^{2J_3}\exp\left(z^{-1}J_+\right), \tag{7.69}$$

where $\hat{w} = \exp(-i\pi J_2)$. We also note that

$$\langle j,j|\hat{w} = (-1)^{2j}\langle j,-j|, \quad \langle j,-j|\hat{w} = \langle j,j|. \tag{7.70}$$

Thus,

$$\begin{aligned} \psi^{(1)}_m(z) &= \langle j,-j|e^{zJ_-}|j,m\rangle \\ &= (-1)^{2j}\langle j,j|\hat{w}\, e^{zJ_-}|j,m\rangle \\ &= (-1)^{2j}\langle j,j|e^{-z^{-1}J_-}(-z)^{2J_3}e^{z^{-1}J_+}|j,m\rangle \\ &= (-1)^{2j}(-z)^{2j}\langle j,j|e^{z^{-1}J_+}|j,m\rangle \\ &= z^{2j}\psi^{(2)}_m(z^{-1}). \end{aligned} \tag{7.71}$$

The transition function z^{2j} that relates $\psi^{(1)}_m(z)$ to $\psi^{(2)}_m(\zeta \equiv 1/z)$ depends only on z. We therefore say that the wavefunctions $\psi^{(1)}_m(z)$ and $\psi^{(2)}_m(\zeta)$ are the local components of a global section $\psi_m \leftrightarrow |j,m\rangle$ of a *holomorphic line bundle*. The requirement that the transition function and its inverse be holomorphic and single valued in the overlap of the z and ζ coordinate patches forces $2j$ to be an integer. The ψ_m form a basis for the space of global holomorphic sections of this bundle.

Borel, Weil and Bott showed that any finite-dimensional representation of a semi-simple Lie group G can be realized as the space of global holomorphic sections of a line bundle over $G_{\mathbb{C}}/B_+$. This bundle is constructed from the

highest (or lowest) weight vectors in the representation by a natural generalization of the method we have used for spin. This idea has been extended by Ed Witten and others to infinite dimensional Lie groups, where it can be used, for example, to quantize two-dimensional gravity.

Exercise 7.2: Normalize the states $|\overline{z}\rangle$, $\langle z|$, by multiplying them by $N = (1+|z|^2)^{-j}$. Show that

$$N^2\langle z|J_3|\overline{z}\rangle = j\frac{|z|^2-1}{|z|^2+1},$$
$$N^2\langle z|J_+|\overline{z}\rangle = j\frac{2z}{|z|^2+1},$$
$$N^2\langle z|J_-|\overline{z}\rangle = j\frac{2\overline{z}}{|z|^2+1},$$

thus confirming the identification of z, $\overline{z}$ with the complex stereographic coordinates on the sphere.

7.3 Working in the Total Space

We have mostly considered a bundle to be a collection of mathematical objects attached to a base space, rather than treating the bundle as a geometric object in its own right. In this section we will demonstrate the advantages to be gained from the latter viewpoint.

7.3.1 Principal Bundles and Associated bundles

The fibre bundles that arise in a gauge theory with Lie group G are called *principal G-Bundles*, and the fields and wavefunctions are sections of *associated* bundles. A principal G-bundle comprises the total space, which we here call P, together with the projection, π, to the base space M. The fibre can be regarded as a copy of G

$$\pi : P \to M, \qquad \pi^{-1}(x) \cong G. \tag{7.72}$$

Strictly speaking, the fibre is only required to be a homogeneous space on which G acts freely and transitively on the *right*; $x \to xg$. Such a set can be identified with G after we have selected a fiducial point $f_0 \in F$ to be the group identity. There is no canonical choice for f_0 and, if the bundle is

twisted, there can be no globally smooth choice. This is because a smooth choice for f_0 in the fibres above an open subset $U \subseteq M$ makes P locally into a product $U \times G$. Being able to extend U to the entirety of M means that P is trivial. We will, however, make use of local assignments $f_0 \mapsto e$ to introduce bundle co-ordinate charts in which P is locally a product, and therefore parametrized by ordered pairs (x, g) with $x \in U$ and $g \in G$.

To understand the bundles *associated* with P, it is simplest to define the sections of the associated bundle. Let $\varphi_i(x, g)$ be a function on the total space P with a set of indices i carrying some representation $g \mapsto D(g)$ of G. We say that $\varphi_i(x, g)$ is a section of an associated bundle if it varies in a particular way as we run up and down the fibres by acting on them from the *right* with elements of G. We require

$$\varphi_i(x, gh) = D_{ij}(h^{-1})\varphi_j(x, g). \tag{7.73}$$

These sections can be thought of as wavefunctions for a particle moving in a gauge field on the base space. The choice of representation D plays the role of "charge," and (7.73) are the gauge transformations. Note that we must take h^{-1} as the argument of D in order for the transformation to be consistent under group multiplication:

$$\begin{aligned} \varphi_i(x, gh_1h_2) &= D_{ij}(h_2^{-1})\varphi_j(x, gh_1) \\ &= D_{ij}(h_2^{-1})D_{jk}(h_1^{-1})\varphi_k(x, g) \\ &= D_{ik}(h_2^{-1}h_1^{-1})\varphi_k(x, g) \\ &= D_{ik}((h_1h_1)^{-1})\varphi_k(x, g). \end{aligned} \tag{7.74}$$

The construction of the associated bundle itself requires rather more abstraction. Suppose that the matrices $D(g)$ act on the vector space V. Then the total space P_V of the associated bundle consists of equivalence classes of $P \times V$ under the relation $((x, g), \mathbf{v}) \sim ((x, gh), D(h^{-1})\mathbf{v})$ for all $\mathbf{v} \in V$, $(x, g) \in P$ and $h \in G$. The set of G-action equivalence classes in a Cartesian product $A \times B$ is usually denoted by $A \times_G B$. Our total space is therefore

$$P_V = P \times_G V. \tag{7.75}$$

We find it conceptually easier to work with the sections as defined above, rather than with these equivalence classes.

7.3.2 Connections

A gauge field is a connection on a principal bundle. The formal definition of a connection is a decomposition of the tangent space TP_p of P at $p \in P$ into a *horizontal subspace* $H_p(P)$ and a *vertical subspace* $V_p(P)$. We require that $V_p(P)$ be the tangent space to the fibres and $H_p(P)$ to be a complementary subspace, *i.e.*, the direct sum should be the whole tangent space

$$TP_p = H_p(P) \oplus V_p(P). \tag{7.76}$$

The horizontal subspaces must also be invariant under the push-forward induced from the action on the fibres from the *right* of a fixed element of G. More formally, if $R[g] : P \to P$ acts to take $p \to pg$, *i.e.* by $R[g](x, g') = (x, g'g)$, we require

$$R[g]_* H_p(P) = H_{pg}(P). \tag{7.77}$$

Thus, we get to chose one horizontal subspace in each fibre, the rest being determined by the right-invariance condition.

Given a curve $x(t)$ in the base space we can, by solving the equation

$$\dot{g} + \frac{\partial x^\mu}{\partial t} \mathcal{A}_\mu(x) g = 0, \tag{7.78}$$

lift it to a curve $(x(t), g(t))$ in the total space, whose tangent is everywhere horizontal. This lifting operation corresponds to parallel transporting the initial value $g(0)$ along the curve $x(t)$ to get $g(t)$. The $\mathcal{A}_\mu = i\hat{\lambda}_a \mathcal{A}^a_\mu$ are a set of Lie-algebra-valued functions that are determined by our choice of horizontal subspace. They are defined so that the vector $(\delta x, -\mathcal{A}_\mu \delta x^\mu g)$ is horizontal for each small displacement δx^μ in the tangent space of M. Here $-\mathcal{A}_\mu \delta x^\mu g$ is to be understood as the displacement that takes $g \to (1 - \mathcal{A}_\mu \delta x^\mu) g$. Because we are multiplying $\mathcal{A}$ in from the *left*, the lifted curve can be slid rigidly up and down the fibres by the right action of any fixed group element. The right-invariance condition is therefore automatically satisfied.

The directional derivative along the lifted curve is

$$\dot{x}^\mu \mathcal{D}_\mu = \dot{x}^\mu \left(\left(\frac{\partial}{\partial x^\mu} \right)_g - \mathcal{A}^a_\mu R_a \right), \tag{7.79}$$

where R_a is a right-invariant vector field on G, *i.e.*, a differential operator on functions defined on the fibres. The $\mathcal{D}_\mu$ are a set of vector fields in TP. These

covariant derivatives span the horizontal subspace at each point $p \in P$, and have Lie brackets

$$[\mathcal{D}_\mu, \mathcal{D}_\nu] = -\mathcal{F}^a_{\mu\nu} R_a. \tag{7.80}$$

Here $\mathcal{F}_{\mu\nu}$, is given in terms of the structure constants appearing in the Lie brackets $[R_a, R_b] = f^c_{ab} R_c$ by

$$\mathcal{F}^c_{\mu\nu} = \partial_\mu \mathcal{A}^c_\nu - \partial_\nu \mathcal{A}^c_\mu - f^c_{ab} \mathcal{A}^a_\mu \mathcal{A}^b_\nu. \tag{7.81}$$

We can also write

$$\mathcal{F}_{\mu\nu} = \partial_\mu \mathcal{A}_\nu - \partial_\nu \mathcal{A}_\mu + [\mathcal{A}_\mu, \mathcal{A}_\nu]. \tag{7.82}$$

where $\mathcal{F}_{\mu\nu} = i\hat{\lambda}_a \mathcal{F}^a_{\mu\nu}$ and $[\hat{\lambda}_a, \hat{\lambda}_b] = i f^c_{ab} \hat{\lambda}_c$.

Because the Lie bracket of the $\mathcal{D}_\mu$ is a linear combination of the R_a, it lies entirely in the vertical subspace. Consequently, when $\mathcal{F}_{\mu\nu} \neq 0$, the $\mathcal{D}_\mu$ are not in involution, and Frobenius' theorem tells us that the horizontal subspaces cannot fit together to form the tangent spaces to a smooth foliation of P.

We make contact with the more familiar definitions of covariant derivatives by remembering that *right* invariant vector fields are derivatives that involve infinitesimal multiplication from the *left*. Their definition is

$$R_a \varphi_i(x, g) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\varphi_i(x, (1 + i\epsilon \hat{\lambda}_a) g) - \varphi_i(x, g) \right), \tag{7.83}$$

where $[\hat{\lambda}_a, \hat{\lambda}_b] = i f^c_{ab} \hat{\lambda}_c$.

Since $\varphi_i(x, g)$ is a section of the associated bundle, we know how it varies when we multiply group elements in on the right. We therefore write

$$(1 + i\epsilon \hat{\lambda}_a) g = g\, g^{-1} (1 + i\epsilon \hat{\lambda}_a) g, \tag{7.84}$$

and from this, (and writing g for $D(g)$ where it makes for compact notation) we find

$$\begin{aligned} R_a \varphi_i(x, g) &= \lim_{\epsilon \to 0} \left(D_{ij}(g^{-1}(1 - i\epsilon \hat{\lambda}_a) g) \varphi_j(x, g) - \varphi_i(x, g) \right) / \epsilon \\ &= -D_{ij}(g^{-1})(i\hat{\lambda}_a)_{jk} D_{kl}(g) \varphi_l(x, g) \\ &= -i(g^{-1} \hat{\lambda}_a g)_{ij} \varphi_j. \end{aligned} \tag{7.85}$$

Here $i(\hat{\lambda}_a)_{ij}$ is the matrix representing the Lie algebra generator $i\hat{\lambda}_a$ in the representation $g \mapsto D(g)$. Acting on sections, we therefore have

$$\mathcal{D}_\mu \varphi = (\partial_\mu \varphi)_g + (g^{-1} \mathcal{A}_\mu g) \varphi. \tag{7.86}$$

This still does not look too familiar because the derivatives with respect to x_μ are being taken at *fixed* g. We normally *fix a gauge* by making a choice of $g = \sigma(x)$ for each x_μ. The conventional wavefunction $\varphi(x)$ is then $\varphi(x, \sigma(x))$. We can use $\varphi(x, \sigma(x)) = \sigma^{-1}(x)\varphi(x, e)$, to obtain

$$\partial_\mu \varphi = (\partial_\mu \varphi)_\sigma + \left(\partial_\mu \sigma^{-1}\right) \sigma\varphi = (\partial_\mu \varphi)_\sigma - \left(\sigma^{-1}\partial_\mu \sigma\right) \varphi. \tag{7.87}$$

From this we get a derivative

$$\nabla_\mu \stackrel{\text{def}}{=} \partial_\mu + (\sigma^{-1}\mathcal{A}_\mu \sigma + \sigma^{-1}\partial_\mu \sigma) = \partial_\mu + A_\mu. \tag{7.88}$$

on functions $\varphi(x) \equiv \varphi(x, \sigma(x))$ defined (locally) on the base space M. This is the conventional covariant derivative, now containing gauge fields $A_\mu(x)$ that are gauge transformations of our g-independent $\mathcal{A}_\mu$. The derivative has been constructed so that

$$\nabla_\mu \varphi(x) = \mathcal{D}_\mu \varphi(x, g)|_{g=\sigma(x)}\,, \tag{7.89}$$

and has commutator

$$[\nabla_\mu, \nabla_\nu] = \sigma^{-1}\mathcal{F}_{\mu\nu}\sigma = F_{\mu\nu}. \tag{7.90}$$

Note the sign change *vis-a-vis* equation (7.80).

It is the curvature tensor $F_{\mu\nu}$ that we have met previously. Recall that it provides a Lie algebra valued two-form

$$F \equiv \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu = dA + A^2 \tag{7.91}$$

on the base space. The connection $A \equiv A_\mu dx^\mu$ is a one-form on the base space, and both F and A have been defined only in the region $U \subset M$ where the smooth gauge-choice section $\sigma(x)$ has been selected.

7.3.3 Monople harmonics

The total-space operations and definitions seem rather abstract. We demonstrate their power by solving the Schrödinger problem for a charged particle confined to a unit sphere surrounding a magnetic monopole. The conventional approach to this problem involves first selecting a gauge for vector the potential A, which, because of the monopole, is necessarily singular at a

Dirac string located somewhere on the sphere, and then delving into properties of Gegenbauer polynomials. Eventually we find the gauge-dependent wavefunction. By working with the total space, however, we can solve the problem in all gauges *at once*, and the problem becomes a simple exercise in Lie group geometry.

Recall that the SU(2) representation matrices $D^J_{mn}(\theta,\phi,\psi)$ form a complete orthonormal set of functions on the group manifold S^3. There will be a similar complete orthonormal set of representation matrices on the manifold of any compact Lie group G. Given a subgroup $H \in G$, we will use these matrices to construct bundles associated to a principal H-bundle that has G as its total space, and the coset space G/H as its base space. The fibres will be copies of H, and the projection π the usual projection $G \to G/H$.

The functions $D^J(g)$ are not in general functions on the coset space G/H as they depend on the choice of representative. Instead, because of the representation property, they vary with the choice of representative in a well-defined way,

$$D^J_{mn}(gh) = D^J_{mn'}(g)D^J_{n'n}(h). \tag{7.92}$$

Since we are dealing with compact groups, the representations can be taken to be unitary and

$$\begin{aligned} [D^J_{mn}(gh)]^* &= [D^J_{mn'}(g)]^*[D^J_{n'n}(h)]^* &(7.93)\\ &= D^J_{nn'}(h^{-1})[D^J_{mn'}(g)]^*. &(7.94)\end{aligned}$$

This is the correct variation under the right action of the group H for the set of functions $[D^J_{mn}(gh)]^*$ to be sections of a bundle associated with the principal fibre bundle $G \to G/H$. The representation $h \mapsto D(h)$ of H is not necessarily that defined by the label J because irreducible representations of G may be reducible under H; D depends on what representation of H the index n belongs to. If D is the identity representation, then the functions are functions on G/H in the ordinary sense. For $G = \mathrm{SU}(2)$ and H the $U(1)$ subgroup generated by J_3, the quotient space is just S^2, and projection is the Hopf map: $S^3 \to S^2$. The resulting bundle can be called the Hopf bundle. It is not a really new object however, because it is a generalization of the monopole bundle of the preceding section. Parameterizing SU(2) with Euler angles, so that

$$D^J_{mn}(\theta,\phi,\psi) = \langle J,m|e^{-i\phi J_3}e^{-i\theta J_2}e^{-i\psi J_3}|J,n\rangle, \tag{7.95}$$

shows that the Hopf map consists of simply forgetting about ψ, so

$$\text{Hopf} : [(\theta, \phi, \psi) \in S^3] \mapsto [(\theta, \phi) \in S^2]. \tag{7.96}$$

The bundle is twisted because S^3 is not a product $S^2 \times S^1$. Taking $n = 0$ gives us functions independent of ψ, and we obtain the well-known identification of the spherical harmonics with representation matrices

$$Y^L_m(\theta, \phi) = \sqrt{\frac{2L+1}{4\pi}} [D^{(L)}_{m0}(\theta, \phi, 0)]^*. \tag{7.97}$$

For $n = \Lambda \neq 0$ we get sections of a bundle with Chern number 2Λ. These sections are the *monopole harmonics*

$$\mathcal{Y}^J_{m;\Lambda}(\theta, \phi, \psi) = \sqrt{\frac{2J+1}{4\pi}} [D^J_{m\Lambda}(\theta, \phi, \psi)]^* \tag{7.98}$$

for a monopole of flux $\int eB \, d(\text{Area}) = 4\pi\Lambda$. The integrality of the Chern number tells us that the flux $4\pi\Lambda$ must be an integer multiple of 2π. This gives us a geometric reason for why the eigenvalues m of J_3 can only be an integer or half integer.

The monopole harmonics have a non-trivial $\propto e^{i\psi\Lambda}$ dependence on the choice we make for ψ at each point on S^2, and we cannot make a globally smooth choice; we always encounter a point where there is a singularity. These sections of the twisted bundle have to be constructed in patches and glued together transition functions.

We now show that the monopole harmonics are eigenfunctions of the Schrödinger operator, $-\nabla^2$, containing the gauge field connection, just as the spherical harmonics are eigenfunctions of the Laplacian on the sphere. This is a simple geometrical exercise. Because they are irreducible representations, the $D^J(g)$ are automatically eigenfunctions of the quadratic Casimir operator

$$(J_1^2 + J_2^2 + J_3^2) D^J(g) = J(J+1) D^J(g). \tag{7.99}$$

The J_i can be either right or left-invariant vector fields on G; the quadratic Casimir is the same second-order differential operator in either case, and it is a good guess that it is proportional to the Laplacian on the group manifold. Taking a locally geodesic co-ordinate system (in which the connection vanishes) confirms this: $J^2 = -\nabla^2$ on the three-sphere. The operator in (7.99) is not the Laplacian we want, however. What we need is the ∇^2 on

the two-sphere $S^2 = G/H$, including the the connection. This ∇^2 operator differs from the one on the total space since it must contain only differential operators lying in the horizontal subspaces. There is a natural notion of orthogonality in the Lie group, deriving from the Killing form, and it is natural to choose the horizontal subspaces to be orthogonal to the fibres of G/H. Since multiplication on the right by the subgroup generated by J_3 moves one up and down the fibres, the orthogonal displacements are obtained by multiplication on the right by infinitesimal elements made by exponentiating J_1 and J_2. The desired ∇^2 is thus made out of the left-invariant vector fields (which act by multiplication on the right), J_1 and J_2 only. The wave operator must be

$$-\nabla^2 = J_1^2 + J_2^2 = J^2 - J_3^2. \tag{7.100}$$

Applying this to the $\mathcal{Y}^J_{m;\Lambda}$ we see that they are eigenfunctions of $-\nabla^2$ on S^2 with eigenvalues $J(J+1) - \Lambda^2$. The Laplace eigenvalues for our flux $4\pi\Lambda$ monopole problem are therefore

$$E_{J,m} = (J(J+1) - \Lambda^2), \quad J \geq |\Lambda|, \quad -J \leq m \leq J. \tag{7.101}$$

The utility of the monopole Harmonics is not restricted to exotic monopole physics. They occur in molecular and nuclear physics as the wavefunctions for the rotational degrees of freedom of diatomic molecules and uniaxially deformed nuclei that possess angular momentum Λ about their axis of symmetry.[3]

Exercise 7.3: Compare these energy levels for a particle on a sphere with those of the Landau level problem on the plane. Show that for any fixed flux the low-lying energies remain close to $E = (eB/m_{\text{particle}})(n + 1/2)$, n zero or a positive integer, but their degeneracy is is equal to the number of flux units penetrating the sphere *plus one*.

7.3.4 Bundle connection and curvature forms

Recall that in section 7.3.2 we introduced the Lie-Algebra-valued functions $\mathcal{A}_\mu(x)$. We now use these functions to introduce the *bundle connection form* $\mathbb{A}$ that lives in T^*P. We set

$$\mathcal{A} = \mathcal{A}_\mu \, dx^\mu \tag{7.102}$$

[3]This is explained, with chararacteristic terseness, in a footnote on page 317 of Landau and Lifshitz' *Quantum Mechanics* (Third Edition).

and

$$\mathbb{A} \stackrel{\text{def}}{=} g^{-1}\left(\mathcal{A} + \delta g\, g^{-1}\right) g. \tag{7.103}$$

In these definitions, x and g are the local co-ordinates in which points in the total space are labelled as (x, g), and d acts on functions of x, and the "δ" is used to denote the exterior derivative acting on the fibre.[4] We have, then, that $\delta x^\mu = 0$ and $dg = 0$. The combinations $\delta g\, g^{-1}$ and $g^{-1}\delta g$ are respectively the right- and left-invariant Maurer-Cartan form on the group.

The complete exterior derivative in the total space requires us to differentiate both with respect to g and with respect to x, and is given by $d_{\text{tot}} = d+\delta$. Because d^2 , δ^2 and $(d+\delta)^2 = d^2 + \delta^2 + d\delta + \delta d$ are all zero, we must have

$$\delta d + d\delta = 0. \tag{7.104}$$

We now define the *bundle curvature form* in terms of $\mathbb{A}$ to be

$$\mathbb{F} \stackrel{\text{def}}{=} d_{\text{tot}}\mathbb{A} + \mathbb{A}^2. \tag{7.105}$$

To compute $\mathbb{F}$ in terms of $\mathcal{A}(x)$ and g we need the ingredients

$$d\mathbb{A} = g^{-1}(d\mathcal{A})g, \tag{7.106}$$

and

$$\delta\mathbb{A} = -(g^{-1}\delta g)\mathbb{A} - \mathbb{A}(g^{-1}\delta g) - (g^{-1}\delta g)^2. \tag{7.107}$$

We find that

$$\begin{aligned}\mathbb{F} = (d+\delta)\mathbb{A} + \mathbb{A}^2 &= g^{-1}\left(d\mathcal{A} + \mathcal{A}^2\right) g \\ &= g^{-1}\mathcal{F} g,\end{aligned} \tag{7.108}$$

where

$$\mathcal{F} = \frac{1}{2}\mathcal{F}_{\mu\nu} dx^\mu dx^\nu, \tag{7.109}$$

and

$$\mathcal{F}_{\mu\nu} = \partial_\mu \mathcal{A}_\nu - \partial_\nu \mathcal{A}_\mu + [\mathcal{A}_\mu, \mathcal{A}_\nu]. \tag{7.110}$$

Although we have defined the connection form $\mathbb{A}$ in terms of the local bundle co-ordinates (x, g), it is, in fact, an intrinsic quantity, *i.e.* it is has a global existence independent of the choice of these co-ordinates. $\mathbb{A}$ has been constructed so that

[4] It is *not* therefore to be confused with the Hodge $\delta = d^\dagger$ operator.

- A vector is annihilated by $\mathbb{A}$ if and only if it is horizontal. In particular $\mathbb{A}(\mathcal{D}_\mu) = 0$ for all covariant derivatives $\mathcal{D}_\mu$.
- The connection form is constant on *left*-invariant vector fields on the fibres. In particular $\mathbb{A}(L_a) = i\hat{\lambda}_a$.

Between them, the globally defined fields $\mathcal{D}_\mu \in H_p(P)$ and $L_a \in V_p(P)$ span the tangent space TP_p. Consequently the two properties listed above tell us how to evaluate $\mathbb{A}$ on any vector, and so define it uniquely and globally.

From the globally defined and gauge invariant $\mathbb{A}$ and its associated curvature $\mathbb{F}$, and for any local gauge-choice section $\sigma : (U \subset M) \to P$, we can recover the gauge-dependent base-space forms A and F as the pull-backs

$$A = \sigma^* \mathbb{A}, \quad F = \sigma^* \mathbb{F}, \tag{7.111}$$

to $U \subset M$ of the total-space forms. The resulting forms are

$$A = \left(\sigma^{-1}\mathcal{A}_\mu \sigma + \sigma^{-1}\partial_\mu \sigma\right) dx^\mu, \quad F = \frac{1}{2}\left(\sigma^{-1}\mathcal{F}_{\mu\nu}\sigma\right)\, dx^\mu dx^\nu, \tag{7.112}$$

and coincide with the equations connecting A_μ with $\mathcal{A}_\mu$ and $F_{\mu\nu}$ with $\mathcal{F}_{\mu\nu}$ that we obtained in section 7.3.2. We should take care to note that the dx^μ that appear in A and F are differential forms on M, while the dx^μ that appear in $\mathcal{A}$ and $\mathcal{F}$ are differential forms on P. Now the projection π is a left inverse of the gauge-choice section σ, *i.e.* $\pi \circ \sigma =$ identity. The associated pull-backs are also inverses, but with the order reversed: $\sigma^* \circ \pi^* =$ identity. These maps relate the two sets of "dx^μ" by

$$dx^\mu|_M = \sigma^*\left(dx^\mu|_P\right), \quad \text{or} \quad dx^\mu|_P = \pi^*\left(dx^\mu|_M\right). \tag{7.113}$$

We now explain the advantage of knowing the total space connection and curvature forms. Consider the Chern character $\propto \operatorname{tr} F^2$ on the base-space M. We can use the bundle projection π to pull this form back to total space. From

$$\mathbb{F}_{\mu\nu} = (g\sigma^{-1})^{-1} F_{\mu\nu} (g\sigma^{-1}), \tag{7.114}$$

we find that

$$\pi^*\left(\operatorname{tr} F^2\right) = \operatorname{tr} \mathbb{F}^2. \tag{7.115}$$

Now $\mathbb{A}$, $\mathbb{F}$ and d_{tot} have the same calculus properties as A , F and d. The manipulations that give

$$\operatorname{tr} F^2 = d \operatorname{tr}\left(A\, dA + \frac{2}{3} A^3\right)$$

also show, therefore, that

$$\operatorname{tr}\mathbb{F}^2 = d_{\text{tot}}\operatorname{tr}\left(\mathbb{A}\,d_{\text{tot}}\mathbb{A} + \frac{2}{3}\mathbb{A}^3\right). \tag{7.116}$$

There is a big difference in the significance of the computation, however. The bundle connection $\mathbb{A}$ is globally defined. Consequently, the form

$$\omega_3(\mathbb{A}) \equiv \operatorname{tr}\left(\mathbb{A}\,d_{\text{tot}}\mathbb{A} + \frac{2}{3}\mathbb{A}^3\right) \tag{7.117}$$

is also globally defined. The pull-back to the total space of the Chern character is d_{tot} exact! This miracle works for all characteristic classes: on the base-space they are exact only when the bundle is trivial; on the total space they are always exact.

We have seen this phonomenon before, for example in exercise 6.7. The area form $d[\text{Area}] = \sin\theta\, d\theta d\phi$ is closed but not exact on S^2. When pulled back to S^3 by the Hopf map, the area form becomes exact:

$$\text{Hopf}^* d[\text{Area}] = \sin\theta\, d\theta d\phi = d(-\cos\theta d\phi + d\psi). \tag{7.118}$$

7.3.5 Characteristic classes as obstructions

The generalized Gauss-Bonnet theorem states that, for a compact orientable even-dimensional manifold M, the integral of the Euler class over M is equal to the Euler character $\chi(M)$. Shiing-Shen Chern used the exactness of the pull-back of the Euler class to give an elegant intrinsic proof[5] of this theorem. Chern showed that the integral of the Euler class over M was equal to the sum of the Poincare-Hopf indices of any tangent vector field on M, a sum we independently know to equal the Euler character $\chi(M)$. We illustrate his strategy by showing how a non-zero $\text{ch}_2(F)$ provides a similar index sum for the singularities of any section of an SU(2)-bundle over a four-dimensional base space. This result provides an interpretation of characteristic classes as obstructions to the existence of global sections.

Let $\sigma : M \to P$ be a section of an SU(2) principal bundle P over a four-dimensional compact orientable manifold M without boundary. For any SU(n) group we have $\text{ch}_1(F) \equiv 0$, but

$$\int_M \text{ch}_2(F) = -\frac{1}{8\pi^2}\int_M \operatorname{tr}(F^2) = n, \tag{7.119}$$

[5] S-J. Chern, *Ann. Math.* **47** (1946) 85-121. This paper is a readable classic.

can be non-zero.

The section σ will, in general, have points x_i where it becomes singular. We punch infinitesimal holes in M surrounding the singular points. The manifold $M' = (M \setminus \text{holes})$ will have as its boundary $\partial M'$ a disjoint union of small three-spheres. We denote by Σ the image of M' under the map $\sigma : M' \to P$. This Σ will be a submanifold of P, whose boundary will be equal in homology to a linear combination of the boundary components of M' with integer coefficients. We show that the Chern number n is equal to the sum of these coefficients.

We begin by using the projection π to pull back $\mathrm{ch}_2(F)$, to the bundle, where we know that

$$\pi^* \mathrm{ch}_2(F) = -\frac{1}{8\pi^2} d_{\mathrm{tot}}\, \omega_3(\mathbb{A}). \tag{7.120}$$

Now we can decompose $\omega_3(\mathbb{A})$ into terms of different bi-degree, *i.e.* into terms that are p-forms in d and q-forms in δ.

$$\omega_3(\mathbb{A}) = \omega_3^0 + \omega_2^1 + \omega_1^2 + \omega_0^3. \tag{7.121}$$

Here the superscript counts the form-degree in δ, and the subscript the form-degree in d. The only term we need to know explicitly is ω_0^3. This comes from the $g^{-1}\delta g$ part of $\mathbb{A}$, and is

$$\begin{aligned} \omega_0^3 &= \mathrm{tr}\left((g^{-1}\delta g)\, \delta(g^{-1}\delta g) + \frac{2}{3}(g^{-1}\delta g)^3 \right) \\ &= \mathrm{tr}\left(-(g^{-1}\delta g)^3 + \frac{2}{3}(g^{-1}\delta g)^3 \right) \\ &= -\frac{1}{3}(g^{-1}\delta g)^3. \end{aligned} \tag{7.122}$$

We next use the map $\sigma : M' \to P$ to pull the right-hand side of (7.120) back from P to M'. We recall that acting on forms on M' we have $\sigma^* \circ \pi^* =$ identity. Thus

$$\begin{aligned} \int_M \mathrm{ch}_2(F) = \int_{M'} \mathrm{ch}_2(F) &= \int_{M'} \sigma^* \circ \pi^* \mathrm{ch}_2(F) \\ &= -\frac{1}{8\pi^2} \int_{M'} \sigma^* d_{\mathrm{tot}}\, \omega_3(\mathbb{A}) \\ &= -\frac{1}{8\pi^2} \int_{\Sigma} d_{\mathrm{tot}}\, \omega_3(\mathbb{A}) \end{aligned}$$

$$\begin{aligned} &= -\frac{1}{8\pi^2}\int_{\partial\Sigma}\omega_3(\mathbb{A}) \\ &= \frac{1}{24\pi^2}\int_{\partial\Sigma}(g^{-1}\delta g)^3. \end{aligned} \tag{7.123}$$

At the first step we have observed that the omitted spheres make a negligeable contribution to the integral over M, and at the last step we have used the fact that the boundary of Σ, has significant extent only along the fibres, so all contributions to the integral over $\partial\Sigma$ come from the purely vertical component of $\omega_3(\mathbb{A})$, which is $\omega_0^3 = -\frac{1}{3}(g^{-1}dg)$.

We know (see exercise 6.8) that for maps $g \mapsto U \in \mathrm{SU}(2)$ we have

$$\int \mathrm{tr}\,(g^{-1}dg)^3 = 24\pi^2 \times \text{winding number}$$

We conclude that

$$\int_M \mathrm{ch}_2(F) = \frac{1}{24\pi^2}\int_{\partial\Sigma}(g^{-1}\delta g)^3 = \sum_{\text{singularities } x_i} N_i \tag{7.124}$$

where N_i is the Brouwer degree of the map $\sigma : S^3 \to \mathrm{SU}(2) \cong S^3$ on the small sphere surrounding x_i.

It turns out that for any $\mathrm{SU}(n)$ the integral of $\mathrm{tr}\,(g^{-1}\delta g)^3$ is $24\pi^2$ times an integer winding number of g about homology spheres. The second Chern number of a $\mathrm{SU}(n)$-bundle is therefore also equal to the sum of the winding-number indices of the section about its singularities. Chern's strategy can be used to relate other characteristic classes to obstructions to the existence of global sections of appropriate bundles.

7.3.6 Stora-Zumino descent equations

In the previous sections we met the forms

$$\mathbb{A} = g^{-1}\mathcal{A}g + g^{-1}\delta g \tag{7.125}$$

and

$$A = \sigma^{-1}\mathcal{A}\sigma + \sigma^{-1}d\sigma. \tag{7.126}$$

The group element g labeled points on the fibres and was independent x, while $\sigma(x)$ was the gauge-choice section of the bundle and depended on x.

The two quantities $\mathbb{A}$ and A look similar, but are not identical. A third superficially similar but distinct object is met with in the BRST (Becchi-Rouet-Stora-Tyutin) approach to quantizing gauge theories, and also in the geometric theory of anomalies. We describe it here to alert the reader to the potential for confusion.

Rather than attempting to define this new differential form rigorously, we will first explain how to calculate with it, and only then indicate what it is. We begin by considering a fixed connection form A on M, and its orbit under the action of the group $\mathcal{G}$ of gauge transformations. This elements of this infinite dimensional group are maps $g : M \to G$ equipped with pointwise product $g_1 g_2(x) = g_1(x) g_2(x)$. This $g(x)$ is neither the fibre co-ordinate g, nor the gauge choice section $\sigma(x)$. The gauge transformation $g(x)$ acts on A to give A^g where

$$A^g = g^{-1} A g + g^{-1} dg. \tag{7.127}$$

We now introduce an object

$$v(x) = g^{-1} \delta g, \tag{7.128}$$

and consider

$$\mathfrak{A} = A^g + v = g^{-1} A g + g^{-1} dg + g^{-1} \delta g. \tag{7.129}$$

This 1-form appears to be a hybrid of the earlier quantities, but we will see that it has to be considered as something new. The essential difference from what has gone before is that we want v to behave like $g^{-1}\delta g$, in that $\delta v = -v^2$, and yet to depend on x. In particular we want δ to behave as an exterior derivative that implements an infinitesimal gauge transformation that takes $g \to g + \delta g$. Thus,

$$\begin{aligned} \delta(g^{-1} dg) &= -(g^{-1}\delta g)(g^{-1} dg) + g^{-1} \delta dg \\ &= -(g^{-1}\delta g)(g^{-1} dg) - (g^{-1} dg)(g^{-1}\delta g) + (g^{-1} dg)(g^{-1}\delta g) - g^{-1} d\delta g \\ &= -v(g^{-1} dg) - (g^{-1} dg) v - dv, \end{aligned} \tag{7.130}$$

and hence

$$\delta A^g = -v A^g - A^g v - dv. \tag{7.131}$$

Previously $g^{-1} dg \equiv 0$, and so there was no "dv" in δ(gauge field).

We can define a curvature associated with $\mathfrak{A}$

$$\mathfrak{F} \stackrel{\text{def}}{=} d_{\text{tot}} \mathfrak{A} + \mathfrak{A}^2, \tag{7.132}$$

and compute

$$\begin{aligned}\mathfrak{F} &= (d+\delta)(A^g+v)+(A^g+v)^2\\ &= dA^g+dv+\delta A^g+\delta v+(A^g)^2+A^gv+vA^g+v^2\\ &= dA^g+(A^g)^2\\ &= g^{-1}Fg,\end{aligned}\tag{7.133}$$

Stora calls (7.133) the *Russian formula*.

Because $\mathfrak{F}$ is yet another gauge transform of F, we have

$$\operatorname{tr}F^2=\operatorname{tr}\mathfrak{F}^2=(d+\delta)\operatorname{tr}\left(\mathfrak{A}(d+\delta)\mathfrak{A}+\frac{2}{3}\mathfrak{A}^3\right)\tag{7.134}$$

and can decompose the right-hand side into terms that are simultaneously p-foms in d and q-forms in δ.

The left hand side, $\operatorname{tr}\mathfrak{F}^2=\operatorname{tr}F^2$, of (7.134) is independent of v. The right hand side of (7.134) contains $\omega_3(\mathfrak{A})$ which we expand as

$$\omega_3(A^g+v)=\omega_3^0(A^g)+\omega_2^1(v,A^g)+\omega_1^2(v,A^g)+\omega_0^3(v).\tag{7.135}$$

As in the previous section, the superscript counts the form-degree in δ, and the subscript the form-degree in d. Explicit computation shows that

$$\begin{aligned}\omega_3^0(A^g) &= \operatorname{tr}\left(A^g\,dA^g+\tfrac{2}{3}(A^g)^3\right),\\ \omega_2^1(v,A^g) &= \operatorname{tr}\left(v\,dA^g\right),\\ \omega_1^2(v,A^g) &= -\operatorname{tr}\left(A^gv^2\right),\\ \omega_0^3(v) &= -\tfrac{1}{3}v^3\end{aligned}\tag{7.136}$$

For example,

$$\omega_0^3(v)=\operatorname{tr}\left(v\,\delta v+\frac{2}{3}v^3\right)=\operatorname{tr}\left(v(-v^2)+\frac{2}{3}v^3\right)=-\frac{1}{3}v^3.\tag{7.137}$$

With this decomposition, (7.116) falls apart into the chain of *descent equations*

$$\begin{aligned}\operatorname{tr}F^2 &= d\omega_3^0(A^g),\\ \delta\omega_3^0(A^g) &= -d\,\omega_2^1(v,A^g),\\ \delta\omega_2^1(v,A^g) &= -d\,\omega_1^2(v,A^g),\\ \delta\omega_1^2(v,A^g) &= -d\,\omega_0^3(v),\\ \delta\omega_0^3(v) &= 0.\end{aligned}\tag{7.138}$$

Let us verify, for example, the penultimate equation $\delta\omega_1^2(v, A^g) = -d\,\omega_0^3(v)$. The left-hand side is

$$-\delta\,\mathrm{tr}\,(A^g v^2) = -\mathrm{tr}\,(-Av^3 - vA^g v^2 - dv\,v^2) = \mathrm{tr}\,(dv\,v^2), \tag{7.139}$$

the terms involving A^g having cancelled *via* the cyclic property of the trace and the fact that A^g anticommutes with v. The right-hand side is

$$-d\left(-\tfrac{1}{3}\mathrm{tr}\,v^3\right) = \mathrm{tr}\,(dv\,v^2) \tag{7.140}$$

as required.

The descent equations were introduced by Raymond Stora and Bruno Zumino as a tool for obtaining and systematizing information about *anomalies* in the quantum field theory of fermions interacting with the gauge field A^g. The $\omega_p^q(v, A^g)$ are p-forms in the dx^μ, and before use they are integrated over p-cycles in M. This process is understood to produce local functionals of A^g that remain q-forms in δg. For example, in $2n$ space-time dimensions, the integral

$$I[g^{-1}\delta g, A^g] = \int_M \omega_{2n}^1(g^{-1}\delta g, A^g) \tag{7.141}$$

has the properties required for it to be a candidate for the anomalous variation $\delta S[A^g]$ of the fermion effective action due to an infinitesimal gauge transformation $g \to g + \delta g$. In particular, when $\partial M = \emptyset$, we have

$$\delta I[g^{-1}\delta g, A^g] = \int_M \delta\omega_{2n}^1(v, A^g) = -\int_M d\omega_{2n-1}^2(v, A^g) = 0. \tag{7.142}$$

This is the *Wess-Zumino consistency condition* that $\delta(\delta S)$ must obey as a consequence of $\delta^2 = 0$.

In addition to producing a convenient solution of the Wess-Zumino condition, the descent equations provide a compact derivation of the gauge transformation properties of useful differential forms. We will not seek to explain further the physical meaning of these forms, leaving this to a field theory course.

The similarity between $\mathbb{A}$ and $\mathfrak{A}$ lead various authors to attempt to identify them, and in particular to identify $v(x)$ with the $g^{-1}\delta g$ Maurer-cartan form appearing in $\mathbb{A}$. However the physical meaning of expressions such as $d(g^{-1}\delta g)$ precludes such a simple interpretation. In evaluating $dv \sim d(g^{-1}\delta g)$ on a vector field $\xi^a(x)L_a$ representing an infinitesimal gauge transformation,

we are to first to insert the field into $v \sim g^{-1}\delta g$ to obtain the x dependent Lie algebra element $i\xi^a(x)\hat{\lambda}_a$, and only then to take the exterior derivative to obtain $i\hat{\lambda}_a\partial_\mu\xi^a\,dx^\mu$. The result therefore involves derivatives of the components $\xi^a(x)$. The evaluation of an ordinary differential form on a vector field never produces derivatives of the vector components.

To understand what the Stora-Zumino forms are, imagine that we equip a two dimensional fibre bundle $E = M \times F$ with base-space co-ordinate x and fibre co-ordinate y. A $p = 1$, $q = 1$ form on E will then be $F = f(x, y)\,dx\,\delta y$ for some function $f(x, y)$. There is only one object δy, and there is no meaning to integrating F over x to leave a 1-form in δy on E. The space of forms introduced by Stora and Zumino, on the other hand, would contain elements such as

$$J = \int_M j(x, y)\,dx\,\delta y_x \tag{7.143}$$

where there is a distinct δy_x for each $x \in M$. If we take, for example, $j(x, y) = \delta'(x - a)$. we evaluate J on the vector field $Y(x, y)\partial_y$ as

$$J[Y(x, y)\partial_y] = \int \delta'(x - a)Y(x, y)\,dx = -Y'(a, y). \tag{7.144}$$

The conclusion is that that the 1-form form field $v(x) \sim g^{-1}\delta g$ must be considered as the left-invariant Maurer-Cartan form on the infinite dimensional Lie group $\mathcal{G}$, rather than a Maurer-Cartan form on the finite dimensional Lie group G. The $\int_M \omega^q_{2n}(v, A^g)$ are therefore elements of the cohomology group $H^q(A^{\mathcal{G}})$ of the $\mathcal{G}$ orbit of A, a rather complicated object. For a thorough discussion see: J. A. de Azcárraga, J. M. Izquierdo, *Lie groups, Lie Algebras, Cohomology and some Applications in Physics*, published by Cambridge University Press.

Chapter 8

Complex Analysis I

Although this chapter is called complex *analysis*, we will try to develop the subject as complex *calculus* — meaning that we shall follow the calculus course tradition of telling you how to do things, and explaining why theorems are true, with arguments that would not pass for rigorous proofs in a course on real analysis. We try, however, to tell no lies.

This chapter will focus on the basic ideas that need to be understood before we apply complex methods to evaluating integrals, analysing data, and solving differential equations.

8.1 Cauchy-Riemann equations

We focus on functions, $f(z)$, of a single complex variable, z, where $z = x+iy$. We can think of these as being complex valued functions of two real variables, x and y. For example

$$\begin{aligned} f(z) = \sin z \equiv \sin(x+iy) &= \sin x \cos iy + \cos x \sin iy \\ &= \sin x \cosh y + i \cos x \sinh y. \end{aligned} \tag{8.1}$$

Here, we have used

$$\begin{aligned} \sin x &= \frac{1}{2i}\left(e^{ix} - e^{-ix}\right), \quad \sinh x = \frac{1}{2}\left(e^{x} - e^{-x}\right), \\ \cos x &= \frac{1}{2}\left(e^{ix} + e^{-ix}\right), \quad \cosh x = \frac{1}{2}\left(e^{x} + e^{-x}\right), \end{aligned}$$

to make the connection between the circular and hyperbolic functions. We shall often write $f(z) = u + iv$, where u and v are real functions of x and y. In the present example, $u = \sin x \cosh y$ and $v = \cos x \sinh y$.

If all four partial derivatives

$$\frac{\partial u}{\partial x}, \quad \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x}, \quad \frac{\partial u}{\partial y}, \tag{8.2}$$

exist and are continuous then $f = u + iv$ is differentiable as a complex-valued function of two real variables. This means that we can approximate the variation in f as

$$\delta f = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y + \cdots, \tag{8.3}$$

where the dots represent a remainder that goes to zero faster than linearly as δx, δy go to zero. We now regroup the terms, setting $\delta z = \delta x + i\delta y$, $\delta \overline{z} = \delta x - i\delta y$, so that

$$\delta f = \frac{\partial f}{\partial z}\delta z + \frac{\partial f}{\partial \overline{z}}\delta \overline{z} + \cdots, \tag{8.4}$$

where we have defined

$$\begin{aligned} \frac{\partial f}{\partial z} &\equiv \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \\ \frac{\partial f}{\partial \overline{z}} &\equiv \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right). \end{aligned} \tag{8.5}$$

Now our function $f(z)$ does not depend on $\overline{z}$, and so it must satisfy

$$\frac{\partial f}{\partial \overline{z}} = 0. \tag{8.6}$$

Thus, with $f = u + iv$,

$$\frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)(u + iv) = 0 \tag{8.7}$$

i.e.

$$\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right) + i\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) = 0. \tag{8.8}$$

Since the vanishing of a complex number requires the real and imaginary parts to be separately zero, this implies that

$$\begin{aligned} \frac{\partial u}{\partial x} &= +\frac{\partial v}{\partial y}, \\ \frac{\partial v}{\partial x} &= -\frac{\partial u}{\partial y}. \end{aligned} \tag{8.9}$$

These two relations between u and v are known as the *Cauchy-Riemann equations*, although they were probably discovered by Gauss. If our continuous partial derivatives satisfy the Cauchy-Riemann equations at $z_0 = x_0 + iy_0$ then we say that the function is *complex differentiable* (or just differentiable) at that point. By taking $\delta z = z - z_0$, we have

$$\delta f \equiv f(z) - f(z_0) = \frac{\partial f}{\partial z}(z - z_0) + \cdots, \tag{8.10}$$

where the remainder, represented by the dots, tends to zero faster than $|z - z_0|$ as $z \to z_0$. This validity of this linear approximation to the variation in $f(z)$ is equivalent to the statement that the ratio

$$\frac{f(z) - f(z_0)}{z - z_0} \tag{8.11}$$

tends to a definite limit as $z \to z_0$ from any direction. It is the direction-independence of this limit that provides a proper meaning to the phrase "does not depend on $\overline{z}$." Since we are not allowing dependence on $\bar{z}$, it is natural to drop the partial derivative signs and write the limit as an ordinary derivative

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \frac{df}{dz}. \tag{8.12}$$

We will also use Newton's fluxion notation

$$\frac{df}{dz} \equiv f'(z). \tag{8.13}$$

The complex derivative obeys exactly the same calculus rules as ordinary real derivatives:

$$\begin{aligned} \frac{d}{dz} z^n &= n z^{n-1}, \\ \frac{d}{dz} \sin z &= \cos z, \\ \frac{d}{dz}(fg) &= \frac{df}{dz} g + f \frac{dg}{dz}, \quad \textit{etc.} \end{aligned} \tag{8.14}$$

If the function is differentiable at all points in an arcwise-connected[1] open set, or *domain*, D, the function is said to be *analytic* there. The words *regular* or *holomorphic* are also used.

8.1.1 Conjugate pairs

The functions u and v comprising the real and imaginary parts of an analytic function are said to form a pair of *harmonic conjugate functions*. Such pairs have many properties that are useful for solving physical problems.

From the Cauchy-Riemann equations we deduce that

$$\begin{aligned} \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)u &= 0, \\ \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)v &= 0. \end{aligned} \tag{8.15}$$

and so both the real and imaginary parts of $f(z)$ are automatically *harmonic* functions of x, y.

Further, from the Cauchy-Riemann conditions, we deduce that

$$\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}=0. \tag{8.16}$$

This means that $\nabla u \cdot \nabla v = 0$. We conclude that, provided that neither of these gradients vanishes, the pair of curves $u = const.$ and $v = const.$ intersect at right angles. If we regard u as the potential ϕ solving some electrostatics problem $\nabla^2\phi = 0$, then the curves $v = const.$ are the associated field lines.

Another application is to fluid mechanics. If $\mathbf{v}$ is the velocity field of an irrotational ($\nabla \times \mathbf{v} = \mathbf{0}$) flow, then we can (perhaps only locally) write the flow field as a gradient

$$\begin{aligned} v_x &= \partial_x\phi, \\ v_y &= \partial_y\phi, \end{aligned} \tag{8.17}$$

where ϕ is a *velocity potential*. If the flow is incompressible ($\nabla \cdot \mathbf{v} = 0$), then we can (locally) write it as a curl

$$\begin{aligned} v_x &= \partial_y\chi, \\ v_y &= -\partial_x\chi, \end{aligned} \tag{8.18}$$

[1] *Arcwise connected* means that any two points in D can be joined by a continuous path that lies wholely within D.

where χ is a *stream function*. The curves $\chi = const.$ are the flow streamlines. If the flow is both irrotational and incompressible, then we may use either ϕ or χ to represent the flow, and, since the two representations must agree, we have

$$\begin{aligned} \partial_x \phi &= +\partial_y \chi, \\ \partial_y \phi &= -\partial_x \chi. \end{aligned} \tag{8.19}$$

Thus ϕ and χ are harmonic conjugates, and so the complex combination $\Phi = \phi + i\chi$ is an analytic function called the *complex stream function*.

A conjugate v exists (at least locally) for any harmonic function u. To see why, assume first that we have a (u, v) pair obeying the Cauchy-Riemann equations. Then we can write

$$\begin{aligned} dv &= \frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy \\ &= -\frac{\partial u}{\partial y} dx + \frac{\partial u}{\partial x} dy. \end{aligned} \tag{8.20}$$

This observation suggests that if we are given a harmonic function u in some simply connected domain D, we can *define* a v by setting

$$v(z) = \int_{z_0}^{z} \left(-\frac{\partial u}{\partial y} dx + \frac{\partial u}{\partial x} dy \right) + v(z_0), \tag{8.21}$$

for some real constant $v(z_0)$ and point z_0. The integral does not depend on choice of path from z_0 to z, and so $v(z)$ is well defined. The path independence comes about because the curl

$$\frac{\partial}{\partial y} \left(-\frac{\partial u}{\partial y} \right) - \frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} \right) = -\nabla^2 u \tag{8.22}$$

vanishes, and because in a simply connected domain all paths connecting the same endpoints are homologous.

We now verify that this candidate $v(z)$ satisfies the Cauchy-Riemann realtions. The path independence, allows us to make our final approach to $z = x + iy$ along a straight line segment lying on either the x or y axis. If we approach along the x axis, we have

$$v(z) = \int^{x} \left(-\frac{\partial u}{\partial y} \right) dx' + \text{ rest of integral}, \tag{8.23}$$

and may use

$$\frac{d}{dx}\int^x f(x',y)\,dx' = f(x,y) \tag{8.24}$$

to see that

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{8.25}$$

at (x,y). If, instead, we approach along the y axis, we may similarly compute

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}. \tag{8.26}$$

Thus $v(z)$ does indeed obey the Cauchy-Riemann equations.

Because of the utility the harmonic conjugate it is worth giving a practical recipe for finding it, and so obtaining $f(z)$ when given only its real part $u(x,y)$. The method we give below is one we learned from John d'Angelo. It is more efficient than those given in most textbooks. We first observe that if f is a function of z only, then $\overline{f(z)}$ depends only on $\overline{z}$. We can therefore define a function $\overline{f}$ of $\overline{z}$ by setting $\overline{f(z)} = \overline{f}(\overline{z})$. Now

$$\frac{1}{2}\left(f(z) + \overline{f(z)}\right) = u(x,y). \tag{8.27}$$

Set

$$x = \frac{1}{2}(z+\overline{z}), \quad y = \frac{1}{2i}(z-\overline{z}), \tag{8.28}$$

so

$$u\left(\frac{1}{2}(z+\overline{z}), \frac{1}{2i}(z-\overline{z})\right) = \frac{1}{2}\left(f(z) + \overline{f}(\overline{z})\right). \tag{8.29}$$

Now set $\overline{z} = 0$, while keeping z fixed! Thus

$$f(z) + \overline{f}(0) = 2u\left(\frac{z}{2}, \frac{z}{2i}\right). \tag{8.30}$$

The function f is not completely determined of course, because we can always add a constant to v, and so we have the result

$$f(z) = 2u\left(\frac{z}{2}, \frac{z}{2i}\right) + iC, \quad C \in \mathbb{R}. \tag{8.31}$$

For example, let $u = x^2 - y^2$. We find

$$f(z) + \overline{f}(0) = 2\left(\frac{z}{2}\right)^2 - 2\left(\frac{z}{2i}\right)^2 = z^2, \tag{8.32}$$

or

$$f(z) = z^2 + iC, \quad C \in \mathbb{R}. \tag{8.33}$$

The business of setting setting $\overline{z} = 0$, while keeping z fixed, may feel like a dirty trick, but it can be justified by the (as yet to be proved) fact that f has a convergent expansion as a power series in $z = x + iy$. In this expansion it is meaningful to let x and y themselves be complex, and so allow z and $\overline{z}$ to become two independent complex variables. Anyway, you can always check *ex post facto* that your answer is correct.

8.1.2 Conformal Mapping

An analytic function $w = f(z)$ maps subsets of its domain of definition in the "z" plane on to subsets in the "w" plane. These maps are often useful for solving problems in two dimensional electrostatics or fluid flow. Their simplest property is geometrical: such maps are *conformal*.

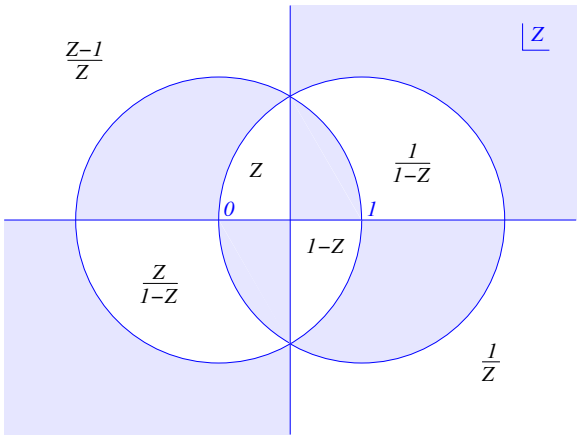

Figure 8.1: *An illustration of conformal mapping. The unshaded "triangle" marked z is mapped into the other five unshaded regions by the functions labeling them. Observe that although the regions are distorted, the angles of the "triangle" are preserved by the maps (with the exception of those corners that get mapped to infinity).*

Suppose that the derivative of $f(z)$ at a point z_0 is non-zero. Then, for z near z_0 we have

$$f(z) - f(z_0) \approx A(z - z_0), \tag{8.34}$$

where

$$A = \left.\frac{df}{dz}\right|_{z_0} . \tag{8.35}$$

If you think about the geometric interpretation of complex multiplication (multiply the magnitudes, add the arguments) you will see that the "f" image of a small neighbourhood of z_0 is stretched by a factor $|A|$, and rotated through an angle $\arg A$ — but relative angles are not altered. The map $z \mapsto f(z) = w$ is therefore *isogonal*. Our map also preserves orientation (the sense of rotation of the relative angle) and these two properties, isogonality and orientation-preservation, are what make the map conformal[2]. The conformal property fails at points where the derivative vanishes or becomes infinite.

If we can find a conformal map $z \ (\equiv x + iy) \mapsto w \ (\equiv u + iv)$ of some domain D to another D' then a function $f(z)$ that solves a potential theory problem (a Dirichlet boundary-value problem, for example) in D will lead to $f(z(w))$ solving an analogous problem in D'.

Consider, for example, the map $z \mapsto w = z + e^z$. This map takes the strip $-\infty < x < \infty$, $-\pi \leq y \leq \pi$ to the entire complex plane with cuts from $-\infty + i\pi$ to $-1 + i\pi$ and from $-\infty - i\pi$ to $-1 - i\pi$. The cuts occur because the images of the lines $y = \pm\pi$ get folded back on themselves at $w = -1 \pm i\pi$, where the derivative of $w(z)$ vanishes. (See figure 8.2)

In this case, the imaginary part of the function $f(z) = x + iy$ trivially solves the Dirichlet problem $\nabla^2_{x,y}\, y = 0$ in the infinite strip, with $y = \pi$ on the upper boundary and $y = -\pi$ on the lower boundary. The function $y(u, v)$, now quite non-trivially, solves $\nabla^2_{u,v}\, y = 0$ in the entire w plane, with $y = \pi$ on the half-line running from $-\infty + i\pi$ to $-1 + i\pi$, and $y = -\pi$ on the half-line running from $-\infty - i\pi$ to $-1 - i\pi$. We may regard the images of the lines $y = const.$ (solid curves) as being the streamlines of an irrotational and incompressible flow out of the end of a tube into an infinite region, or as the equipotentials near the edge of a pair of capacitor plates. In the latter case, the images of the lines $x = const.$ (dotted curves) are the corresponding field-lines

Example: The Joukowski map. This map is famous in the history of aeronautics because it can be used to map the exterior of a circle to the exterior of an aerofoil-shaped region. We can use the *Milne-Thomson circle theorem* (see 8.3.2) to find the streamlines for the flow past a circle in the z plane,

[2]If f were a function of $\overline{z}$ only, then the map would still be isogonal, but would reverse the orientation. We call such maps *antiholomorphic* or *anti-conformal*.

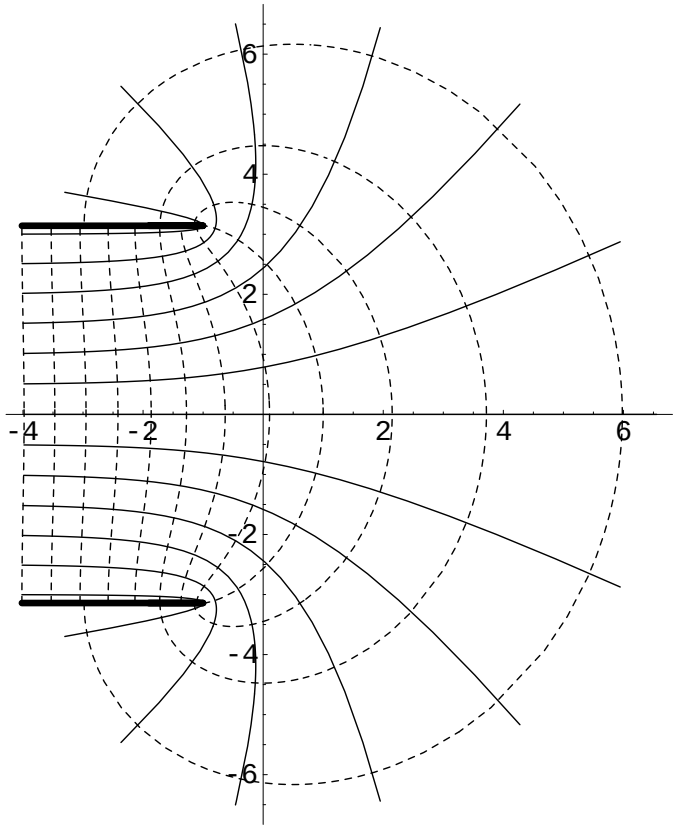

Figure 8.2: *Image of part of the strip* $-\pi \leq y \leq \pi$, $-\infty < x < \infty$ *under the map* $z \mapsto w = z + e^z$.

and then use Joukowski's transformation,

$$w = f(z) = \frac{1}{2}\left(z + \frac{1}{z}\right), \tag{8.36}$$

to map this simple flow to the flow past the aerofoil. To produce an aerofoil shape, the circle must go through the point $z = 1$, where the derivative of f vanishes, and the image of this point becomes the sharp trailing edge of the aerofoil.

The Riemann Mapping Theorem

There are tables of conformal maps for D, D' pairs, but an underlying principle is provided by the Riemann mapping theorem:
Theorem: The interior of any simply connected domain D in $\mathbb{C}$ whose boundary consists of more that one point can be mapped conformally one-to-one and onto the interior of the unit circle. It is possible to choose an arbitrary interior point w_0 of D and map it to the origin, and to take an arbitrary direction through w_0 and make it the direction of the real axis. With these two choices the mapping is unique.

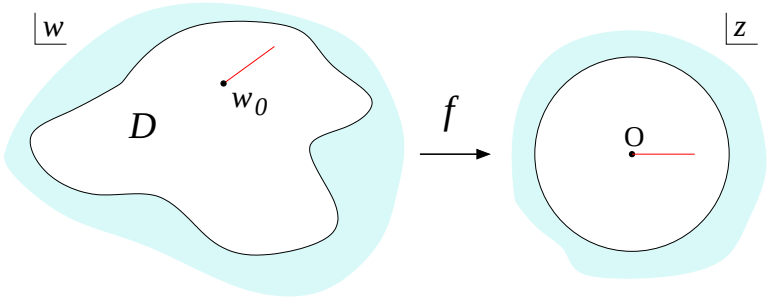

Figure 8.3: *The Riemann mapping theorem.*

This theorem was first stated in Riemann's PhD thesis in 1851. He regarded it as "obvious" for the reason that we will give as a physical "proof." Riemann's argument is not rigorous, however, and it was not until 1912 that a real proof was obtained by Constantin Carathéodory. A proof that is both shorter and more in spirit of Riemann's ideas was given by Leopold Fejér and Frigyes Riesz in 1922.

For the physical "proof," observe that in the function

$$-\frac{1}{2\pi}\ln z = -\frac{1}{2\pi}\left\{\ln|z| + i\theta\right\}, \tag{8.37}$$

the real part $\phi = -\frac{1}{2\pi}\ln|z|$ is the potential of a unit charge at the origin, and with the additive constant chosen so that $\phi = 0$ on the circle $|z| = 1$. Now imagine that we have solved the two-dimensional electrostatics problem of finding the potential for a unit charge located at $w_0 \in D$, also with the boundary of D being held at zero potential. We have

$$\nabla^2\phi_1 = -\delta^2(w - w_0), \quad \phi_1 = 0 \quad \text{on} \quad \partial D. \tag{8.38}$$

Now find the ϕ_2 that is harmonically conjugate to ϕ_1. Set

$$\phi_1 + i\phi_2 = \Phi(w) = -\frac{1}{2\pi}\ln(ze^{i\alpha}) \tag{8.39}$$

where α is a real constant. We see that the transformation $w \mapsto z$, or

$$z = e^{-i\alpha}e^{-2\pi\Phi(w)}, \tag{8.40}$$

does the job of mapping the interior of D into the interior of the unit circle, and the boundary of D to the boundary of the unit circle. Note how our freedom to choose the constant α is what allows us to "take an arbitrary direction through w_0 and make it the direction of the real axis."

Example: To find the map that takes the upper half-plane into the unit circle, with the point $z = i$ mapping to the origin, we use the method of images to solve for the complex potential of a unit charge at $w = i$:

$$\begin{aligned}\phi_1 + i\phi_2 &= -\frac{1}{2\pi}\left(\ln(w - i) - \ln(w + i)\right) \\ &= -\frac{1}{2\pi}\ln(e^{i\alpha}z).\end{aligned}$$

Therefore

$$z = e^{-i\alpha}\frac{w - i}{w + i}. \tag{8.41}$$

We immediately verify that that this works: we have $|z| = 1$ when w is real, and $z = 0$ at $w = i$.

The difficulty with the physical argument is that it is not clear that a solution to the point-charge electrostatics problem exists. In three dimensions,

for example, there is no solution when the boundary has a sharp inward directed spike. (We cannot physically realize such a situation either: the electric field becomes unboundedly large near the tip of a spike, and boundary charge will leak off and neutralize the point charge.) There might well be analogous difficulties in two dimensions if the boundary of D is pathological. However, the fact that there *is* a proof of the Riemann mapping theorem shows that the two-dimensional electrostatics problem does always have a solution, at least in the *interior* of D — even if the boundary is an infinite-length fractal. However, unless ∂D is reasonably smooth the resulting Riemann map cannot be continuously extended to the boundary. When the boundary of D *is* a smooth closed curve, then the the boundary of D *will* map one-to-one and continuously onto the boundary of the unit circle.

Exercise 8.1: *Van der Pauw's Theorem.*[3] This problem explains a practical method of for determining the conductivity σ of a material, given a sample in the form of of a wafer of uniform thickness d, but of irregular shape. In practice at the Phillips company in Eindhoven, this was a wafer of semiconductor cut from an unmachined boule.

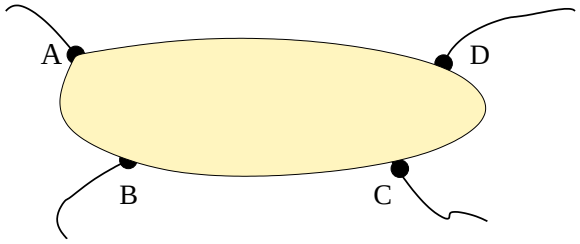

Figure 8.4: A thin semiconductor wafer with attached leads.

We attach leads to point contacts A, B, C, D, taken in anticlockwise order, on the periphery of the wafer and drive a current I_{AB} from A to B. We record the potential difference $V_D - V_C$ and so find $R_{AB,DC} = (V_D - V_C)/I_{AB}$. Similarly we measure $R_{BC,AD}$. The current flow in the wafer is assumed to be two dimensional, and to obey

$$\mathbf{J} = -(\sigma d)\nabla V, \qquad \nabla \cdot \mathbf{J} = 0,$$

[3]L. J. Van der Pauw, *Phillips Research Reps*. **13** (1958) 1. See also A. M. Thompson, D. G. Lampard, *Nature* **177** (1956) 888, and D. G. Lampard. *Proc. Inst. Elec. Eng. C.* **104** (1957) 271, for the "Calculable Capacitor."

and $\mathbf{n} \cdot \mathbf{J} = 0$ at the boundary (except at the current source and drain). The potential V is therefore harmonic, with Neumann boundary conditions.

Van der Pauw claims that

$$\exp\{-\pi\sigma d R_{AB,DC}\} + \exp\{-\pi\sigma d R_{BC,AD}\} = 1.$$

From this σd can be found numerically.

a) First show that Van der Pauw's claim is true if the wafer were the entire upper half-plane with A, B, C, D on the real axis with $x_A < x_B < x_C < x_D$.
b) Next, taking care to consider the transformation of the current source terms and the Neumann boundary conditions, show that the claim is invariant under conformal maps, and, by mapping the wafer to the upper half-plane, show that it is true in general.

8.2 Complex Integration: Cauchy and Stokes

In this section we will define the integral of an analytic function, and make contact with the exterior calculus from chapters 2-4. The most obvious difference between the real and complex integral is that in evaluating the definite integral of a function in the complex plane we must specify the path along which we integrate. When this path of integration is the boundary of a region, it is often called a *contour* from the use of the word in the graphic arts to describe the outline of something. The integrals themselves are then called *contour integrals*.

8.2.1 The Complex Integral

The complex integral

$$\int_\Gamma f(z)dz \tag{8.42}$$

over a path Γ may be defined by expanding out the real and imaginary parts

$$\int_\Gamma f(z)dz \equiv \int_\Gamma (u+iv)(dx+idy) = \int_\Gamma (udx - vdy) + i\int_\Gamma (vdx + udy). \tag{8.43}$$

and treating the two integrals on the right hand side as standard vector-calculus line-integrals of the form $\int \mathbf{v}\cdot d\mathbf{r}$, one with $\mathbf{v} \to (u, -v)$ and and one with $\mathbf{v} \to (v, u)$.

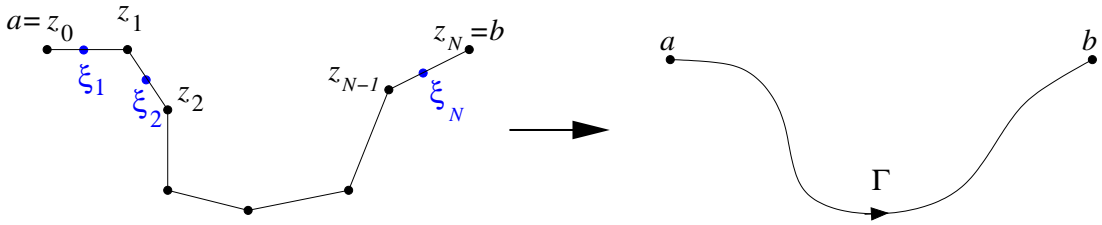

Figure 8.5: *A chain approximation to the curve* Γ.

The complex integral can also be constructed as the limit of a Riemann sum in a manner parallel to the definition of the real-variable Riemann integral of elementary calculus. Replace the path Γ with a chain composed of of N line-segments z_0-to-z_1, z_1-to-z_2, all the way to z_{N-1}-to-z_N. Now let ξ_m lie on the line segment joining z_{m-1} and z_m. Then the integral $\int_\Gamma f(z)dz$ is the limit of the (Riemann) sum

$$\sum_{m=1}^{N} f(\xi_m)(z_m - z_{m-1}) \tag{8.44}$$

as N gets large and all the $|z_m - z_{m-1}| \to 0$. For this definition to make sense and be useful, the limit must be independent of both how we chop up the curve and how we select the points ξ_m. This may be shown to be the case when the integration path is smooth and the function being integrated is continuous.

The Riemann-sum definition of the integral leads to a useful inequality: combining the triangle inequality $|a+b| \le |a| + |b|$ with $|ab| = |a|\,|b|$ we deduce that

$$\begin{aligned} \left| \sum_{m=1}^{N} f(\xi_m)(z_m - z_{m-1}) \right| &\le \sum_{m=1}^{N} |f(\xi_m)(z_m - z_{m-1})| \\ &= \sum_{m=1}^{N} |f(\xi_m)|\,|(z_m - z_{m-1})|. \end{aligned} \tag{8.45}$$

For sufficiently smooth curves the last sum converges to the real integral $\int_\Gamma |f(z)|\,|dz|$, and we deduce that

$$\left| \int_\Gamma f(z)\,dz \right| \le \int_\Gamma |f(z)|\,|dz|. \tag{8.46}$$

For curves Γ that are smooth enough to have a well-defined length $|\Gamma|$, we will have $\int_\Gamma |dz| = |\Gamma|$. From this we conclude that if $|f| \leq M$ on Γ, then we have the *Darboux inequality*

$$\left| \int_\Gamma f(z)\, dz \right| \leq M|\Gamma|. \tag{8.47}$$

We shall find many uses for this inequality.

The Riemann sum definition also makes it clear that if $f(z)$ is the derivative of another analytic function $g(z)$, *i.e.*

$$f(z) = \frac{dg}{dz}, \tag{8.48}$$

then, for Γ a smooth path from $z = a$ to $z = b$, we have

$$\int_\Gamma f(z)dz = g(b) - g(a). \tag{8.49}$$

This follows by approximating $f(\xi_m) \approx (g(z_m) - g(z_{m-1}))/(z_m - z_{m-1})$, and observing that the resultant Riemann sum

$$\sum_{m=1}^{N} \Big(g(z_m) - g(z_{m-1}) \Big) \tag{8.50}$$

telescopes. The approximation to the derivative will become exact in the limit $|z_m - z_{m-1}| \to 0$. Thus, when $f(z)$ is the derivative of another function, the integral is independent of the route that Γ takes from a to b.

We shall see that any analytic function is (at least locally) the derivative of another analytic function, and so this path independence holds generally — provided that we do not try to move the integration contour over a place where f ceases to be differentiable. This is the essence of what is known as *Cauchy's Theorem* — although, as with much of complex analysis, the result was known to Gauss.

8.2.2 Cauchy's theorem

Before we state and prove Cauchy's theorem, we must introduce an orientation convention and some traditional notation. Recall that a p-chain is a finite formal sum of p-dimensional oriented surfaces or curves, and that a

p-cycle is a p-chain Γ whose boundary vanishes: $\partial\Gamma = 0$. A 1-cycle that consists of only a single connected component is a closed curve. We will mostly consider integrals over *simple* closed curves — these being curves that do not self intersect — or 1-cycles consisting of finite formal sums of such curves. The orientation of a simple closed curve can be described by the sense, clockwise or anticlockwise, in which we traverse it. We will adopt the convention that a positively oriented curve is one such that the integration is performed in a *anticlockwise* direction. The integral over a chain Γ of oriented simple closed curves will be denoted by the symbol $\oint_\Gamma f\,dz$.

We now establish Cauchy's theorem by relating it to our previous work with exterior derivatives: Suppose that f is analytic with a a domain D, so that $\partial_{\overline{z}} f = 0$ within D. We therefore have that the the exterior derivative of f is

$$df = \partial_z f\,dz + \partial_{\overline{z}} f\,d\overline{z} = \partial_z f\,dz. \tag{8.51}$$

Now suppose that the simple closed curve Γ is the boundary of a region $\Omega \subset D$. We can exploit Stokes' theorem to deduce that

$$\oint_{\Gamma=\partial\Omega} f(z)dz = \int_\Omega d(f(z)dz) = \int_\Omega (\partial_z f)\,dz \wedge dz = 0. \tag{8.52}$$

The last integral is zero because $dz \wedge dz = 0$. We may state our result as: *Theorem (Cauchy, in modern language): The integral of an analytic function over a 1-cycle that is homologous to zero vanishes.*

The zero result is only guaranteed if the function f is analytic throughout the region Ω. For example, if Γ is the unit circle $z = e^{i\theta}$ then

$$\oint_\Gamma \left(\frac{1}{z}\right) dz = \int_0^{2\pi} e^{-i\theta}\,d\left(e^{i\theta}\right) = i\int_0^{2\pi} d\theta = 2\pi i. \tag{8.53}$$

Cauchy's theorem is not applicable because $1/z$ is *singular*, *i.e.* not differentiable, at $z = 0$. The formula (8.53) will hold for Γ any contour homologous to the unit circle in $\mathbb{C} \setminus 0$, the complex plane punctured by the removal of the point $z = 0$. Thus

$$\oint_\Gamma \left(\frac{1}{z}\right) dz = 2\pi i \tag{8.54}$$

for any contour Γ that encloses the origin. We can deduce a rather remarkable formula from (8.54): Writing $\Gamma = \partial\Omega$ with anticlockwise orientation, we use Stokes' theorem to obtain

$$\oint_{\partial\Omega} \left(\frac{1}{z}\right) dz = \int_\Omega \partial_{\overline{z}} \left(\frac{1}{z}\right) d\overline{z} \wedge dz = \begin{cases} 2\pi i, & 0 \in \Omega, \\ 0, & 0 \notin \Omega. \end{cases} \tag{8.55}$$

Since $d\overline{z} \wedge dz = 2i dx \wedge dy$, we have established that

$$\partial_{\overline{z}} \left(\frac{1}{z} \right) = \pi \delta^2(x, y). \tag{8.56}$$

This rather cryptic formula encodes one of the most useful results in mathematics.

Perhaps perversely, functions that are more singular than $1/z$ have vanishing integrals about their singularities. With Γ again the unit circle, we have

$$\oint_\Gamma \left(\frac{1}{z^2} \right) dz = \int_0^{2\pi} e^{-2i\theta}\, d\left(e^{i\theta}\right) = i \int_0^{2\pi} e^{-i\theta}\, d\theta = 0. \tag{8.57}$$

The same is true for all higher integer powers:

$$\oint_\Gamma \left(\frac{1}{z^n} \right) dz = 0, \quad n \geq 2. \tag{8.58}$$

We can understand this vanishing in another way, by evaluating the integral as

$$\oint_\Gamma \left(\frac{1}{z^n} \right) dz = \oint_\Gamma \frac{d}{dz} \left(-\frac{1}{n-1} \frac{1}{z^{n-1}} \right) dz = \left[-\frac{1}{n-1} \frac{1}{z^{n-1}} \right]_\Gamma = 0, \quad n \neq 1. \tag{8.59}$$

Here, the notation $[A]_\Gamma$ means the difference in the value of A at two ends of the integration path Γ. For a closed curve the difference is zero because the two ends are at the same point. This approach reinforces the fact that the complex integral can be computed from the "anti-derivative" in the same way as the real-variable integral. We also see why $1/z$ is special. It is the derivative of $\ln z = \ln |z| + i \arg z$, and $\ln z$ is not really a function, as it is multivalued. In evaluating $[\ln z]_\Gamma$ we must follow the continuous evolution of $\arg z$ as we traverse the contour. As the origin is within the contour, this angle increases by 2π, and so

$$[\ln z]_\Gamma = [i \arg z]_\Gamma = i \left(\arg e^{2\pi i} - \arg e^{0i} \right) = 2\pi i. \tag{8.60}$$

Exercise 8.2: Suppose $f(z)$ is analytic in a simply-connected domain D, and $z_0 \in D$. Set $g(z) = \int_{z_0}^{z} f(z)\, dz$ along some path in D from z_0 to z. Use the path-independence of the integral to compute the derivative of $g(z)$ and show that

$$f(z) = \frac{dg}{dz}.$$

This confirms our earlier claim that any analytic function is the derivative of some other analytic function.

Exercise 8.3:*The "D-bar" problem*: Suppose we are given a simply-connected domain Ω, and a function $f(z,\overline{z})$ defined on it, and wish to find a function $F(z,\overline{z})$ such that

$$\frac{\partial F(z,\overline{z})}{\partial \overline{z}} = f(z,\overline{z}), \quad (z,\overline{z}) \in \Omega.$$

Use (8.56) to argue formally that the general solution is

$$F(\zeta,\bar{\zeta}) = -\frac{1}{\pi}\int_\Omega \frac{f(z,\overline{z})}{z-\zeta} dx \wedge dy + g(\zeta),$$

where $g(\zeta)$ is an arbitrary analytic function. This result can be shown to be correct by more rigorous reasoning.

8.2.3 The residue theorem

The essential tool for computations with complex integrals is provided by the *residue theorem*. With the aid of this theorem, the evaluation of contour integrals becomes easy. All one has to do is identify points at which the function being integrated blows up, and examine just how it blows up.

If, near the point z_i, the function can be written

$$f(z) = \left\{ \frac{a_N^{(i)}}{(z-z_i)^N} + \cdots + \frac{a_2^{(i)}}{(z-z_i)^2} + \frac{a_1^{(i)}}{(z-z_i)} \right\} g^{(i)}(z), \tag{8.61}$$

where $g^{(i)}(z)$ is analytic and non-zero at z_i, then $f(z)$ has a *pole* of order N at z_i. If $N=1$ then $f(z)$ is said to have a *simple pole* at z_i. We can normalize $g^{(i)}(z)$ so that $g^{(i)}(z_i)=1$, and then the coefficient, $a_1^{(i)}$, of $1/(z-z_i)$ is called the *residue* of the pole at z_i. The coefficients of the more singular terms do not influence the result of the integral, but N must be finite for the singularity to be called a pole.

Theorem: Let the function $f(z)$ be analytic within and on the boundary $\Gamma = \partial D$ of a simply connected domain D, with the exception of finite number of points at which $f(z)$ has poles. Then

$$\oint_\Gamma f(z)\, dz = \sum_{\text{poles} \in D} 2\pi i \,(\text{residue at pole}), \tag{8.62}$$

the integral being traversed in the positive (anticlockwise) sense.

We prove the residue theorem by drawing small circles C_i about each singular point z_i in D.

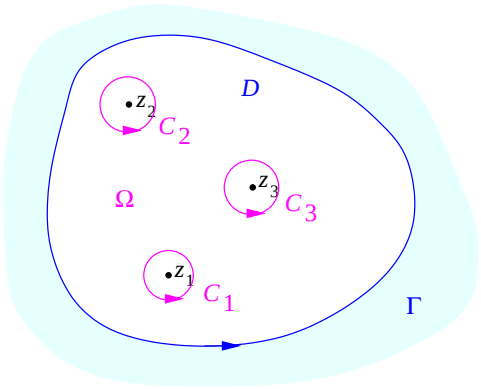

Figure 8.6: *Circles for the residue theorem.*

We now assert that

$$\oint_\Gamma f(z)\,dz = \sum_i \oint_{C_i} f(z)\,dz, \tag{8.63}$$

because the 1-cycle

$$C \equiv \Gamma - \sum_i C_i = \partial\Omega \tag{8.64}$$

is the boundary of a region Ω in which f is analytic, and hence C is homologous to zero. If we make the radius R_i of the circle C_i sufficiently small, we may replace each $g^{(i)}(z)$ by its limit $g^{(i)}(z_i) = 1$, and so take

$$\begin{aligned} f(z) \;&\to\; \left\{ \frac{a_1^{(i)}}{(z-z_i)} + \frac{a_2^{(i)}}{(z-z_i)^2} + \cdots + \frac{a_N^{(i)}}{(z-z_i)^N} \right\} g^{(i)}(z_i) \\ &= \frac{a_1^{(i)}}{(z-z_i)} + \frac{a_2^{(i)}}{(z-z_i)^2} + \cdots + \frac{a_N^{(i)}}{(z-z_i)^N}, \end{aligned} \tag{8.65}$$

on C_i. We then evaluate the integral over C_i by using our previous results to get

$$\oint_{C_i} f(z)\,dz = 2\pi i a_1^{(i)}. \tag{8.66}$$

The integral around Γ is therefore equal to $2\pi i \sum_i a_1^{(i)}$.

The restriction to contours containing only finitely many poles arises for two reasons: Firstly, with infinitely many poles, the sum over i might not converge; secondly, there may be a point whose every neighbourhood contains infinitely many of the poles, and there our construction of drawing circles around each individual pole would not be possible.

Exercise 8.4: *Poisson's Formula.* The function $f(z)$ is analytic in $|z| < R'$. Prove that if $|a| < R < R'$,

$$f(a) = \frac{1}{2\pi i} \oint_{|z|=R} \frac{R^2 - \bar{a}a}{(z-a)(R^2 - \bar{a}z)} f(z) dz.$$

Deduce that, for $0 < r < R$,

$$f(re^{i\theta}) = \frac{1}{2\pi} \int_0^{2\pi} \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \phi) + r^2} f(Re^{i\phi}) d\phi.$$

Show that this formula solves the boundary-value problem for Laplace's equation in the disc $|z| < R$.

Exercise 8.5: *Bergman Kernel.* The Hilbert space of analytic functions on a domain D with inner product

$$\langle f, g \rangle = \int_D \bar{f} g \, dxdy$$

is called the Bergman[4] space of D.

a) Suppose that $\varphi_n(z)$, $n = 0, 1, 2, \ldots$, are a complete set of orthonormal functions on the Bergman space. Show that

$$K(\zeta, z) = \sum_{m=0}^{\infty} \varphi_m(\zeta) \overline{\varphi_m(z)}.$$

has the property that

$$g(\zeta) = \iint_D K(\zeta, z) g(z) \, dxdy.$$

[4]This space should not be confused with the Bargmann-Fock space of analytic functions on the entirety of $\mathbb{C}$ with inner product

$$\langle f, g \rangle = \int_{\mathbb{C}} e^{-|z|^2} \bar{f} g \, d^2 z.$$

Stefan Bergman and Valentine Bargmann are two different people.

for any function g analytic in D. Thus $K(\zeta, z)$ plays the role of the delta function on the space of analytic functions on D. This object is called the *reproducing* or *Bergman kernel*. By taking $g(z) = \varphi_n(z)$, show that it is the unique integral kernel with the reproducing property.

b) Consider the case of D being the unit circle. Use the Gramm-Schmidt procedure to construct an orthonormal set from the functions z^n, $n = 0, 1, 2, \ldots$. Use the result of part a) to conjecture (because we have not proved that the set is complete) that, for the unit circle,

$$K(\zeta, z) = \frac{1}{\pi} \frac{1}{(1 - \zeta\overline{z})^2}.$$

c) For any smooth, complex valued, function g defined on a domain D and its boundary, use Stokes' theorem to show that

$$\iint_D \partial_{\overline{z}} g(z, \overline{z}) dx dy = \frac{1}{2i} \oint_{\partial D} g(z, \overline{z}) dz.$$

Use this to verify that this the $K(\zeta, z)$ you constructed in part b) is indeed a (and hence "the") reproducing kernel.

d) Now suppose that D is a simply connected domain whose boundary ∂D is a smooth curve. We know from the Riemann mapping theorem that there exists an analytic function $f(z) = f(z; \zeta)$ that maps D onto the interior of the unit circle in such a way that $f(\zeta) = 0$ and $f'(\zeta)$ is real and non-zero. Show that if we set $K(\zeta, z) = \overline{f'(z)} f'(\zeta)/\pi$, then, by using part c) together with the residue theorem to evaluate the integral over the boundary, we have

$$g(\zeta) = \iint_D K(\zeta, z) g(z)\, dx dy.$$

This $K(\zeta, z)$ must therefore be the reproducing kernel. We see that if we know K we can recover the map f from

$$f'(z; \zeta) = \sqrt{\frac{\pi}{K(\zeta, \zeta)}} K(z, \zeta).$$

e) Apply the formula from part d) to the unit circle, and so deduce that

$$f(z; \zeta) = \frac{z - \zeta}{1 - \bar{\zeta} z}$$

is the unique function that maps the unit circle onto itself with the point ζ mapping to the origin and with the horizontal direction through ζ remaining horizontal.

8.3 Applications

We now know enough about complex variables to work through some interesting applications, including the mechanism by which an aeroplane flies.

8.3.1 Two-dimensional vector calculus

It is often convenient to use complex co-ordinates for vectors and tensors. In these co-ordinates the standard metric on $\mathbb{R}^2$ becomes

$$\begin{aligned} \text{“}ds^2\text{”} &= dx \otimes dx + dy \otimes dy \\ &= d\overline{z} \otimes dz \\ &= g_{zz} dz \otimes dz + g_{\overline{z}z} d\overline{z} \otimes dz + g_{z\overline{z}} dz \otimes d\overline{z} + g_{\overline{z}\overline{z}} d\overline{z} \otimes d\overline{z}, \end{aligned} \quad (8.67)$$

so the complex co-ordinate components of the metric tensor are $g_{zz} = g_{\overline{z}\overline{z}} = 0$, $g_{z\overline{z}} = g_{\overline{z}z} = \frac{1}{2}$. The inverse metric tensor is $g^{z\overline{z}} = g^{\overline{z}z} = 2$, $g^{zz} = g^{\overline{z}\overline{z}} = 0$.

In these co-ordinates the Laplacian is

$$\nabla^2 = g^{ij}\partial^2_{ij} = 2(\partial_z \partial_{\overline{z}} + \partial_{\overline{z}} \partial_z). \quad (8.68)$$

When f has singularities, it is not safe to assume that $\partial_z \partial_{\overline{z}} f = \partial_{\overline{z}} \partial_z f$. For example, from

$$\partial_{\overline{z}} \left(\frac{1}{z} \right) = \pi \delta^2(x, y), \quad (8.69)$$

we deduce that

$$\partial_{\overline{z}} \partial_z \ln z = \pi \delta^2(x, y). \quad (8.70)$$

When we evaluate the derivatives in the opposite order, however, we have

$$\partial_z \partial_{\overline{z}} \ln z = 0. \quad (8.71)$$

To understand the source of the non-commutativity, take real and imaginary parts of these last two equations. Write $\ln z = \ln |z| + i\theta$, where $\theta = \arg z$, and add and subtract. We find

$$\begin{aligned} \nabla^2 \ln |z| &= 2\pi \delta^2(x, y), \\ (\partial_x \partial_y - \partial_y \partial_x)\theta &= 2\pi \delta^2(x, y). \end{aligned} \quad (8.72)$$

The first of these shows that $\frac{1}{2\pi} \ln |z|$ is the Green function for the Laplace operator, and the second reveals that the vector field $\nabla\theta$ is singular, having a delta function “curl” at the origin.

If we have a vector field $\mathbf{v}$ with contravariant components (v^x, v^y) and (numerically equal) covariant components (v_x, v_y) then the covariant components in the complex co-ordinate system are $v_z = \frac{1}{2}(v_x - iv_y)$ and $v_{\overline{z}} = \frac{1}{2}(v_x + iv_y)$. This can be obtained by a using the change of co-ordinates rule, but a quicker route is to observe that

$$\mathbf{v} \cdot d\mathbf{r} = v_x dx + v_y dy = v_z dz + v_{\overline{z}} d\overline{z}. \tag{8.73}$$

Now

$$\partial_{\overline{z}} v_z = \frac{1}{4}(\partial_x v_x + \partial_y v_y) + i\frac{1}{4}(\partial_y v_x - \partial_x v_y). \tag{8.74}$$

Thus the statement that $\partial_{\overline{z}} v_z = 0$ is equivalent to the vector field $\mathbf{v}$ being both solenoidal (incompressible) and irrotational. This can also be expressed in form language by setting $\eta = v_z\, dz$ and saying that $d\eta = 0$ means that the corresponding vector field is both solenoidal and irrotational.

8.3.2 Milne-Thomson Circle Theorem

As we mentioned earlier, we can describe an irrotational and incompressible fluid motion either by a velocity potential

$$v_x = \partial_x \phi, \quad v_y = \partial_y \phi, \tag{8.75}$$

where $\mathbf{v}$ is automatically irrotational but incompressibilty requires $\nabla^2 \phi = 0$, or by a stream function

$$v_x = \partial_y \chi, \quad v_y = -\partial_x \chi, \tag{8.76}$$

where $\mathbf{v}$ is automatically incompressible but irrotationality requires $\nabla^2 \chi = 0$. We can combine these into a single *complex stream function* $\Phi = \phi + i\chi$ which, for an irrotational incompressible flow, satisfies the Cauchy-Riemann equations and is therefore an analytic function of z. We see that

$$2v_z = \frac{d\Phi}{dz}, \tag{8.77}$$

ϕ and χ making equal contributions.

The Milne-Thomson theorem says that if Φ is the complex stream function for a flow in unobstructed space, then

$$\widetilde{\Phi} = \Phi(z) + \overline{\Phi}\left(\frac{a^2}{z}\right) \tag{8.78}$$

is the stream function after the cylindrical obstacle $|z| = a$ is inserted into the flow. Here $\overline{\Phi}(z)$ denotes the analytic function defined by $\overline{\Phi}(z) = \overline{\Phi(\overline{z})}$. To see that this works, observe that $a^2/z = \overline{z}$ on the curve $|z| = a$, and so on this curve $\operatorname{Im}\widetilde{\Phi} = \chi = 0$. The surface of the cylinder has therefore become a streamline, and so the flow does not penetrate into the cylinder. If the original flow is created by souces and sinks exterior to $|z| = a$, which will be singularities of Φ, the additional term has singularites that lie only within $|z| = a$. These will be the "images" of the sources and sinks in the sense of the "method of images."

Example: A uniform flow with speed U in the x direction has $\Phi(z) = Uz$. Inserting a cylinder makes this

$$\tilde{\Phi}(z) = U\left(z + \frac{a^2}{z}\right). \tag{8.79}$$

Because v_z is the derivative of this, we see that the perturbing effect of the obstacle on the velocity field falls off as the square of the distance from the cylinder. This is a general result for obstructed flows.

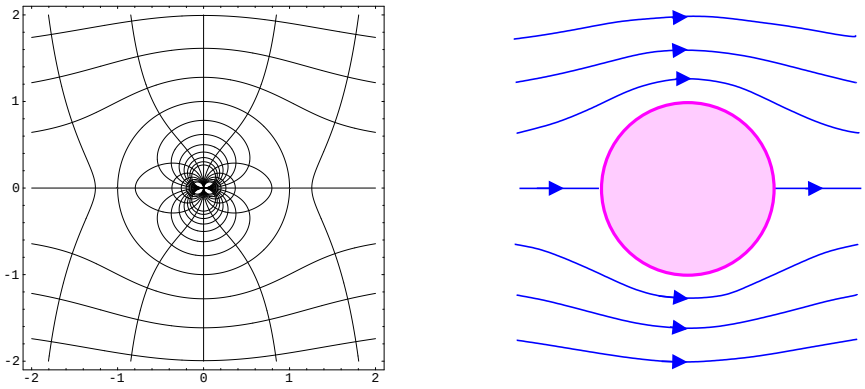

Figure 8.7: *The real and imaginary parts of the function $z + z^{-1}$ provide the velocity potentials and streamlines for irrotational incompressible flow past a cylinder of unit radius.*

8.3.3 Blasius and Kutta-Joukowski Theorems

We now derive the celebrated result, discovered independently by Martin Wilhelm Kutta (1902) and Nikolai Egorovich Joukowski (1906), that the

lift per unit span of an aircraft wing is equal to the product of the density of the air ρ, the circulation $\kappa \equiv \oint \mathbf{v} \cdot d\mathbf{r}$ about the wing, and the forward velocity U of the wing through the air. Their theory treats the air as being incompressible—a good approximation unless the flow-velocities approach the speed of sound—and assumes that the wing is long enough that the flow can be regarded as being two dimensional.

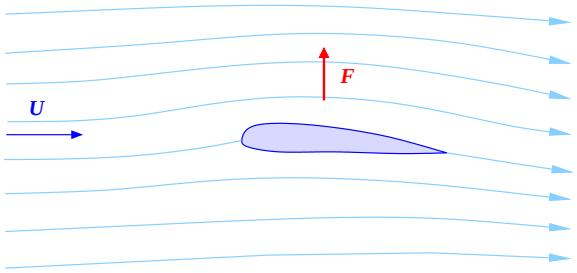

Figure 8.8: *Flow past an aerofoil.*

Begin by recalling how the momentum flux tensor

$$T_{ij} = \rho v_i v_j + g_{ij} P \tag{8.80}$$

enters fluid mechanics. In cartesian co-ordinates, and in the presence of an external body force f_i acting on the fluid, the Euler equation of motion for the fluid is

$$\rho(\partial_t v_i + v^j \partial_j v_i) = -\partial_i P + f_i. \tag{8.81}$$

Here P is the pressure and we are distinguishing between co and contravariant components, although at the moment $g_{ij} \equiv \delta_{ij}$. We can combine Euler's equation with the law of mass conservation,

$$\partial_t \rho + \partial^i(\rho v_i) = 0, \tag{8.82}$$

to obtain

$$\partial_t(\rho v_i) + \partial^j(\rho v_j v_i + g_{ij} P) = f_i. \tag{8.83}$$

This momemtum-tracking equation shows that the external force acts as a source of momentum, and that for steady flow f_i is equal to the divergence of the momentum flux tensor:

$$f_i = \partial^l T_{li} = g^{kl} \partial_k T_{li}. \tag{8.84}$$

As we are interested in steady, irrotational motion with uniform density we may use Bernoulli's theorem, $P + \frac{1}{2}\rho|v|^2 = const.$, to substitute $-\frac{1}{2}\rho|v|^2$ in place of P. (The constant will not affect the momentum flux.) With this substitution T_{ij} becomes a traceless symmetric tensor:

$$T_{ij} = \rho(v_i v_j - \frac{1}{2} g_{ij}|v|^2). \tag{8.85}$$

Using $v_z = \frac{1}{2}(v_x - iv_y)$ and

$$T_{zz} = \frac{\partial x^i}{\partial z}\frac{\partial x^j}{\partial z} T_{ij}, \tag{8.86}$$

together with

$$x \equiv x^1 = \frac{1}{2}(z + \overline{z}), \qquad y \equiv x^2 = \frac{1}{2i}(z - \overline{z}) \tag{8.87}$$

we find

$$T \equiv T_{zz} = \frac{1}{4}(T_{xx} - T_{yy} - 2iT_{xy}) = \rho(v_z)^2. \tag{8.88}$$

This is the only component of T_{ij} that we will need to consider. $T_{\overline{z}\overline{z}}$ is simply $\overline{T}$, whereas $T_{z\overline{z}} = 0 = T_{\overline{z}z}$ because T_{ij} is traceless.

In our complex co-ordinates, the equation

$$f_i = g^{kl}\partial_k T_{li} \tag{8.89}$$

reads

$$f_z = g^{\overline{z}z}\partial_{\overline{z}} T_{zz} + g^{z\overline{z}}\partial_z T_{\overline{z}z} = 2\partial_{\overline{z}} T. \tag{8.90}$$

We see that in steady flow the net momentum flux $\dot{P}_i$ out of a region Ω is given by

$$\dot{P}_z = \int_\Omega f_z \, dxdy = \frac{1}{2i}\int_\Omega f_z \, d\overline{z}dz = \frac{1}{i}\int_\Omega \partial_{\overline{z}} T \, d\overline{z}dz = \frac{1}{i}\oint_{\partial\Omega} T \, dz. \tag{8.91}$$

We have used Stokes' theorem at the last step. In regions where there is no external force, T is analytic, $\partial_{\overline{z}} T = 0$, and the integral will be independent of the choice of contour $\partial\Omega$. We can subsititute $T = \rho v_z^2$ to get

$$\dot{P}_z = -i\rho \oint_{\partial\Omega} v_z^2 \, dz. \tag{8.92}$$

To apply this result to our aerofoil we take can take $\partial\Omega$ to be its boundary. Then $\dot{P}_z$ is the total force exerted on the fluid by the wing, and, by Newton's third law, this is minus the force exerted by the fluid on the wing. The total force on the aerofoil is therefore

$$F_z = i\rho \oint_{\partial\Omega} v_z^2 \, dz. \tag{8.93}$$

The result (8.93) is often called *Blasius' theorem*.

Evaluating the integral in (8.93) is not immediately possible because the velocity $\mathbf{v}$ on the boundary will be a complicated function of the shape of the body. We can, however, exploit the contour independence of the integral and evaluate it over a path encircling the aerofoil at large distance where the flow field takes the asymptotic form

$$v_z = U_z + \frac{\kappa}{4\pi i}\frac{1}{z} + O\left(\frac{1}{z^2}\right). \tag{8.94}$$

The $O(1/z^2)$ term is the velocity perturbation due to the air having to flow round the wing, as with the cylinder in a free flow. To confirm that this flow has the correct circulation we compute

$$\oint \mathbf{v}\cdot d\mathbf{r} = \oint v_z dz + \oint v_{\overline{z}}\, d\overline{z} = \kappa. \tag{8.95}$$

Substituting v_z in (8.93) we find that the $O(1/z^2)$ term cannot contribute as it cannot affect the residue of any pole. The only part that does contribute is the cross term that arises from multiplying U_z by $\kappa/(4\pi i z)$. This gives

$$F_z = i\rho\left(\frac{U_z\kappa}{2\pi i}\right)\oint \frac{dz}{z} = i\rho\kappa U_z \tag{8.96}$$

so that

$$\frac{1}{2}(F_x - iF_y) = i\rho\kappa\frac{1}{2}(U_x - iU_y). \tag{8.97}$$

Thus, in conventional co-ordinates, the reaction force on the body is

$$\begin{aligned} F_x &= \rho\kappa U_y, \\ F_y &= -\rho\kappa U_x. \end{aligned} \tag{8.98}$$

The fluid therefore provides a lift force proportional to the product of the circulation with the asymptotic velocity. The force is at right angles to the incident airstream, so there is no *drag*.

The circulation around the wing is determined by the *Kutta condition* that the velocity of the flow at the sharp trailing edge of the wing be finite. If the wing starts moving into the air and the requisite circulation is not yet established then the flow under the wing does not leave the trailing edge smoothly but tries to whip round to the topside. The velocity gradients become very large and viscous forces become important and prevent the air from making the sharp turn. Instead, a *starting vortex* is shed from the trailing edge. Kelvin's theorem on the conservation of vorticity shows that this causes a circulation of equal and opposite strength to be induced about the wing.

For finite wings, the path independence of $\oint \mathbf{v} \cdot d\mathbf{r}$ means that the wings must leave a pair of trailing *wingtip vortices* of strength κ that connect back to the starting vortex to form a closed loop. The velocity field induced by the trailing vortices cause the airstream incident on the aerofoil to come from a slighly different direction than the asymptotic flow. Consequently, the lift is not quite perpendicular to the motion of the wing. For finite-length wings, therefore, lift comes at the expense of an inevitable *induced drag* force. The work that has to be done against this drag force in driving the wing forwards provides the kinetic energy in the trailing vortices.

8.4 Applications of Cauchy's Theorem

Cauchy's theorem provides the Royal Road to complex analysis. It is possible to develop the theory without it, but the path is harder going.

8.4.1 Cauchy's Integral Formula

If $f(z)$ is analytic within and on the boundary of a simply connected domain Ω, with $\partial\Omega = \Gamma$, and if ζ is a point in Ω, then, noting that the the integrand has a simple pole at $z = \zeta$ and applying the residue formula, we have *Cauchy's integral formula*

$$f(\zeta) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z - \zeta}\, dz, \quad \zeta \in \Omega. \tag{8.99}$$

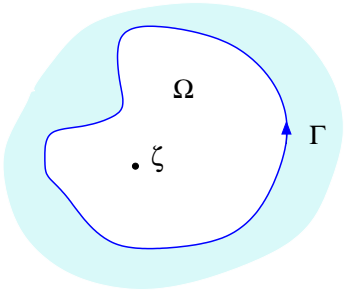

Figure 8.9: *Cauchy contour.*

This formula holds only if ζ lies within Ω. If it lies outside, then the integrand is analytic everywhere inside Ω, and so the integral gives zero.

We may show that it is legitimate to differentiate under the integral sign in Cauchy's formula. If we do so n times, we have the useful corollary that

$$f^{(n)}(\zeta) = \frac{n!}{2\pi i} \oint_\Gamma \frac{f(z)}{(z-\zeta)^{n+1}}\, dz. \tag{8.100}$$

This shows that being *once* differentiable (analytic) in a region automatically implies that $f(z)$ is differentiable *arbitrarily many times*!

Exercise 8.6: *The generalized Cauchy formula.* Suppose that we have solved a D-bar problem (see exercise 8.3), and so found an $F(z,\overline{z})$ with $\partial_{\overline{z}} F = f(z,\overline{z})$ in a region Ω. Compute the exterior derivative of

$$\frac{F(z,\overline{z})}{z-\zeta}$$

using (8.56). Now, manipulating formally with delta functions, apply Stokes' theorem to show that, for $(\zeta,\bar{\zeta})$ in the interior of Ω, we have

$$F(\zeta,\bar{\zeta}) = \frac{1}{2\pi i} \oint_{\partial\Omega} \frac{F(z,\overline{z})}{z-\zeta}\, dz - \frac{1}{\pi} \int_\Omega \frac{f(z,\overline{z})}{z-\zeta} dx\, dy.$$

This is called the *generalized Cauchy formula*. Note that the first term on the right, unlike the second, is a function only of ζ, and so is analytic.

Liouville's Theorem

A dramatic corollary of Cauchy's integral formula is provided by

Liouville's theorem: *If $f(z)$ is analytic in all of $\mathbb{C}$, and is bounded there, meaning that there is a positive real number K such that $|f(z)| < K$, then $f(z)$ is a constant.*

This result provides a powerful strategy for proving that two formulæ, $f_1(z)$ and $f_2(z)$, represent the same analytic function. If we can show that the difference $f_1 - f_2$ is analytic and tends to zero at infinity then Liouville's theorem tells us that $f_1 = f_2$.

Because the result is perhaps unintuitive, and because the methods are typical, we will spell out in detail how Liouville's theorem works. We select any two points, z_1 and z_2, and use Cauchy's formula to write

$$f(z_1) - f(z_2) = \frac{1}{2\pi i} \oint_\Gamma \left(\frac{1}{z - z_1} - \frac{1}{z - z_2} \right) f(z)\, dz. \tag{8.101}$$

We take the contour Γ to be circle of radius ρ centered on z_1. We make $\rho > 2|z_1 - z_2|$, so that when z is on Γ we are sure that $|z - z_2| > \rho/2$.

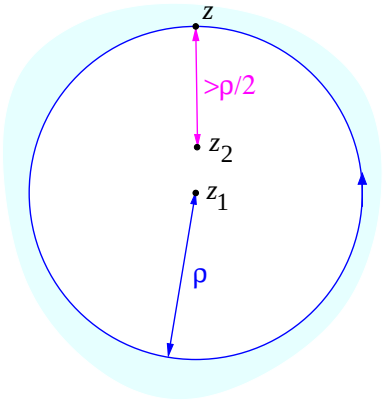

Figure 8.10: *Contour for Liouville' theorem.*

Then, using $|\int f(z)dz| \leq \int |f(z)||dz|$, we have

$$\begin{aligned} |f(z_1) - f(z_2)| &= \frac{1}{2\pi} \left| \oint_\Gamma \frac{(z_1 - z_2)}{(z - z_1)(z - z_2)} f(z)\, dz \right| \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} \frac{|z_1 - z_2| K}{\rho/2}\, d\theta = \frac{2|z_1 - z_2|K}{\rho}. \end{aligned} \tag{8.102}$$

The right hand side can be made arbitrarily small by taking ρ large enough, so we we must have $f(z_1) = f(z_2)$. As z_1 and z_2 were any pair of points, we deduce that $f(z)$ takes the same value everywhere.

8.4.2 Taylor and Laurent Series

We have defined a function to be analytic in a domain D if it is (once) complex differentiable at all points in D. It turned out that this apparently mild requirement automatically implied that the function is differentiable *arbitrarily many times* in D. In this section we shall see that knowledge of all derivatives of $f(z)$ at any single point in D is enough to completely determine the function at any other point in D. Compare this with functions of a real variable, for which it is easy to construct examples that are once but not twice differentiable, and where complete knowledge of function at a point, or in even in a neighbourhood of a point, tells us absolutely nothing of the behaviour of the function away from the point or neighbourhood.

The key ingredient in these almost magical properties of complex analytic functions is that any analytic function has a Taylor series expansion that actually converges to the function. Indeed an alternative definition of analyticity is that $f(z)$ be representable by a convergent power series. For real variables this is the definition of a *real analytic* function.

To appreciate the utility of power series representations we do need to discuss some basic properties of power series. Most of these results are extensions to the complex plane of what we hope are familiar notions from real analysis.

Consider the power series

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n \equiv \lim_{N\to\infty} S_N, \tag{8.103}$$

where S_N are the *partial sums*

$$S_N = \sum_{n=0}^{N} a_n (z - z_0)^n. \tag{8.104}$$

Suppose that this limit exists (i.e the series is convergent) for some $z = \zeta$; then it turns out that the series is *absolutely convergent*[5] for any $|z - z_0| < |\zeta - z_0|$.

[5]Recall that absolute convergence of $\sum a_n$ means that $\sum |a_n|$ converges. Absolute convergence implies convergence, and also allows us to rearrange the order of terms in the series without changing the value of the sum. Compare this with *conditional convergence*, where $\sum a_n$ converges, but $\sum |a_n|$ does not. You may remember that Riemann showed that the terms of a conditionally convergent series can be rearranged so as to *get any answer whatsoever*!

To establish this absolute convergence we may assume, without loss of generality, that $z_0 = 0$. Then, convergence of the sum $\sum a_n \zeta^n$ requires that $|a_n \zeta^n| \to 0$, and thus $|a_n \zeta^n|$ is bounded. In other words, there is a B such that $|a_n \zeta^n| < B$ for any n. We now write

$$|a_n z^n| = |a_n \zeta^n| \left|\frac{z}{\zeta}\right|^n < B \left|\frac{z}{\zeta}\right|^n . \tag{8.105}$$

The sum $\sum |a_n z^n|$ therefore converges for $|z/\zeta| < 1$, by comparison with a geometric progression.

This result, that if a power series in $(z - z_0)$ converges at a point then it converges at all points closer to z_0, shows that a power series possesses some *radius of convergence* R. The series converges for all $|z - z_0| < R$, and diverges for all $|z - z_0| > R$. (What happens *on* the circle $|z - z_0| = R$ is usually delicate, and harder to establish.) We soon show that the radius of convergence of a power series is the distance from z_0 to the nearest singularity of the function that it represents.

By comparison with a geometric progression, we may establish the following useful formulæ giving R for the series $\sum a_n z^n$:

$$\begin{aligned} R &= \lim_{n\to\infty} \frac{|a_{n-1}|}{|a_n|} \\ &= \lim_{n\to\infty} |a_n|^{1/n}. \end{aligned} \tag{8.106}$$

The proof of these formulæ is identical the real-variable version.

When we differentiate the terms in a power series, and thus take $a_n z^n \to n a_n z^{n-1}$, this does not alter R. This observation suggests that it is legitimate to evaluate the derivative of the function represented by the powers series by differentiating term-by-term. As step on the way to justifying this, observe that if the series converges at $z = \zeta$ and D_r is the domain $|z| < r < |\zeta|$ then, using the same bound as in the proof of absolute convergence, we have

$$|a_n z^n| < B \frac{|z^n|}{|\zeta|^n} < B \frac{r^n}{|\zeta|^n} = M_n \tag{8.107}$$

where $\sum M_n$ is convergent. As a consequence $\sum a_n z^n$ is *uniformly convergent* in D_r by the Weierstrass "M" test. You probably know that uniform convergence allows the interchange the order of sums and *integrals*: $\int (\sum f_n(x))dx = \sum \int f_n(x)dx$. For real variables, uniform convergence is

not a strong enough a condition for us to to safely interchange order of sums and *derivatives*: $(\sum f_n(x))'$ is not necessarily equal to $\sum f_n'(x)$. For complex analytic functions, however, Cauchy's integral formula reduces the operation of differentiation to that of integration, and so this interchange *is* permitted. In particular we have that if

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \tag{8.108}$$

and R is defined by $R = |\zeta|$ for any ζ for which the series converges, then $f(z)$ is analytic in $|z| < R$ and

$$f'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}, \tag{8.109}$$

is also analytic in $|z| < R$.

Morera's Theorem

There is is a partial converse of Cauchy's theorem:
Theorem (Morera): If $f(z)$ is defined and continuous in a domain D, and if $\oint_\Gamma f(z)\,dz = 0$ for all closed contours, then $f(z)$ is analytic in D. To prove this we set $F(z) = \int_P^z f(\zeta)\,d\zeta$. The integral is path-independent by the hypothesis of the theorem, and because $f(z)$ is continuous we can differentiate with respect to the integration limit to find that $F'(z) = f(z)$. Thus $F(z)$ is complex differentiable, and so analytic. Then, by Cauchy's formula for higher derivatives, $F''(z) = f'(z)$ exists, and so $f(z)$ itself is analytic.

A corollary of Morera's theorem is that if $f_n(z) \to f(z)$ uniformly in D, with all the f_n analytic, then

i) $f(z)$ is analytic in D, and
ii) $f_n'(z) \to f'(z)$ uniformly.

We use Morera's theorem to prove (i) (appealing to the uniform convergence to justify the interchange the order of summation and integration), and use Cauchy's theorem to prove (ii).

Taylor's Theorem for analytic functions

Theorem: Let Γ be a circle of radius ρ centered on the point a. Suppose that $f(z)$ is analytic within and on Γ, and and that the point $z = \zeta$ is within Γ.

Then $f(\zeta)$ can be expanded as a Taylor series

$$f(\zeta) = f(a) + \sum_{n=1}^{\infty} \frac{(\zeta - a)^n}{n!} f^{(n)}(a), \tag{8.110}$$

meaning that this series converges to $f(\zeta)$ for all ζ such that $|\zeta - a| < \rho$.

To prove this theorem we use identity

$$\frac{1}{z-\zeta} = \frac{1}{z-a} + \frac{(\zeta-a)}{(z-a)^2} + \cdots + \frac{(\zeta-a)^{N-1}}{(z-a)^N} + \frac{(\zeta-a)^N}{(z-a)^N}\frac{1}{z-\zeta} \tag{8.111}$$

and Cauchy's integral, to write

$$\begin{aligned} f(\zeta) &= \frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{(z-\zeta)}\,dz \\ &= \sum_{n=0}^{N-1} \frac{(\zeta-a)^n}{2\pi i}\oint \frac{f(z)}{(z-a)^{n+1}}\,dz + \frac{(\zeta-a)^N}{2\pi i}\oint \frac{f(z)}{(z-a)^N(z-\zeta)}\,dz \\ &= \sum_{n=0}^{N-1} \frac{(\zeta-a)^n}{n!} f^{(n)}(a) + R_N, \end{aligned} \tag{8.112}$$

where

$$R_N \stackrel{\text{def}}{=} \frac{(\zeta-a)^N}{2\pi i}\oint_\Gamma \frac{f(z)}{(z-a)^N(z-\zeta)}\,dz. \tag{8.113}$$

This is Taylor's theorem with remainder. For real variables this is as far as we can go. Even if a real function is differentiable infinitely many times, there is no reason for the remainder to become small. For analytic functions, however, we can show that $R_N \to 0$ as $N \to \infty$. This means that the complex-variable Taylor series is convergent, and its limit is actually equal to $f(z)$. To show that $R_N \to 0$, recall that Γ is a circle of radius ρ centered on $z = a$. Let $r = |\zeta - a| < \rho$, and let M be an upper bound for $f(z)$ on Γ. (This exists because f is continuous and Γ is a compact subset of $\mathbb{C}$.) Then, estimating the integral using methods similar to those invoked in our proof of Liouville's Theorem, we find that

$$R_N < \frac{r^N}{2\pi}\left(\frac{2\pi\rho M}{\rho^N(\rho - r)}\right). \tag{8.114}$$

As $r < \rho$, this tends to zero as $N \to \infty$.

We can take ρ as large as we like provided there are no singularities of f end up within, or on, the circle. This confirms the claim made earlier: the radius of convergence of the powers series representation of an analytic functionis the distance to the nearest singularity.

Laurent Series

Theorem (Laurent): Let Γ_1 and Γ_2 be two anticlockwise circlular paths with centre a, radii ρ_1 and ρ_2, and with $\rho_2 < \rho_1$. If $f(z)$ is analytic on the circles and within the annulus between them, then, for ζ in the annulus:

$$f(\zeta) = \sum_{n=0}^{\infty} a_n(\zeta - a)^n + \sum_{n=1}^{\infty} b_n(\zeta - a)^{-n}. \tag{8.115}$$

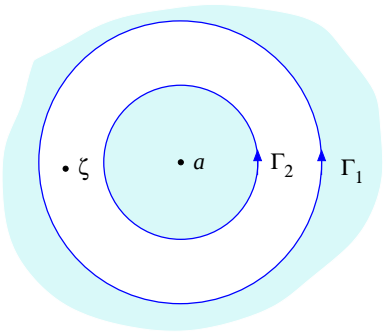

Figure 8.11: *Contours for Laurent's theorem.*

The coefficients a_n and b_n are given by

$$a_n = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{(z-a)^{n+1}}\,dz, \quad b_n = \frac{1}{2\pi i}\oint_{\Gamma_2} f(z)(z-a)^{n-1}\,dz. \tag{8.116}$$

Laurent's theorem is proved by observing that

$$f(\zeta) = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{(z-\zeta)}\,dz - \frac{1}{2\pi i}\oint_{\Gamma_2} \frac{f(z)}{(z-\zeta)}\,dz, \tag{8.117}$$

and using the identities

$$\frac{1}{z-\zeta} = \frac{1}{z-a} + \frac{(\zeta-a)}{(z-a)^2} + \cdots + \frac{(\zeta-a)^{N-1}}{(z-a)^N} + \frac{(\zeta-a)^N}{(z-a)^N}\frac{1}{z-\zeta}, \tag{8.118}$$

and

$$-\frac{1}{z-\zeta} = \frac{1}{\zeta - a} + \frac{(z-a)}{(\zeta-a)^2} + \cdots + \frac{(z-a)^{N-1}}{(\zeta-a)^N} + \frac{(z-a)^N}{(\zeta-a)^N}\frac{1}{\zeta - z}. \quad (8.119)$$

Once again we can show that the remainder terms tend to zero.
Warning: Although the coefficients a_n are given by the same integrals as in Taylor's theorem, they are not interpretable as derivatives of f unless $f(z)$ is analytic within the inner circle, in which case all the b_n are zero.

8.4.3 Zeros and Singularities

This section is something of a *nosology* — a classification of diseases — but you should study it carefully as there is some tight reasoning here, and the conclusions are the essential foundations for the rest of subject.

First a review and some definitions:

a) If $f(z)$ is analytic with a domain D, we have seen that f may be expanded in a Taylor series about any point $z_0 \in D$:

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n. \quad (8.120)$$

If $a_0 = a_1 = \cdots = a_{n-1} = 0$, and $a_n \neq 0$, so that the first non-zero term in the series is $a_n(z-z_0)^n$, we say that $f(z)$ has a *zero* of order n at z_0.

b) A *singularity* of $f(z)$ is a point at which $f(z)$ ceases to be differentiable. If $f(z)$ has no singularities at finite z (for example, $f(z) = \sin z$) then it is said to be an *entire* function.

c) If $f(z)$ is analytic in D except at $z = a$, an *isolated singularity*, then we may draw two concentric circles of centre a, both within D, and in the annulus between them we have the Laurent expansion

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n + \sum_{n=1}^{\infty} b_n (z-a)^{-n}. \quad (8.121)$$

The second term, consisting of negative powers, is called the *principal part* of $f(z)$ at $z = a$. It may happen that $b_m \neq 0$ but $b_n = 0$, $n > m$. Such a singularity is called a pole of order m at $z = a$. The coefficient b_1, which may be 0, is called the residue of f at the pole $z = a$. If the series of negative powers does not terminate, the singularity is called an *isolated essential singularity*

Now some observations:

i) Suppose $f(z)$ is analytic in a domain D containing the point $z = a$. Then we can expand: $f(z) = \sum a_n(z-a)^n$. If $f(z)$ is zero at $z = 0$, then there are exactly two possibilities: a) all the a_n vanish, and then $f(z)$ is identically zero; b) there is a first non-zero coefficient, a_m say, and so $f(z) = z^m\varphi(z)$, where $\varphi(a) \neq 0$. In the second case f is said to possess a *zero of order* m at $z = a$.

ii) If $z = a$ is a zero of order m, of $f(z)$ then the zero is *isolated – i.e.* there is a neighbourhood of a which contains no other zero. To see this observe that $f(z) = (z-a)^m\varphi(z)$ where $\varphi(z)$ is analytic and $\varphi(a) \neq 0$. Analyticity implies continuity, and by continuity there is a neighbourhood of a in which $\varphi(z)$ does not vanish.

iii) Limit points of zeros I: Suppose that we know that $f(z)$ is analytic in D and we know that it vanishes at a sequence of points $a_1, a_2, a_3, \ldots \in D$. If these points have a limit point[6] that is interior to D then $f(z)$ must, by continuity, be zero there. But this would be a non-isolated zero, in contradiction to item ii), unless $f(z)$ actually vanishes identically in D. This, then, is the only option.

iv) From the definition of poles, they too are isolated.

v) If $f(z)$ has a pole at $z = a$ then $f(z) \to \infty$ as $z \to a$ in any manner.

vi) Limit points of zeros II: Suppose we know that f is analytic in D, except possibly at $z = a$ which is limit point of zeros as in iii), but we also know that f is not identically zero. Then $z = a$ must be singularity of f — but not a pole (because f would tend to infinity and could not have arbitrarily close zeros) — so a must be an isolated essential singularity. For example $\sin 1/z$ has an isolated essential singularity at $z = 0$, this being a limit point of the zeros at $z = 1/n\pi$.

vii) A limit point of poles or other singularities would be a *non-isolated essential singularity*.

8.4.4 Analytic Continuation

Suppose that $f_1(z)$ is analytic in the (open, arcwise-connected) domain D_1, and $f_2(z)$ is analytic in D_2, with $D_1 \cap D_2 \neq \emptyset$. Suppose further that $f_1(z) = f_2(z)$ in $D_1 \cap D_2$. Then we say that f_2 is an analytic continuation of f_1 to

[6] A point z_0 is a limit point of a set S if for every $\epsilon > 0$ there is some $a \in S$, other than z_0 itself, such that $|a - z_0| \leq \epsilon$. A sequence need not have a limit for it to possess one or more limit points.

D_2. Such analytic continuations are *unique*: if f_3 is also analytic in D_2, and $f_3 = f_1$ in $D_1 \cap D_2$, then $f_2 - f_3 = 0$ in $D_1 \cap D_2$. Because the intersection of two open sets is also open, $f_1 - f_2$ vanishes on an open set and, so by observation iii) of the previous section, it vanishes everywhere in D_2.

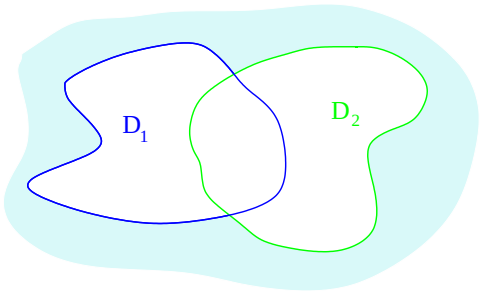

Figure 8.12: *Intersecting domains.*

We can use this uniqueness result, coupled with the circular domains of convergence of the Taylor series, to extend the definition of analytic functions beyond the domain of their initial definition.

The distribution $x_+^{\alpha-1}$

An interesting and useful example of analytic continuation is provided by the distribution $x_+^{\alpha-1}$, which, for real positive α, is defined by its evaluation on a test function $\varphi(x)$ as

$$(x_+^{\alpha-1}, \varphi) = \int_0^\infty x^{\alpha-1}\varphi(x)\,dx. \tag{8.122}$$

The pairing $(x_+^{\alpha-1}, \varphi)$ extends to an complex analytic function of α provided the integral converges. Test functions are required to decrease at infinity faster than any power of x, and so the integral always converges at the upper limit. It will converge at the lower limit provided $\mathrm{Re}\,(\alpha) > 0$. Assume that this is so, and integrate by parts using

$$\frac{d}{dx}\left(\frac{x^\alpha}{\alpha}\varphi(x)\right) = x^{\alpha-1}\varphi(x) + \frac{x^\alpha}{\alpha}\varphi'(x). \tag{8.123}$$

We find that, for $\epsilon > 0$,

$$\left[\frac{x^\alpha}{\alpha}\varphi(x)\right]_\epsilon^\infty = \int_\epsilon^\infty x^{\alpha-1}\varphi(x)\,dx + \int_\epsilon^\infty \frac{x^\alpha}{\alpha}\varphi'(x)\,dx. \tag{8.124}$$

The integrated-out part on the left-hand-side of (8.124) tends to zero as we take ϵ to zero, and both of the integrals converge in this limit as well. Consequently

$$I_1(\alpha) \equiv -\frac{1}{\alpha}\int_0^\infty x^\alpha \varphi'(x)\,dx \tag{8.125}$$

is equal to $(x_+^{\alpha-1}, \varphi)$ for $0 < \mathrm{Re}\,(\alpha) < \infty$. However, the integral defining $I_1(\alpha)$ converges in the larger region $-1 < \mathrm{Re}\,(\alpha) < \infty$. It therefore provides an analytic continuation to this larger domain. The factor of $1/\alpha$ reveals that the analytically-continued function possesses a pole at $\alpha = 0$, with residue

$$-\int_0^\infty \varphi'(x)\,dx = \varphi(0). \tag{8.126}$$

We can repeat the integration by parts, and find that

$$I_2(\alpha) \equiv \frac{1}{\alpha(\alpha+1)}\int_0^\infty x^{\alpha+1}\varphi''(x)\,dx \tag{8.127}$$

provides an analytic continuation to the region $-2 < \mathrm{Re}\,(\alpha) < \infty$. By proceeding in this manner, we can continue $(x_+^{\alpha-1}, \varphi)$ to a function analytic in the entire complex α plane with the exception of zero and the negative integers, at which it has simple poles. The residue of the pole at $\alpha = -n$ is $\varphi^{(n)}(0)/n!$.

There is another, much more revealing, way of expressing these analytic continuations. To obtain this, suppose that $\phi \in C^\infty[0,\infty]$ and $\phi \to 0$ at infinity as least as fast as $1/x$. (Our test function φ decreases much more rapidly than this, but $1/x$ is all we need for what follows.) Now the function

$$I(\alpha) \equiv \int_0^\infty x^{\alpha-1}\phi(x)\,dx \tag{8.128}$$

is convergent and analytic in the strip $0 < \mathrm{Re}\,(\alpha) < 1$. By the same reasoning as above, $I(\alpha)$ is there equal to

$$-\int_0^\infty \frac{x^\alpha}{\alpha}\phi'(x)\,dx. \tag{8.129}$$

Again this new integral provides an analytic continuation to the larger strip $-1 < \mathrm{Re}\,(\alpha) < 1$. But in the left-hand half of this strip, where $-1 <$

$\mathrm{Re}\,(\alpha) < 0$, we can write

$$\begin{aligned} -\int_0^\infty \frac{x^\alpha}{\alpha}\phi'(x)\,dx &= \lim_{\epsilon\to 0}\left\{\int_\epsilon^\infty x^{\alpha-1}\phi(x)\,dx - \left[\frac{x^\alpha}{\alpha}\phi(x)\right]_\epsilon^\infty\right\} \\ &= \lim_{\epsilon\to 0}\left\{\int_\epsilon^\infty x^{\alpha-1}\phi(x)\,dx + \phi(\epsilon)\frac{\epsilon^\alpha}{\alpha}\right\} \\ &= \lim_{\epsilon\to 0}\left\{\int_\epsilon^\infty x^{\alpha-1}[\phi(x)-\phi(\epsilon)]\,dx\right\}, \\ &= \int_0^\infty x^{\alpha-1}[\phi(x)-\phi(0)]\,dx. \end{aligned} \tag{8.130}$$

Observe how the integrated out part, which tends to zero in $0 < \mathrm{Re}\,(\alpha) < 1$, becomes divergent in the strip $-1 < \mathrm{Re}\,(\alpha) < 0$. This divergence is there craftily combined with the integral to cancel *its* divergence, leaving a finite remainder. As a consequence, for $-1 < \mathrm{Re}\,(\alpha) < 0$, the analytic continuation is given by

$$I(\alpha) = \int_0^\infty x^{\alpha-1}[\phi(x)-\phi(0)]\,dx. \tag{8.131}$$

Next we observe that $\chi(x) = [\phi(x)-\phi(0)]/x$ tends to zero as $1/x$ for large x, and at $x = 0$ can be defined by its limit as $\chi(0) = \phi'(0)$. This $\chi(x)$ then satisfies the same hypotheses as $\phi(x)$. With $I(\alpha)$ denoting the analytic continuation of the original I, we therefore have

$$\begin{aligned} I(\alpha) &= \int_0^\infty x^{\alpha-1}[\phi(x)-\phi(0)]\,dx, \qquad -1 < \mathrm{Re}\,(\alpha) < 0 \\ &= \int_0^\infty x^{\beta-1}\left[\frac{\phi(x)-\phi(0)}{x}\right]dx, \quad \text{where } \beta = \alpha+1, \\ &\to \int_0^\infty x^{\beta-1}\left[\frac{\phi(x)-\phi(0)}{x} - \phi'(0)\right]dx, \quad -1 < \mathrm{Re}\,(\beta) < 0 \\ &= \int_0^\infty x^{\alpha-1}[\phi(x)-\phi(0)-x\phi'(0)]\,dx, \quad -2 < \mathrm{Re}\,(\alpha) < -1, \end{aligned} \tag{8.132}$$

the arrow denoting the same analytic continuation process that we used with ϕ.

We can now apply this machinary to our original $\varphi(x)$, and so deduce

that the analytically-continued distribution is given by

$$(x_+^{\alpha-1},\varphi)=\begin{cases}\displaystyle\int_0^\infty x^{\alpha-1}\varphi(x)\,dx, & 0<\mathrm{Re}\,(\alpha)<\infty,\\ \displaystyle\int_0^\infty x^{\alpha-1}[\varphi(x)-\varphi(0)]\,dx, & -1<\mathrm{Re}\,(\alpha)<0,\\ \displaystyle\int_0^\infty x^{\alpha-1}[\varphi(x)-\varphi(0)-x\varphi'(0)]\,dx, & -2<\mathrm{Re}\,(\alpha)<-1,\end{cases} \tag{8.133}$$

and so on. The analytic continuation automatically subtracts more and more terms of the Taylor series of $\varphi(x)$ the deeper we penetrate into the left-hand half-plane. This property, that analytic continuation covertly subtracts the minimal number of Taylor-series terms required ensure convergence, lies behind a number of physics applications, most notably the method of *dimensional regularization* in quantum field theory.

The following exercise illustrates some standard techniques of reasoning *via* analytic continuation.

Exercise 8.7: Define the *dilogarithm* function by the series

$$\mathrm{Li}_2(z)=\frac{z}{1^2}+\frac{z^2}{2^2}+\frac{z^3}{3^2}+\cdots.$$

The radius of convergence of this series is unity, but the domain of $\mathrm{Li}_2(z)$ can be extended to $|z|>1$ by analytic continuation.

a) Observe that the series converges at $z=\pm1$, and at $z=1$ is

$$\mathrm{Li}_2(1)=1+\frac{1}{2^2}+\frac{1}{3^2}+\cdots=\frac{\pi^2}{6}.$$

Rearrange the series to show that

$$\mathrm{Li}_2(-1)=-\frac{\pi^2}{12}.$$

b) Identify the derivative of the power series for $\mathrm{Li}_2(z)$ with that of an elementary function. Exploit your identification to extend the definition of $[\mathrm{Li}_2(z)]'$ outside $|z|<1$. Use the properties of this derivative function, together with part a), to prove that

$$\mathrm{Li}_2(-z)+\mathrm{Li}_2\left(-\frac{1}{z}\right)=-\frac{1}{2}(\ln z)^2-\frac{\pi^2}{6}.$$

This formula allows us to calculate values of the dilogarithm for $|z|>1$ in terms of those with $|z|<1$.

Many weird identities involving dilogarithms exist. Some, such as

$$\mathrm{Li}_2\left(-\frac{1}{2}\right)+\frac{1}{6}\mathrm{Li}_2\left(\frac{1}{9}\right)=-\frac{1}{18}\pi^2+\ln 2\ln 3-\frac{1}{2}(\ln 2)^2-\frac{1}{3}(\ln 3)^2,$$

were found by Ramanujan. Others, originally discovered by sophisticated numerical methods, have been given proofs based on techniques from quantum mechanics. *Polylogarithms*, defined by

$$\mathrm{Li}_k(z)=\frac{z}{1^k}+\frac{z^2}{2^k}+\frac{z^3}{3^k}+\cdots,$$

occur frequently when evaluating Feynman diagrams.

8.4.5 Removable Singularities and the Weierstrass-Casorati Theorem

Sometimes we are given a definition that makes a function analytic in a region with the exception of a single point. Can we extend the definition to make the function analytic in the entire region? Provided that the function is well enough behaved near the point, the answer is yes, and the extension is unique. Curiously, the proof that this is so gives us insight into the wild behaviour of functions near essential singularities.

Removable singularities

Suppose that $f(z)$ is analytic in $D\setminus a$, but that $\lim_{z\to a}(z-a)f(z)=0$, then f may be extended to a function analytic in all of D — *i.e.* $z=a$ is a *removable singularity*. To see this, let ζ lie between two simple closed contours Γ_1 and Γ_2, with a within the smaller, Γ_2. We use Cauchy to write

$$f(\zeta)=\frac{1}{2\pi i}\oint_{\Gamma_1}\frac{f(z)}{z-\zeta}\,dz-\frac{1}{2\pi i}\oint_{\Gamma_2}\frac{f(z)}{z-\zeta}\,dz. \tag{8.134}$$

Now we can shrink Γ_2 down to be very close to a, and because of the condition on $f(z)$ near $z=a$, we see that the second integral vanishes. We can also arrange for Γ_1 to enclose any chosen point in D. Thus, if we set

$$\tilde{f}(\zeta)=\frac{1}{2\pi i}\oint_{\Gamma_1}\frac{f(z)}{z-\zeta}\,dz \tag{8.135}$$

within Γ_1, we see that $\tilde{f}=f$ in $D\setminus a$, and is analytic in all of D. The extension is unique because any two analytic functions that agree everywhere except for a single point, must also agree at that point.

Weierstrass-Casorati

We apply the idea of removable singularities to show just how pathological a beast is an isolated essential singularity:
Theorem (Weierstrass-Casorati): Let $z = a$ be an isolated essential singularity of $f(z)$, then in any neighbourhood of a the function $f(z)$ comes arbitrarily close to any assigned valued in $\mathbb{C}$.

To prove this, define $N_\delta(a) = \{z \in \mathbb{C} : |z - a| < \delta\}$, and $N_\epsilon(\zeta) = \{z \in \mathbb{C} : |z - \zeta| < \epsilon\}$. The claim is then that there is an $z \in N_\delta(a)$ such that $f(z) \in N_\epsilon(\zeta)$. Suppose that the claim is *not* true. Then we have $|f(z) - \zeta| > \epsilon$ for all $z \in N_\delta(a)$. Therefore

$$\left| \frac{1}{f(z) - \zeta} \right| < \frac{1}{\epsilon} \tag{8.136}$$

in $N_\delta(a)$, while $1/(f(z) - \zeta)$ is analytic in $N_\delta(a) \setminus a$. Therefore $z = a$ is a removable singularity of $1/(f(z) - \zeta)$, and there is an an analytic $g(z)$ which coincides with $1/(f(z) - \zeta)$ at all points except a. Therefore

$$f(z) = \zeta + \frac{1}{g(z)} \tag{8.137}$$

except at a. Now $g(z)$, being analytic, may have a zero at $z = a$ giving a pole in f, but it cannot give rise to an essential singularity. The claim is true, therefore.

Picard's Theorems

Weierstrass-Casorati is elementary. There are much stronger results:
Theorem (Picard's little theorem): Every nonconstant entire function attains every complex value with at most one exception.
Theorem (Picard's big theorem): In any neighbourhood of an isolated essential singularity, $f(z)$ takes every complex value with at most one exception.
The proofs of these theorems are hard.

As an illustration of Picard's little theorem, observe that the function $\exp z$ is entire, and takes all values except 0. For the big theorem observe that function $f(z) = \exp(1/z)$. has an essential singularity at $z = 0$, and takes all values, with the exception of 0, in any neighbourhood of $z = 0$.

8.5 Meromorphic functions and the Winding-Number

A function whose only singularities in D are poles is said to be *meromorphic* there. These functions have a number of properties that are essentially topological in character.

8.5.1 Principle of the Argument

If $f(z)$ is meromorphic in D with $\partial D = \Gamma$, and $f(z) \neq 0$ on Γ, then

$$\frac{1}{2\pi i} \oint_\Gamma \frac{f'(z)}{f(z)}\, dz = N - P \tag{8.138}$$

where N is the number of zero's in D and P is the number of poles. To show this, we note that if $f(z) = (z-a)^m \varphi(z)$ where φ is analytic and non-zero near a, then

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{\varphi'(z)}{\varphi(z)} \tag{8.139}$$

so f'/f has a simple pole at a with residue m. Here m can be either positive or negative. The term $\varphi'(z)/\varphi(z)$ is analytic at $z = a$, so collecting all the residues from each zero or pole gives the result.

Since $f'/f = \frac{d}{dz} \ln f$ the integral may be written

$$\oint_\Gamma \frac{f'(z)}{f(z)}\, dz = \Delta_\Gamma \ln f(z) = i \Delta_\Gamma \arg f(z), \tag{8.140}$$

the symbol Δ_Γ denoting the total change in the quantity after we traverse Γ. Thus

$$N - P = \frac{1}{2\pi} \Delta_\Gamma \arg f(z). \tag{8.141}$$

This result is known as the principle of the argument.

Local mapping theorem

Suppose the function $w = f(z)$ maps a region Ω holomorphicly onto a region Ω', and a simple closed curve $\gamma \subset \Omega$ onto another closed curve $\Gamma \subset \Omega'$, which will in general have self intersections. Given a point $a \in \Omega'$, we can ask

ourselves how many points within the simple closed curve γ map to a. The answer is given by the *winding number* of the image curve Γ about a.

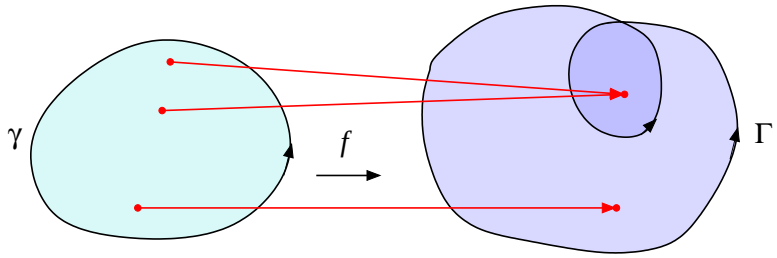

Figure 8.13: *An analytic map is one-to-one where the winding number is unity, but two-to-one at points where the image curve winds twice.*

To that this is so, we appeal to the principal of the argument as

$$\begin{aligned} \#\text{ of zeros of } (f-a) \text{ within } \gamma &= \frac{1}{2\pi i}\oint_\gamma \frac{f'(z)}{f(z)-a}\,dz, \\ &= \frac{1}{2\pi i}\oint_\Gamma \frac{dw}{w-a}, \\ &= n(\Gamma, a), \end{aligned} \tag{8.142}$$

where $n(\Gamma, a)$ is called the winding number of the image curve Γ about a. It is equal to

$$n(\Gamma, a) = \frac{1}{2\pi}\Delta_\gamma \arg(w-a), \tag{8.143}$$

and is the number of times the image point w encircles a as z traverses the original curve γ.

Since the number of pre-image points cannot be negative, these winding numbers must be positive. This means that the holomorphic image of curve winding in the anticlockwise direction is also a curve winding anticlockwise.

For mathematicians, another important consequence of this result is that a holomorphic map is *open*– *i.e.* the holomorphic image of an open set is itself an open set. The local mapping theorem is therefore sometime called the *open mapping theorem*.

8.5.2 Rouché's theorem

Here we provide an effective tool for locating zeros of functions.

Theorem (Rouché): Let $f(z)$ and $g(z)$ be analytic within and on a simple closed contour γ. Suppose further that $|g(z)| < |f(z)|$ everywhere on γ, then $f(z)$ and $f(z) + g(z)$ have the same number of zeros within γ.

Before giving the proof, we illustrate Rouché's theorem by giving its most important corollary: the algebraic completeness of the complex numbers, a result otherwise known as the *fundamental theorem of algebra*. This asserts that, if R is sufficiently large, a polynomial $P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0$ has exactly n zeros, when counted with their multiplicity, lying within the circle $|z| = R$. To prove this note that we can take R sufficiently big that

$$\begin{aligned} |a_n z^n| &= |a_n| R^n \\ &> |a_{n-1}| R^{n-1} + |a_{n-2}| R^{n-2} \cdots + |a_0| \\ &> |a_{n-a} z^{n-1} + a_{n-2} z^{n-2} \cdots + a_0|, \end{aligned} \tag{8.144}$$

on the circle $|z| = R$. We can therefore take $f(z) = a_n z^n$ and $g(z) = a_{n-a} z^{n-1} + a_{n-2} z^{n-2} \cdots + a_0$ in Rouché. Since $a_n z^n$ has exactly n zeros, all lying at $z = 0$, within $|z| = R$, we conclude that so does $P(z)$.

The proof of Rouché is a corollary of the principle of the argument. We observe that

$$\begin{aligned} \#\text{ of zeros of } f+g &= n(\Gamma, 0) \\ &= \frac{1}{2\pi} \Delta_\gamma \arg(f+g) \\ &= \frac{1}{2\pi i} \Delta_\gamma \ln(f+g) \\ &= \frac{1}{2\pi i} \Delta_\gamma \ln f + \frac{1}{2\pi i} \Delta_\gamma \ln(1+g/f) \\ &= \frac{1}{2\pi} \Delta_\gamma \arg f + \frac{1}{2\pi} \Delta_\gamma \arg(1+g/f). \end{aligned} \tag{8.145}$$

Now $|g/f| < 1$ on γ, so $1 + g/f$ cannot circle the origin as we traverse γ. As a consequence $\Delta_\gamma \arg(1 + g/f) = 0$. Thus the number of zeros of $f + g$ inside γ is the same as that of f alone. (Naturally, they are not usually in the same places.)

The geometric part of this argument is often illustrated by a dog on a lead. If the lead has length L, and the dog's owner stays a distance $R > L$ away from a lamp post, then the dog cannot run round the lamp post unless the owner does the same.

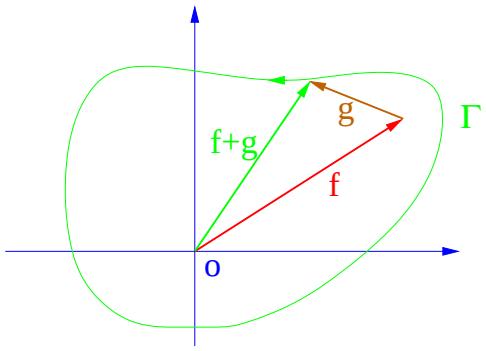

Figure 8.14: *The curve* Γ *is the image of* γ *under the map* $f+g$*. If* $|g|<|f|$*, then, as* z *traverses* γ*,* $f+g$ *winds about the origin the same number of times that* f *does.*

Exercise 8.8: *Jacobi Theta Function.* The function $\theta(z|\tau)$ is defined for $\mathrm{Im}\,\tau > 0$ by the sum

$$\theta(z|\tau) = \sum_{n=-\infty}^{\infty} e^{i\pi\tau n^2} e^{2\pi i n z}.$$

Show that $\theta(z+1|\tau) = \theta(z|\tau)$, and $\theta(z+\tau|\tau) = e^{-i\pi\tau - 2\pi i z}\theta(z|\tau)$. Use this information and the principle of the argument to show that $\theta(z|\tau)$ has exactly one zero in each unit cell of the Bravais lattice comprising the points $z = m + n\tau$; $m, n \in \mathbb{Z}$. Show that these zeros are located at $z = (m+1/2) + (n+1/2)\tau$.

Exercise 8.9: Use Rouché's theorem to find the number of roots of the equation $z^5 + 15z + 1 = 0$ lying within the circles, i) $|z| = 2$, ii) $|z| = 3/2$.

8.6 Analytic Functions and Topology

8.6.1 The Point at Infinity

Some functions, $f(z) = 1/z$ for example, tend to a fixed limit (here 0) as z become large, independently of in which direction we set off towards infinity. Others, such as $f(z) = \exp z$, behave quite differently depending on what direction we take as $|z|$ becomes large.

To accommodate the former type of function, and to be able to legitimately write $f(\infty) = 0$ for $f(z) = 1/z$, it is convenient to add "∞" to the set of complex numbers. Technically, what we are doing is to constructing

the *one-point compactification* of the locally compact space $\mathbb{C}$. We often portray this extended complex plane as a sphere S^2 (the Riemann sphere), using stereographic projection to locate infinity at the north pole, and 0 at the south pole.

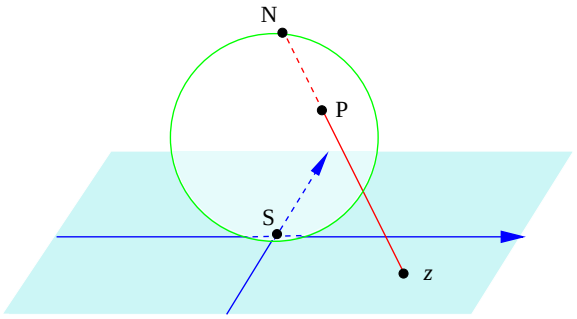

Figure 8.15: *Stereographic mapping of the complex plane to the 2-Sphere.*

By the phrase a *neighbourhood* of z, we mean an open set containing z. We use the stereographic map to define a *neighbourhood of infinity* as the stereographic image of a neighbourhood of the north pole. With this definition, the extended complex plane $\mathbb{C} \cup \{\infty\}$ becomes topologically a sphere, and in particular, becomes a compact set.

If we wish to study the behaviour of a function "at infinity," we use the map $z \mapsto \zeta = 1/z$ to bring ∞ to the origin, and study the behaviour of the function there. Thus the polynomial

$$f(z) = a_0 + a_1 z + \cdots + a_N z^N \tag{8.146}$$

becomes

$$f(\zeta) = a_0 + a_1 \zeta^{-1} + \cdots + a_N \zeta^{-N}, \tag{8.147}$$

and so has a pole of order N at infinity. Similarly, the function $f(z) = z^{-3}$ has a zero of order three at infinity, and $\sin z$ has an isolated essential singularity there.

We must be a careful about defining *residues* at infinity. The residue is more a property of the 1-form $f(z)\,dz$ than of the function $f(z)$ alone, and to find the residue we need to transform the dz as well as $f(z)$. For example, if we set $z = 1/\zeta$ in dz/z we have

$$\frac{dz}{z} = \zeta\, d\left(\frac{1}{\zeta}\right) = -\frac{d\zeta}{\zeta}, \tag{8.148}$$

so the 1-form $(1/z)\,dz$ has a pole at $z = 0$ with residue 1, and has a pole with residue -1 at infinity—even though the *function* $1/z$ has no pole there. This 1-form viewpoint is required for compatability with the residue theorem: The integral of $1/z$ around the positively oriented unit circle is simultaneously minus the integral of $1/z$ about the oppositely oriented unit circle, now regarded as a a positively oriented circle enclosing the point at infinity. Thus if $f(z)$ has of pole of order N at infinity, and

$$\begin{aligned} f(z) &= \cdots + a_{-2}z^{-2} + a_{-1}z^{-1} + a_0 + a_1 z + a_2 z^2 + \cdots + A_N z^N \\ &= \cdots + a_{-2}\zeta^{2} + a_{-1}\zeta + a_0 + a_1\zeta^{-1} + a_2\zeta^{-2} + \cdots + A_N\zeta^{-N} \end{aligned} \tag{8.149}$$

near infinity, then the residue at infinity must be defined to be $-a_{-1}$, and not a_1 as one might naïvely have thought.

Once we have allowed ∞ as a point in the set we map *from*, it is only natural to add it to the set we map *to* — in other words to allow ∞ as a possible value for $f(z)$. We will set $f(a) = \infty$, if $|f(z)|$ becomes unboundedly large as $z \to a$ in any manner. Thus, if $f(z) = 1/z$ we have $f(0) = \infty$.

The map

$$w = \left(\frac{z - z_0}{z - z_\infty}\right)\left(\frac{z_1 - z_\infty}{z_1 - z_0}\right) \tag{8.150}$$

takes

$$\begin{aligned} z_0 &\to 0, \\ z_1 &\to 1, \\ z_\infty &\to \infty, \end{aligned} \tag{8.151}$$

for example. Using this language, the Möbius maps

$$w = \frac{az + b}{cz + d} \tag{8.152}$$

become one-to-one maps of $S^2 \to S^2$. They are the only such globally conformal one-to-one maps. When the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{8.153}$$

is an element of SU(2), the resulting one–to-one map is a rigid rotation of the Riemann sphere. Stereographic projection is thus revealed to be the geometric origin of the spinor representations of the rotation group.

If an analytic function $f(z)$ has no essential singularities anywhere on the Riemann sphere then f is *rational*, meaning that it can be written as $f(z) = P(z)/Q(z)$ for some polynomials P, Q.

We begin the proof of this fact by observing that $f(z)$ can have only a finite number of poles. If, to the contrary, f had an infinite number of poles then the compactness of S^2 would ensure that the poles would have a limit point somewhere. This would be a non-isolated singularity of f, and hence an essential singularity. Now suppose we have poles at z_1, z_2, ..., z_N with principal parts

$$\sum_{m=1}^{m_n} \frac{b_{n,m}}{(z-z_n)^m}.$$

If one of the z_n is ∞, we first use a Möbius map to move it to some finite point. Then

$$F(z) = f(z) - \sum_{n=1}^{N} \sum_{m=1}^{m_n} \frac{b_{n,m}}{(z-z_n)^m} \tag{8.154}$$

is everywhere analytic, and therefore continuous, on S^2. But S^2 being compact and $F(z)$ being continuous implies that F is bounded. Therefore, by Liouville's theorem, it is a constant. Thus

$$f(z) = \sum_{n=1}^{N} \sum_{m=1}^{m_n} \frac{b_{n,m}}{(z-z_n)^m} + C, \tag{8.155}$$

and this is a rational function. If we made use of a Möbius map to move a pole at infinity, we use the inverse map to restore the original variables. This manoeuvre does not affect the claimed result because Möbius maps take rational functions to rational functions.

The map $z \mapsto f(z)$ given by the rational function

$$f(z) = \frac{P(z)}{Q(z)} = \frac{a_n z^n + a_{n-1} z^{n-1} + \cdots a_0}{b_n z^n + b_{n-1} z^{n-1} + \cdots b_0} \tag{8.156}$$

wraps the Riemann sphere n times around the target S^2. In other words, it is a n-to-one map.

8.6.2 Logarithms and Branch Cuts

The function $y = \ln z$ is defined to be the solution to $z = \exp y$. Unfortunately, since $\exp 2\pi i = 1$, the solution is not unique: if y is a solution, so is

$y + 2\pi i$. Another way of looking at this is that if $z = \rho \exp i\theta$, with ρ real, then $y = \ln \rho + i\theta$, and the angle θ has the same $2\pi i$ ambiguity. Now there is no such thing as a "many valued function." By definition, a function is a machine into which we plug something and get a unique output. To make $\ln z$ into a legitimate function we must select a unique $\theta = \arg z$ for each z. This can be achieved by cutting the z plane along a curve extending from the the *branch point* at $z = 0$ all the way to infinity. Exactly where we put this *branch cut* is not important; what *is* important is that it serve as an impenetrable fence preventing us from following the continuous evolution of the function along a path that winds around the origin.

Similar branch cuts serve to make fractional powers single valued. We define the power z^α for for non-integral α by setting

$$z^\alpha = \exp\{\alpha \ln z\} = |z|^\alpha e^{i\alpha\theta}, \tag{8.157}$$

where $z = |z|e^{i\theta}$. For the square root $z^{1/2}$ we get

$$z^{1/2} = \sqrt{|z|}e^{i\theta/2}, \tag{8.158}$$

where $\sqrt{|z|}$ represents the *positive* square root of $|z|$. We can therefore make this single-valued by a cut from 0 to ∞. To make $\sqrt{(z-a)(z-b)}$ single valued we only need to cut from a to b. (Why? — think this through!).

We can get away without cuts if we imagine the functions being maps *from* some set other than the complex plane. The new set is called a *Riemann surface*. It consists of a number of copies of the complex plane, one for each possible value of our "multivalued function." The map from this new surface is then single-valued, because each possible value of the function is the value of the function evaluated at a point on a different copy. The copies of the complex plane are called *sheets*, and are connected to each other in a manner dictated by the function. The cut plane may now be thought of as a drawing of one level of the multilayered Riemann surface. Think of an architect's floor plan of a spiral-floored multi-story car park: If the architect starts drawing at one parking spot and works her way round the central core, at some point she will find that the floor has become the ceiling of the part already drawn. The rest of the structure will therefore have to be plotted on the plan of the next floor up — but exactly where she draws the division between one floor and the one above is rather arbitrary. The spiral car-park is a good model for the Riemann surface of the $\ln z$ function. See figure 8.16.

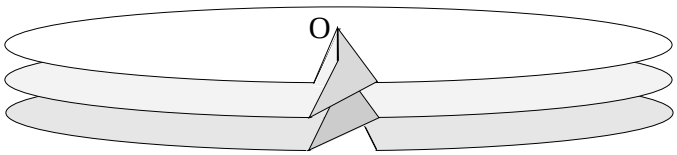

Figure 8.16: *Part of the Riemann surface for* $\ln z$*. Each time we circle the origin, we go up one level.*

To see what happens for a square root, follow $z^{1/2}$ along a curve circling the branch point singularity at $z = 0$. We come back to our starting point with the function having changed sign; A second trip along the same path would bring us back to the original value. The square root thus has only two sheets, and they are cross-connected as shown in figure 8.17.

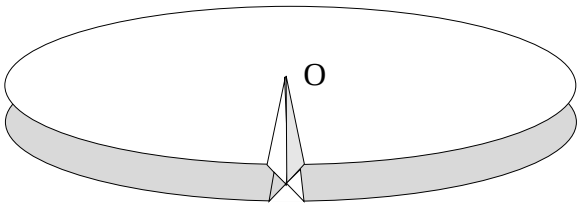

Figure 8.17: *Part of the Riemann surface for* $\sqrt{z}$*. Two copies of* $\mathbb{C}$ *are cross-connected. Circling the origin once takes you to the lower level. A second circuit brings you back to the upper level.*

In figures 8.16 and 8.17, we have shown the cross-connections being made rather abruptly along the cuts. This is not necessary —there is no singularity in the function at the cut — but it is often a convenient way to think about the structure of the surface. For example, the surface for $\sqrt{(z-a)(z-b)}$ also consists of two sheets. If we include the point at infinity, this surface can be thought of as two spheres, one inside the other, and cross connected along the cut from a to b.

8.6.3 Topology of Riemann surfaces

Riemann surfaces often have interesting topology. Indeed much of modern algebraic topology emerged from the need to develop tools to understand multiply-connected Riemann surfaces. As we have seen, the complex numbers, with the point at infinity included, have the topology of a sphere. The

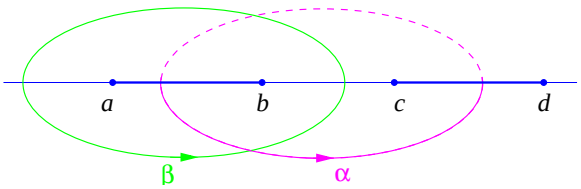

Figure 8.18: *The 1-cycles α and β on the plane with two square-root branch cuts. The dashed part of α lies hidden on the second sheet of the Riemann surface.*

$\sqrt{(z-a)(z-b)}$ surface is still topologically a sphere. To see this imagine continuously deforming the Riemann sphere by pinching it at the equator down to a narrow waist. Now squeeze the front and back of the waist together and (imagining that the the surface can pass freely through itself) fold the upper half of the sphere inside the lower. The result is the precisely the two-sheeted $\sqrt{(z-a)(z-b)}$ surface described above. The Riemann surface of the function $\sqrt{(z-a)(z-b)(z-c)(z-d)}$, which can be thought of a two spheres, one inside the other and connected along two cuts, one from a to b and one from c to d, is, however, a *torus*. Think of the torus as a bicycle inner tube. Imagine using the fingers of your left hand to pinch the front and back of the tube together and the fingers of your right hand to do the same on the diametrically opposite part of the tube. Now fold the tube about the pinch lines through itself so that one half of the tube is inside the other, and connected to the outer half through two square-root cross-connects. If you have difficulty visualizing this process, figures 8.18 and 8.19 show how the two 1-cycles, α and β, that generate the homology group $H_1(T^2)$ appear when drawn on the plane cut from a to b and c to d, and then when drawn on the torus. Observe, in figure 8.18, how the curves in the two-sheeted plane manage to intersect in only one point, just as they do when drawn on the torus in figure 8.19.

That the topology of the twice-cut plane is that of a torus has important consequences. This is because the *elliptic integral*

$$w = I^{-1}(z) = \int_{z_0}^{z} \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}} \tag{8.159}$$

maps the twice-cut z-plane 1-to-1 onto the torus, the latter being considered as the complex w-plane with the points w and $w + n\omega_1 + m\omega_2$ identified. The

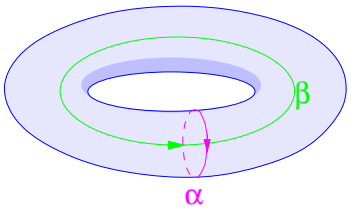

Figure 8.19: *The 1-cycles α and β on the torus.*

two numbers $\omega_{1,2}$ are given by

$$\begin{aligned} \omega_1 &= \oint_\alpha \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}}, \\ \omega_2 &= \oint_\beta \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}}, \end{aligned} \tag{8.160}$$

and are called the *periods* of the *elliptic function* $z = I(w)$. The map $w \mapsto z = I(w)$ is a genuine function because the original z is uniquely determined by w. It is *doubly periodic* because

$$I(w + n\omega_1 + m\omega_2) = I(w), \quad n, m \in \mathbb{Z}. \tag{8.161}$$

The inverse "function" $w = I^{-1}(z)$ is not a genuine function of z, however, because w increases by ω_1 or ω_2 each time z goes around a curve deformable into α or β, respectively. The periods are complicated functions of a, b, c, d.

If you recall our discussion of de Rham's theorem from chapter 4, you will see that the ω_i are the results of pairing the closed holomorphic 1-form.

$$\text{"}dw\text{"} = \frac{dz}{\sqrt{(z-a)(z-b)(z-c)(z-d)}} \in H^1(T^2) \tag{8.162}$$

with the two generators of $H_1(T^2)$. The quotation marks about dw are there to remind us that dw is not an exact form, *i.e.* it is not the exterior derivative of a single-valued function w. This cohomological interpretation of the periods of the elliptic function is the origin of the use of the word "period" in the context of de Rham's theorem. (See section 10.5 for more information on elliptic functions.)

More general Riemann surfaces are oriented 2-manifolds that can be thought of as the surfaces of doughnuts with g holes. The number g is called

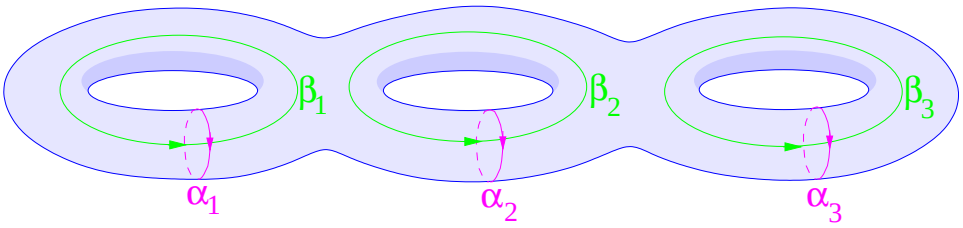

Figure 8.20: *A surface M of genus 3. The non-bounding 1-cycles α_i and β_i form a basis of $H_1(M)$. The entire surface forms the single 2-cycle that spans $H_2(M)$.*

the *genus* of the surface. The sphere has $g = 0$ and the torus has $g = 1$. The Euler character of the Riemann surface of genus g is $\chi = 2(1 - g)$. For example, figure 8.20 shows a surface of genus three. The surface is in one piece, so $\dim H_0(M) = 1$. The other Betti numbers are $\dim H_1(M) = 6$ and $\dim H_2(M) = 1$, so

$$\chi = \sum_{p=0}^{2} (-1)^p \dim H_p(M) = 1 - 6 + 1 = -4, \tag{8.163}$$

in agreement with $\chi = 2(1 - 3) = -4$. For complicated functions, the genus may be infinite.

If we have two complex variables z and w then a polynomial relation $P(z, w) = 0$ defines a *complex algebraic curve*. Except for degenerate cases, this one (complex) dimensional curve is simultaneously a two (real) dimensional Riemann surface. With

$$P(z, w) = z^3 + 3w^2 z + w + 3 = 0, \tag{8.164}$$

for example, we can think of $z(w)$ being a three-sheeted function of w defined by solving this cubic. Alternatively we can consider $w(z)$ to be the two-sheeted function of z obtained by solving the quadratic equation

$$w^2 + \frac{1}{3z} w + \frac{(3 + z^3)}{3z} = 0. \tag{8.165}$$

In each case the branch points will be located where two or more roots coincide. The roots of (8.165), for example, coincide when

$$1 - 12z(3 + z^3) = 0. \tag{8.166}$$

This quartic equation has four solutions, so there are four square-root branch points. Although constructed differently, the Riemann surface for $w(z)$ and the Riemann surface for $z(w)$ will have the same genus (in this case $g = 1$) because they are really are one and the same object — the algebraic curve defined by the original polynomial equation.

In order to capture all its points at infinity, we often consider a complex algebraic curve as being a subset of $\mathbb{C}P^2$. To do this we make the defining equation homogeneous by introducing a third co-ordinate. For example, for (8.164) we make

$$P(z,w) = z^3 + 3w^2z + w + 3 \rightarrow P(z,w,v) = z^3 + 3w^2z + wv^2 + 3v^3. \quad (8.167)$$

The points where $P(z,w,v) = 0$ define[7] a *projective curve* lying in $\mathbb{C}P^2$. Places on this curve where the co-ordinate v is zero are the added points at infinity. Places where v is non-zero (and where we may as well set $v = 1$) constitute the original *affine curve*.

A generic (non-singular) curve

$$P(z,w) = \sum_{r,s} a_{rs} z^r w^s = 0, \quad (8.168)$$

with its points at infinity included, has genus

$$g = \frac{1}{2}(d-1)(d-2). \quad (8.169)$$

Here $d = \max(r+s)$ is the *degree* of the curve. This *degree-genus* relation is due to Plücker. It is not, however, trivial to prove. Also not easy to prove is Riemann's theorem of 1852 that *any* finite genus Riemann surface is the complex algebraic curve associated with some two-variable polynomial.

The two assertions in the previous paragraph seem to contradict each other. "Any" finite genus, must surely include $g = 2$, but how can a genus two surface be a complex algebraic curve? There is no integer value of d such that $(d-1)(d-2)/2 = 2$. This is where the "non-singular" caveat becomes important. An affine curve $P(z,w) = 0$ is said to be *singular* at P $= (z_0, w_0)$ if all of

$$P(z,w), \quad \frac{\partial P}{\partial z}, \quad \frac{\partial P}{\partial w},$$

[7] A homogeneous polynomial $P(z,w,v)$ of degree n does not provide a map from $\mathbb{C}P^2 \rightarrow \mathbb{C}$ because $P(\lambda z, \lambda w, \lambda v) = \lambda^n P(z,w,v)$ usually depends on λ, while the co-ordinates $(\lambda z, \lambda w, \lambda v)$ and (z,w,v) correspond to the same point in $\mathbb{C}P^2$. The *zero set* where $P = 0$ is, however, well-defined in $\mathbb{C}P^2$.

vanish at P. A projective curve is singular at $\mathrm{P} \in \mathbb{C}P^2$ if all of

$$P(z,w,v), \quad \frac{\partial P}{\partial z}, \quad \frac{\partial P}{\partial w}, \quad \frac{\partial P}{\partial v}$$

are zero there. If the curve has a singular point then then it degenerates and ceases to be a manifold. Now Riemann's construction does not guarantee an *embedding* of the surface into $\mathbb{C}P^2$, only an *immersion*. The distinction between these two concepts is that an immersed surface is allowed to self-intersect, while an embedded one is not. Being a double root of the defining equation $P(z,w) = 0$, a point of self-intersection is necessarily a singular point.

As an illustration of a singular curve, consider our earlier example of the curve

$$w^2 = (z-a)(z-b)(z-c)(z-d) \tag{8.170}$$

whose Riemann surface we know to be a torus once two some points are added at infinity, and when a, b, c, d are all distinct. The degree-genus formula applied to this degree four curve gives, however, $g = 3$ instead of the expected $g = 1$. This is because the corresponding projective curve

$$w^2v^2 = (z-av)(z-bv)(z-cv)(z-dv) \tag{8.171}$$

has a *tacnode* singularity at the point $(z,w,v) = (0,1,0)$. Rather than investigate this rather complicated singularity at infinity, we will consider the simpler case of what happens if we allow b to coincide with c. When b and c merge, the finite point $\mathrm{P} = (w_0, z_0) = (0, b)$ becomes a singular. Near the singularity, the equation defining our curve looks like

$$0 = w^2 - ad\,(z-b)^2, \tag{8.172}$$

which is the equation of two lines, $w = \sqrt{ad}\,(z-b)$ and $w = -\sqrt{ad}\,(z-b)$, that intersect at the point $(w,z) = (0,b)$. To understand what is happening topologically it is first necessary to realize that a *complex* line is a copy of $\mathbb{C}$ and hence, after the point at infinity is included, is topologically a sphere. A pair of intersecting complex lines is therefore topologically a pair of spheres sharing a common point. Our degenerate curve only looks like a pair of lines near the point of intersection however. To see the larger picture, look back at the figure of the twice-cut plane where we see that as b approaches c we have an α cycle of zero total length. A zero length cycle means that

the circumference of the torus becomes zero at P, so that it looks like a bent sausage with its two ends sharing the common point P. Instead of two separate spheres, our sausage is equivalent to a single two-sphere with two points identified.

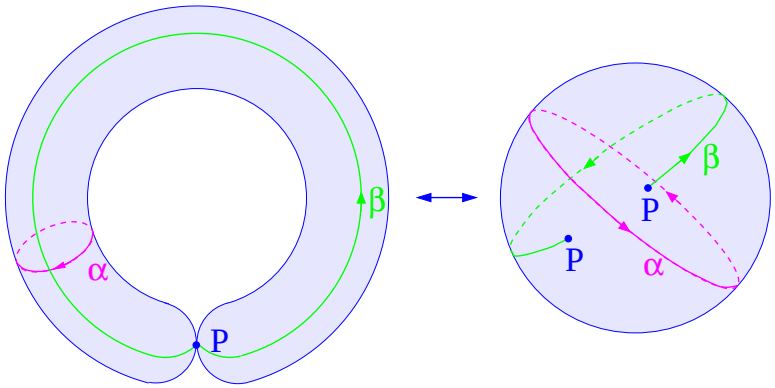

Figure 8.21: *A degenerate torus is topologically the same as a sphere with two points identified.*

As it stands, such a set is no longer a manifold because any neighbourhood of P will contain bits of both ends of the sausage, and therefore cannot be given co-ordinates that make it look like a region in $\mathbb{R}^2$. We can, however, simply agree to delete the common point, and then plug the resulting holes in the sausage ends with two distinct points. The new set is again a manifold, and topologically a sphere. From the viewpoint of the pair of intersecting lines, this construction means that we stay on one line, and ignore the other as it passes through.

A similar *resolution of singularities* allows us to regard immersed surfaces as non-singular manifolds, and it is this sense that Riemann's theorem is to be understood. When n such self-intersection double points are deleted and replaced by pairs of distinct points The degree-genus formula becomes

$$g = \frac{1}{2}(d-1)(d-2) - n, \tag{8.173}$$

and this can take any integer value.

8.6.4 Conformal geometry of Riemann surfaces

In this section we recall Hodge's theory of harmonic forms from section 4.7.1, and see how it looks from a complex variable perspective. This viewpoint reveals a relationship between Riemann surfaces and Riemann manifolds that forms an important ingredient in string and conformal field theory.

Isothermal co-ordinates and complex structure

Suppose we have a two-dimensional orientable Riemann manifold M with metric

$$ds^2 = g_{ij}\, dx^i dx^j. \tag{8.174}$$

In two dimensions g_{ij} has three independent components. When we make a co-ordinate transformation we have two arbitrary functions at our disposal, and so we can use this freedom to select local co-ordinates in which only one independent component remains. The most useful choice is *isothermal* (also called *conformal*) co-ordinates x, y in which the metric tensor is diagonal, $g_{ij} = e^{\sigma}\delta_{ij}$, and so

$$ds^2 = e^{\sigma}(dx^2 + dy^2). \tag{8.175}$$

The e^{σ} is called the *scale factor* or *conformal factor*. If we set $z = x + iy$ and $\overline{z} = x - iy$ the metric becomes

$$ds^2 = e^{\sigma(z,\overline{z})} d\overline{z}\, dz. \tag{8.176}$$

We can construct isothermal co-ordinates for some open neighbourhood of any point in M. If in an overlapping isothermal co-ordinate patch the metric is

$$ds^2 = e^{\tau(\zeta,\overline{\zeta})} d\overline{\zeta}\, d\zeta, \tag{8.177}$$

and if the co-ordinates have the same orientation, then in the overlap region ζ must be a function only of z and $\overline{\zeta}$ a function only of $\overline{z}$. This is so that

$$e^{\tau(\zeta,\overline{\zeta})} d\overline{\zeta}\, d\zeta = e^{\sigma(z,\overline{z})} \left|\frac{dz}{d\zeta}\right|^2 d\overline{\zeta}\, d\zeta \tag{8.178}$$

without any $d\zeta^2$ or $d\overline{\zeta}^2$ terms appearing. A manifold with an atlas of complex charts whose change-of-co-ordinate formulae are holomorphic in this way is said to be a *complex manifold*, and the co-ordinates endow it with a *complex*

structure. The existence of a global complex structure allows to us to define the notion of meromorphic and rational functions on M. Our Riemann manifold is therefore also a Riemann surface.

While any compact, orientable, two-dimensional Riemann manifold has a complex structure that is determined by the metric, the mapping: *metric* $\rightarrow$ *complex structure* is not one-to-one. Two metrics g_{ij}, $\tilde{g}_{ij}$ that are related by a conformal scale factor

$$g_{ij} = \lambda(x^1, x^2)\tilde{g}_{ij} \tag{8.179}$$

give rise to the same complex structure. Conversely, a pair of two-dimensional Riemann manifolds having the same complex structure have metrics that are related by a scale factor.

The use of isothermal co-ordinates simplifies many computations. Firstly, observe that $g^{ij}/\sqrt{g} = \delta_{ij}$, the conformal factor having cancelled. If you look back at its definition, you will see that this means that when the Hodge "$\star$" map acts on one-forms, the result is independent of the metric. If ω is a one-form

$$\omega = p\,dx + q\,dy, \tag{8.180}$$

then

$$\star\omega = -q\,dx + p\,dy. \tag{8.181}$$

Note that, on one-forms,

$$\star\star = -1. \tag{8.182}$$

With $z = x + iy$, $\overline{z} = x - iy$, we have

$$\omega = \frac{1}{2}(p - iq)\,dz + \frac{1}{2}(p + iq)\,d\overline{z}. \tag{8.183}$$

Let us focus on the dz part:

$$A = \frac{1}{2}(p - iq)\,dz = \frac{1}{2}(p - iq)(dx + idy). \tag{8.184}$$

Then

$$\star A = \frac{1}{2}(p - iq)(dy - idx) = -iA. \tag{8.185}$$

Similarly, with

$$B = \frac{1}{2}(p + iq)\,d\overline{z} \tag{8.186}$$

we have

$$\star B = iB. \tag{8.187}$$

Thus the dz and $d\overline{z}$ parts of the original form are separately eigenvectors of $\star$ with different eigenvalues. We use this observation to construct a resolution of the identity Id into the sum of two projection operators

$$\begin{aligned} Id &= \frac{1}{2}(1+i\star) + \frac{1}{2}(1-i\star), \\ &= \quad P \quad + \quad \overline{P}, \end{aligned} \tag{8.188}$$

where P projects on the dz part and $\overline{P}$ onto the $d\overline{z}$ part of the form.

The original form is harmonic if it is both closed $d\omega = 0$, and co-closed $d\star\omega = 0$. Thus, in two dimensions, the notion of being harmonic (*i.e.* a solution of Laplace's equation) is independent of what metric we are given. If ω is a harmonic form, then $(p-iq)dz$ and $(p+iq)d\overline{z}$ are separately closed. Observe that $(p-iq)dz$ being closed means that $\partial_{\overline{z}}(p-iq) = 0$, and so $p-iq$ is a holomorphic (and hence harmonic) function. Since both $(p-iq)$ and dz depend only on z, we will call $(p-iq)dz$ a holomorphic 1-form. The complex conjugate form

$$\overline{(p-iq)dz} = (p+iq)d\overline{z} \tag{8.189}$$

then depends only on $\overline{z}$ and is anti-holomorphic.

Riemann bilinear relations

As an illustration of the interplay of harmonic forms and two-dimensional topology, we derive some famous formuæ due to Riemann. These formulæ have applications in string theory and in conformal field theory.

Suppose that M is a Riemann surface of genus g, with α_i, β_i ,$i = 1, \ldots, g$, the representative generators of $H_1(M)$ that intersect as shown in figure 8.20. By applying Hodge-de Rham to this surface, we know that we can select a set of $2g$ independent, real, harmonic, 1-forms as a basis of $H^1(M, \mathbb{R})$. With the aid of the projector P we can assemble these into g holomorphic closed 1-forms ω_i, together with g anti-holomorphic closed 1-forms $\overline{\omega}_i$, the original $2g$ real forms being recovered from these as $\omega_i + \overline{\omega}_i$ and $\star(\omega_i + \overline{\omega}_i) = i(\overline{\omega}_i - \omega_i)$. A physical interpretation of these forms is as the z and $\overline{z}$ components of irrotational and incompressible fluid flows on the surface M. It is not surprising that such flows form a $2g$ real dimensional, or g complex dimensional, vector space because we can independently specify the

circulation $\oint \mathbf{v}\cdot d\mathbf{r}$ around each of the $2g$ generators of $H_1(M)$. If the flow field has (covariant) components v_x, v_y, then $\omega = v_z dz$ where $v_z = (v_x - iv_y)/2$, and $\overline{\omega} = v_{\overline{z}} d\overline{z}$ where $v_{\overline{z}} = (v_x + iv_y)/2$.

Suppose now that a and b are closed 1-forms on M. Then, either by exploiting the powerful and general intersection-form formula (4.77) or by cutting open the surface along the curves α_i, β_i and using the more direct strategy that gave us (4.79), we find that

$$\int_M a \wedge b = \sum_{i=1}^{g} \left\{ \int_{\alpha_i} a \int_{\beta_i} b - \int_{\beta_i} a \int_{\alpha_i} b \right\}. \tag{8.190}$$

We use this formula to derive two *bilinear relations* associated with a closed holomorphic 1-form ω. Firstly we compute its Hodge inner-product norm

$$\begin{aligned} \|\omega\|^2 \equiv \int_M \omega \wedge \star\overline{\omega} &= \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega \int_{\beta_i} \star\overline{\omega} - \int_{\beta_i} \omega \int_{\alpha_i} \star\overline{\omega} \right\} \\ &= i \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega \int_{\beta_i} \overline{\omega} - \int_{\beta_i} \omega \int_{\alpha_i} \overline{\omega} \right\} \\ &= i \sum_{i=1}^{g} \left\{ A_i \overline{B}_i - B_i \overline{A}_i \right\}, \end{aligned} \tag{8.191}$$

where $A_i = \int_{\alpha_i} \omega$ and $B_i = \int_{\beta_i} \omega$. We have used the fact that $\overline{\omega}$ is an anti-holomorphic 1 form and thus an eigenvector of $\star$ with eigenvalue i. It follows, therefore, that if all the A_i are zero then $\|\omega\| = 0$ and so $\omega = 0$.

Let $A_{ij} = \int_{\alpha_i} \omega_j$. The determinant of the matrix A_{ij} is non-zero: If it were zero, then there would be numbers λ_i, not all zero, such that

$$0 = A_{ij}\lambda_j = \int_{\alpha_i} (\omega_j \lambda_j), \tag{8.192}$$

but, by (8.191), this implies that $\|\omega_j \lambda_j\| = 0$ and hence $\omega_j \lambda_j = 0$, contrary to the linear independence of the ω_i. We can therefore solve the equations

$$A_{ij}\lambda_{jk} = \delta_{ik} \tag{8.193}$$

for the numbers λ_{jk} and use these to replace each of the ω_i by the linear combination $\omega_j \lambda_{ji}$. The new ω_i then obey $\int_{\alpha_i} \omega_j = \delta_{ij}$. From now on we suppose that this has be done.

Define $\tau_{ij} = \int_{\beta_i} \omega_j$. Observe that $dz \wedge dz = 0$ forces $\omega_i \wedge \omega_j = 0$, and therefore we have a second relation

$$\begin{aligned} 0 = \int_M \omega_m \wedge \omega_n &= \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega_m \int_{\beta_i} \omega_n - \int_{\beta_i} \omega_m \int_{\alpha_i} \omega_n \right\} \\ &= \sum_{i=1}^{g} \{\delta_{im}\tau_{in} - \tau_{im}\delta_{in}\} \\ &= \tau_{mn} - \tau_{nm}. \end{aligned} \tag{8.194}$$

The matrix τ_{ij} is therefore symmetric. A similar compuation shows that

$$\|\lambda_i \omega_i\|^2 = 2\overline{\lambda}_i (\operatorname{Im} \tau_{ij}) \lambda_j \tag{8.195}$$

so the matrix $(\operatorname{Im} \tau_{ij})$ is positive definite. The set of such symmetric matrices whose imaginary part is positive definite is called the *Siegel upper half-plane*. Not every such matrix correponds to a Riemann surface, but when it does it encodes all information about the shape of the Riemann manifold M that is left invariant under conformal rescaling.

8.7 Further Exercises and Problems

Exercise 8.10: *Harmonic partners.* Show that the function

$$u = \sin x \cosh y + 2 \cos x \sinh y$$

is harmonic. Determine the corresponding analytic function $u + iv$.

Exercise 8.11: *Möbius Maps.* The Map

$$z \mapsto w = \frac{az + b}{cz + d}$$

is called a Möbius transformation. These maps are important because they are the only one-to-one conformal maps of the Riemann sphere onto itself.

a) Show that two successive Möbius transformations

$$z' = \frac{az + b}{cz + d}, \qquad z'' = \frac{Az' + B}{Cz' + D}$$

give rise to another Möbius transformation, and show that the rule for combining them is equivalent to matrix multiplication.

b) Let z_1, z_2, z_3, z_4 be complex numbers. Show that a necessary and sufficient condition for the four points to be concyclic is that their their *cross-ratio*

$$\{z_1, z_2, z_3, z_4\} \stackrel{\text{def}}{=} \frac{(z_1 - z_4)(z_3 - z_2)}{(z_1 - z_2)(z_3 - z_4)}$$

be real (Hint: use a well-known property of opposite angles of a cyclic quadrilateral). Show that Möbius transformations leave the cross-ratio invariant, and thus take circles into circles.

Exercise 8.12: *Hyperbolic geometry*. The Riemann metric for the Poincaré-disc model of Lobachevski's hyperbolic plane (See exercises **??.??** and 3.13) can be taken to be

$$ds^2 = \frac{4|dz|^2}{(1 - |z|^2)^2}, \quad |z|^2 < 1.$$

a) Show that the Möbius transformation

$$z \mapsto w = e^{i\lambda} \frac{z - a}{\bar{a}z - 1}, \qquad |a| < 1, \quad \lambda \in \mathbb{R}$$

provides a 1-1 map of the interior of the unit disc onto itself. Show that these maps form a group.

b) Show that the hyperbolic-plane metric is left invariant under the group of maps in part (a). Deduce that such maps are orientation-preserving *isometries* of the hyperbolic plane.

c) Use the circle-preserving property of the Möbius maps to deduce that circles in hyperbolic geometry are represented in the Poincaré disc by Euclidean circles that lie entirely within the disc.

The conformal maps of part (a) are in fact the *only* orientation preserving isometries of the hyperbolic plane. With the exception of circles centered at $z = 0$, the center of the hyperbolic circle does not coincide with the center of its representative Euclidean circle. Euclidean circles that are internally tangent to the boundary of the unit disc have infinite hyperbolic radius and their hyperbolic centers lie on the boundary of the unit disc and hence at hyperbolic infinity. They are known as *horocycles.*

Exercise 8.13: *Rectangle to Ellipse.* Consider the map $w \mapsto z = \sin w$. Draw a picture of the image, in the z plane, of the interior of the rectangle with corners $u = \pm\pi/2$, $v = \pm\lambda$. $(w = u + iv)$. Show which points correspond to the corners of the rectangle, and verify that the vertex angles remain $\pi/2$. At what points does the isogonal property fail?

Exercise 8.14: The part of the negative real axis where $x < -1$ is occupied by a conductor held at potential $-V_0$. The positive real axis for $x > +1$ is similarly occupied by a conductor held at potential $+V_0$. The conductors extend to infinity in both directions perpendicular to the $x - y$ plane, and so the potential V satisfies the two-dimensional Laplace equation.

a) Find the image in the ζ plane of the cut z plane where the cuts run from -1 to $-\infty$ and from $+1$ to $+\infty$ under the map $z \mapsto \zeta = \sin^{-1} z$
b) Use your answer from part a) to solve the electrostatic problem and show that the field lines and equipotentials are conic sections of the form $ax^2+by^2 = 1$. Find expressions for a and b for the both the field lines and the equipotentials and draw a labelled sketch to illustrate your results.

Exercise 8.15: Draw the image under the map $z \mapsto w = e^{\pi z/a}$ of the infinite strip S, consisting of those points $z = x + iy \in \mathbb{C}$ for which $0 < y < a$. Label enough points to show which point in the w plane corresponds to which in the z plane. Hence or otherwise show that the Dirichlet Green function $G(x, y; x_0, y_0)$ that obeys

$$\nabla^2 G = \delta(x - x_0)\delta(y - y_0)$$

in S, and $G(x, y; x_0, y_0) = 0$ for (x, y) on the boundary of S, can be written as

$$G(x, y; x_0, y_0) = \frac{1}{2\pi} \ln |\sinh(\pi(z - z_0)/2a)| + \ldots$$

The dots indicate the presence of a second function, similar to the first, that you should find. Assume that $(x_0, y_0) \in S$.

Exercise 8.16: State Laurent's theorem for functions analytic in an annulus. Include formulae for the coefficients of the expansion. Show that, suitably interpreted, this theorem reduces to a form of Fourier's theorem for functions analytic in a neighbourhood of the unit circle.

Exercise 8.17: *Laurent Paradox.* Show that in the annulus $1 < |z| < 2$ the function

$$f(z) = \frac{1}{(z - 1)(2 - z)}$$

has a Laurent expansion in powers of z. Find the coefficients. The part of the series with negative powers of z does not terminate. Does this mean that $f(z)$ has an essential singularity at $z = 0$?

Exercise 8.18: Assuming the following series

$$\frac{1}{\sinh z} = \frac{1}{z} - \frac{1}{6}z + \frac{7}{16}z^3 + \ldots,$$

evaluate the integral

$$I = \oint_{|z|=1} \frac{1}{z^2 \sinh z} dz.$$

Now evaluate the integral

$$I = \oint_{|z|=4} \frac{1}{z^2 \sinh z} dz.$$

(Hint: The zeros of $\sinh z$ lie at $z = n\pi i$.)

Exercise 8.19: State the theorem relating the difference between the number of poles and zeros of $f(z)$ in a region to the winding number of argument of $f(z)$. Hence, or otherwise, evaluate the integral

$$I = \oint_C \frac{5z^4 + 1}{z^5 + z + 1} dz$$

where C is the circle $|z| = 2$. Prove, including a statement of any relevent theorem, any assertions you make about the locations of the zeros of $z^5 + z + 1$.

Exercise 8.20: *Arcsine branch cuts.* Let $w = \sin^{-1} z$. Show that

$$w = n\pi \pm i \ln\{iz + \sqrt{1 - z^2}\}$$

with the $\pm$ being selected depending on whether n is odd or even. Where would you put cuts to ensure that w is a single-valued function?

Problem 8.21: *Cutting open a genus-2 surface.* The Riemann surface for the function

$$y = \sqrt{(z - a_1)(z - a_2)(z - a_3)(z - a_4)(z - a_5)(z - a_6)}$$

has genus $g = 2$. Such a surface M is sketched in figure 8.22, where the four independent 1-cycles $\alpha_{1,2}$ and $\beta_{1,2}$ that generate $H_1(M)$ have been drawn so that they share a common vertex.

a) Realize the genus-2 surface as two copies of $\mathbb{C} \cup \{\infty\}$ cross-connected by three square-root branch cuts. Sketch how the 1-cycles α_i and β_i, $i = 1, 2$ of figure 8.22 appear when drawn on your thrice-cut plane.

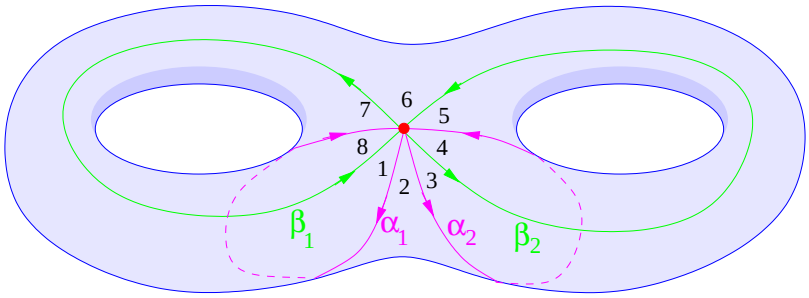

Figure 8.22: *Concurrent 1-cycles on a genus-2 surface.*

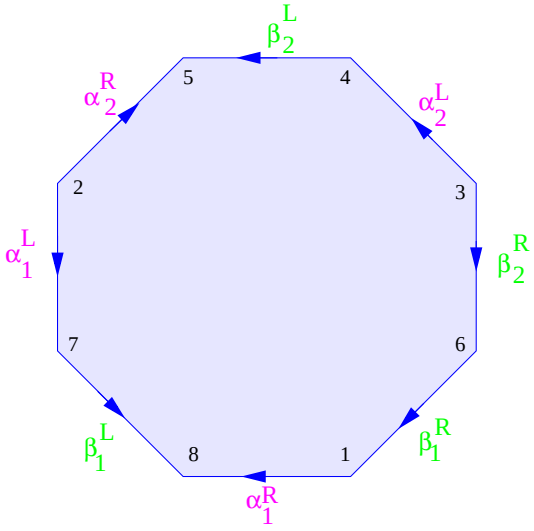

Figure 8.23: *The cut-open genus-2 surface. The superscripts L and R denote respectively the left and right sides of each 1-cycle, viewed from the direction of the arrow orienting the cycle.*

b) Cut the surface open along the four 1-cycles, and show that resulting surface is homeomorphic to the octagonal region appearing in figure 8.23.

c) Apply the direct method that gave us (4.79) to the octagonal region of part b). Hence show that for closed 1-forms a, b, on the surface we have

$$\int_M a \wedge b = \sum_{i=1}^{2} \left\{ \int_{\alpha_i} a \int_{\beta_i} b - \int_{\beta_i} a \int_{\alpha_i} b \right\}.$$

Chapter 9

Complex Analysis II

In this chapter we will apply what we have learned of complex variables. The applications will range from the elementary to the sophisticated.

9.1 Contour Integration Technology

The goal of contour integration technology is to evaluate ordinary, real-variable, definite integrals. We have already met the basic tool, the *residue theorem*:

Theorem: Let $f(z)$ be analytic within and on the boundary $\Gamma = \partial D$ of a simply connected domain D, with the exception of finite number of points at which the function has poles. Then

$$\oint_\Gamma f(z)\,dz = \sum_{\text{poles} \in D} 2\pi i\,(\text{residue at pole}).$$

9.1.1 Tricks of the Trade

The effective application of the residue theorem is something of an art, but there are useful classes of integrals which we can learn to recognize.

Rational Trigonometric Expressions

Integrals of the form

$$\int_0^{2\pi} F(\cos\theta, \sin\theta)\,d\theta \tag{9.1}$$

are dealt with by writing $\cos\theta = \frac{1}{2}(z+\overline{z})$, $\sin\theta = \frac{1}{2i}(z-\overline{z})$ and integrating around the unit circle. For example, let a, b be real and $b < a$, then

$$I = \int_0^{2\pi} \frac{d\theta}{a + b\cos\theta} = \frac{2}{i}\oint_{|z|=1} \frac{dz}{bz^2 + 2az + b} = \frac{2}{ib}\oint \frac{dz}{(z-\alpha)(z-\beta)}. \quad (9.2)$$

Since $\alpha\beta = 1$, only one pole is within the contour. This is at

$$\alpha = (-a + \sqrt{a^2 - b^2})/b. \quad (9.3)$$

The residue is

$$\frac{2}{ib}\frac{1}{\alpha - \beta} = \frac{1}{i}\frac{1}{\sqrt{a^2 - b^2}}. \quad (9.4)$$

Therefore, the integral is given by

$$I = \frac{2\pi}{\sqrt{a^2 - b^2}}. \quad (9.5)$$

These integrals are, of course, also do-able by the "t" substitution $t = \tan(\theta/2)$, whence

$$\sin\theta = \frac{2t}{1+t^2}, \quad \cos\theta = \frac{1-t^2}{1+t^2}, \quad d\theta = \frac{2dt}{1+t^2}, \quad (9.6)$$

followed by a partial fraction decomposition. The labour is perhaps slightly less using the contour method.

Rational Functions

Integrals of the form

$$\int_{-\infty}^{\infty} R(x)\,dx, \quad (9.7)$$

where $R(x)$ is a rational function of x with the degree of the denominator exceeding the degree of the numerator by two or more, may be evaluated by integrating around a rectangle from $-A$ to $+A$, A to $A + iB$, $A + iB$ to $-A + iB$, and back down to $-A$. Because the integrand decreases at least as fast as $1/|z|^2$ as z becomes large, we see that if we let $A, B \to \infty$, the contributions from the unwanted parts of the contour become negligeable. Thus

$$I = 2\pi i \left(\sum \text{Residues of poles in upper half-plane} \right). \quad (9.8)$$

We could also use a rectangle in the lower half-plane with the result

$$I = -2\pi i \left(\sum \text{Residues of poles in lower half-plane} \right), \tag{9.9}$$

This must give the same answer.

For example, let n be a positive integer and consider

$$I = \int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^n}. \tag{9.10}$$

The integrand has an n-th order pole at $z = \pm i$. Suppose we close the contour in the upper half-plane. The new contour encloses the pole at $z = +i$ and we therefore need to compute its residue. We set $z - i = \zeta$ and expand

$$\begin{aligned} \frac{1}{(1+z^2)^n} &= \frac{1}{[(i+\zeta)^2+1]^n} = \frac{1}{(2i\zeta)^n}\left(1 - \frac{i\zeta}{2}\right)^{-n} \\ &= \frac{1}{(2i\zeta)^n}\left(1 + n\left(\frac{i\zeta}{2}\right) + \frac{n(n+1)}{2!}\left(\frac{i\zeta}{2}\right)^2 + \cdots\right). \end{aligned} \tag{9.11}$$

The coefficient of ζ^{-1} is

$$\frac{1}{(2i)^n}\frac{n(n+1)\cdots(2n-2)}{(n-1)!}\left(\frac{i}{2}\right)^{n-1} = \frac{1}{2^{2n-1}i}\frac{(2n-2)!}{((n-1)!)^2}. \tag{9.12}$$

The integral is therefore

$$I = \frac{\pi}{2^{2n-2}}\frac{(2n-2)!}{((n-1)!)^2}. \tag{9.13}$$

These integrals can also be done by partial fractions.

9.1.2 Branch-cut integrals

Integrals of the form

$$I = \int_0^{\infty} x^{\alpha-1} R(x) dx, \tag{9.14}$$

where $R(x)$ is rational, can be evaluated by integration round a slotted circle (or "key-hole") contour.

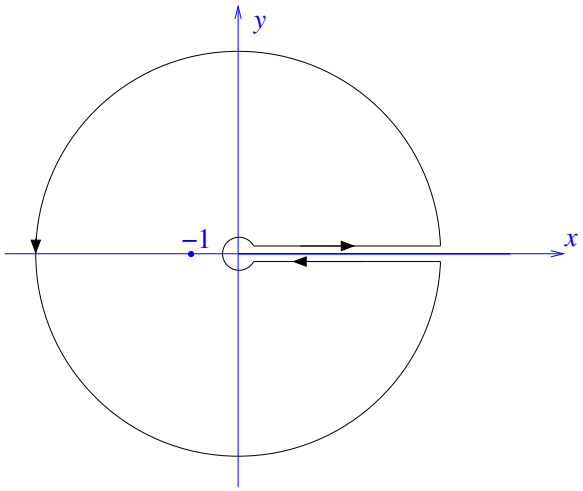

Figure 9.1: *A slotted circle contour* Γ *of outer radius* Λ *and inner radius* ϵ.

A little more work is required to extract the answer, though.

For example, consider

$$I = \int_0^\infty \frac{x^{\alpha-1}}{1+x} dx, \quad 0 < \operatorname{Re}\alpha < 1. \tag{9.15}$$

The restrictions on the range of α are necessary for the integral to converge at its upper and lower limits.

We take Γ to be a circle of radius Λ centred at $z = 0$, with a slot indentation designed to exclude the positive real axis, which we take as the branch cut of $z^{\alpha-1}$, and a small circle of radius ϵ about the origin. The branch of the fractional power is defined by setting

$$z^{\alpha-1} = \exp[(\alpha-1)(\ln|z| + i\theta)], \tag{9.16}$$

where we will take θ to be zero immediately above the real axis, and 2π immediately below it. With this definition the residue at the pole at $z = -1$ is $e^{i\pi(\alpha-1)}$. The residue theorem therefore tells us that

$$\oint_\Gamma \frac{z^{\alpha-1}}{1+z} dz = 2\pi i e^{\pi i(\alpha-1)}. \tag{9.17}$$

The integral decomposes as

$$\oint_\Gamma \frac{z^{\alpha-1}}{1+z} dz = \oint_{|z|=\Lambda} \frac{z^{\alpha-1}}{1+z} dz + (1 - e^{2\pi i(\alpha-1)}) \int_\epsilon^\Lambda \frac{x^{\alpha-1}}{1+x} dx - \oint_{|z|=\epsilon} \frac{z^{\alpha-1}}{1+z} dz. \tag{9.18}$$

As we send Λ off to infinity we can ignore the "1" in the denominator compared to the z, and so estimate

$$\left| \oint_{|z|=\Lambda} \frac{z^{\alpha-1}}{1+z} dz \right| \to \left| \oint_{|z|=\Lambda} z^{\alpha-2} dz \right| \leq 2\pi\Lambda \times \Lambda^{\mathrm{Re}\,(\alpha)-2}. \tag{9.19}$$

This tends to zero provided that $\mathrm{Re}\,\alpha < 1$. Similarly, provided $0 < \mathrm{Re}\,\alpha$, the integral around the small circle about the origin tends to zero with ϵ. Thus

$$-e^{\pi i\alpha} 2\pi i = \left(1 - e^{2\pi i(\alpha-1)}\right) I. \tag{9.20}$$

We conclude that

$$I = \frac{2\pi i}{(e^{\pi i\alpha} - e^{-\pi i\alpha})} = \frac{\pi}{\sin \pi\alpha}. \tag{9.21}$$

Exercise 9.1: Using the slotted circle contour, show that

$$I = \int_0^\infty \frac{x^{p-1}}{1+x^2} dx = \frac{\pi}{2\sin(\pi p/2)} = \frac{\pi}{2}\mathrm{cosec}\,(\pi p/2), \quad 0 < p < 2.$$

Exercise 9.2: Integrate $z^{a-1}/(z-1)$ around a contour Γ_1 consisting of a semicircle in the upper half plane together with the real axis indented at $z = 0$ and $z = 1$

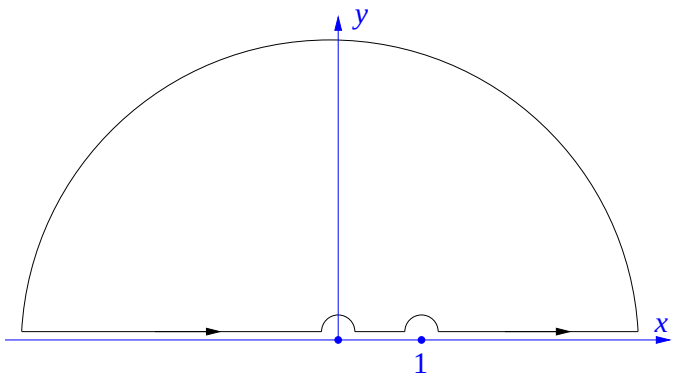

Figure 9.2: *The contour* Γ_1.

to get

$$0 = \oint_\Gamma \frac{z^{a-1}}{z-1}\, dz = P\int_0^\infty \frac{x^{a-1}}{x-1}\, dx - i\pi + (\cos \pi a + i \sin \pi a) \int_0^\infty \frac{x^{a-1}}{x+1}\, dx.$$

As usual, the symbol P in front of the integral sign denotes a *principal part* integral, meaning that we must omit an infinitesimal segment of the contour symmetrically disposed about the pole at $z = 1$. The term $-i\pi$ comes from integrating around the small semicircle about this point. We get $-1/2$ of the residue because we have only a half circle, and that traversed in the "wrong" direction. **Warning**: this fractional residue result is only true when we indent to avoid a *simple pole—i.e.* one that is of order one.

Now take real and imaginary parts and deduce that

$$\int_0^\infty \frac{x^{a-1}}{1+x} dx = \frac{\pi}{\sin \pi\alpha}, \quad 0 < \operatorname{Re} a < 1,$$

and

$$P\int_0^\infty \frac{x^{a-1}}{1-x} dx = \pi \cot \pi a, \quad 0 < \operatorname{Re} a < 1.$$

9.1.3 Jordan's Lemma

We often need to evaluate Fourier integrals

$$I(k) = \int_{-\infty}^{\infty} e^{ikx} R(x)\, dx \tag{9.22}$$

with $R(x)$ a rational function. For example, the Green function for the operator $-\partial_x^2 + m^2$ is given by

$$G(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \frac{e^{ikx}}{k^2 + m^2}. \tag{9.23}$$

Suppose $x \in \mathbb{R}$ and $x > 0$. Then, in contrast to the analogous integral without the exponential function, we have no flexibility in closing the contour in the upper or lower half-plane. The function e^{ikx} grows without limit as we head south in the lower half-plane, but decays rapidly in the upper half-plane. This means that we may close the contour without changing the value of the integral by adding a large upper-half-plane semicircle.

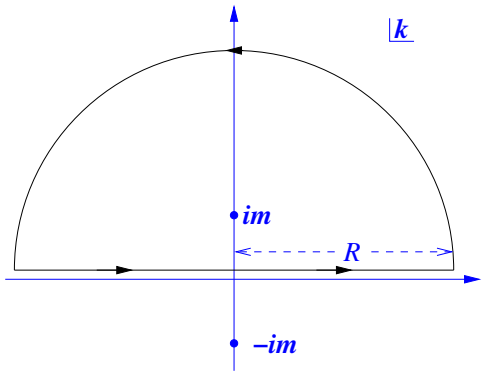

Figure 9.3: *Closing the contour in the upper half-plane.*

The modified contour encloses a pole at $k = im$, and this has residue $i/(2m)e^{-mx}$. Thus

$$G(x) = \frac{1}{2m}e^{-mx}, \quad x > 0. \tag{9.24}$$

For $x < 0$, the situation is reversed, and we must close in the lower half-plane. The residue of the pole at $k = -im$ is $-i/(2m)e^{mx}$, but the minus sign is cancelled because the contour goes the "wrong way" (clockwise). Thus

$$G(x) = \frac{1}{2m}e^{+mx}, \quad x < 0. \tag{9.25}$$

We can combine the two results as

$$G(x) = \frac{1}{2m}e^{-m|x|}. \tag{9.26}$$

The formal proof that the added semicircles make no contribution to the integral when their radius becomes large is known as *Jordan's Lemma*:
Lemma: Let Γ be a semicircle, centred at the origin, and of radius R. Suppose

i) that $f(z)$ is meromorphic in the upper half-plane;
ii) that $f(z)$ tends uniformly to zero as $|z| \to \infty$ for $0 < \arg z < \pi$;
iii) the number λ is real and positive.

Then

$$\int_\Gamma e^{i\lambda z} f(z)\, dz \to 0, \quad \text{as} \quad R \to \infty. \tag{9.27}$$

To establish this, we assume that R is large enough that $|f| < \epsilon$ on the contour, and make a simple estimate

$$\begin{aligned}\left|\int_\Gamma e^{i\lambda z} f(z)\,dz\right| &< 2R\epsilon \int_0^{\pi/2} e^{-\lambda R\sin\theta}\,d\theta \\ &< 2R\epsilon \int_0^{\pi/2} e^{-2\lambda R\theta/\pi}\,d\theta \\ &= \frac{\pi\epsilon}{\lambda}(1-e^{-\lambda R}) < \frac{\pi\epsilon}{\lambda}. \end{aligned} \tag{9.28}$$

In the second inequality we have used the fact that $(\sin\theta)/\theta \geq 2/\pi$ for angles in the range $0 < \theta < \pi/2$. Since ϵ can be made as small as we like, the lemma follows.

Example: Evaluate

$$I(\alpha) = \int_{-\infty}^{\infty} \frac{\sin(\alpha x)}{x}\,dx. \tag{9.29}$$

We have

$$I(\alpha) = \mathrm{Im}\left\{\int_{-\infty}^{\infty} \frac{\exp i\alpha z}{z}\,dz\right\}. \tag{9.30}$$

If we take $\alpha > 0$, we can close in the upper half-plane, but our contour must exclude the pole at $z = 0$. Therefore

$$0 = \int_{|z|=R} \frac{\exp i\alpha z}{z}\,dz - \int_{|z|=\epsilon} \frac{\exp i\alpha z}{z}\,dz + \int_{-R}^{-\epsilon} \frac{\exp i\alpha x}{x}\,dx + \int_{\epsilon}^{R} \frac{\exp i\alpha x}{x}\,dx. \tag{9.31}$$

As $R \to \infty$, we can ignore the big semicircle, the rest, after letting $\epsilon \to 0$, gives

$$0 = -i\pi + P\int_{-\infty}^{\infty} \frac{e^{i\alpha x}}{x}\,dx. \tag{9.32}$$

Again, the symbol P denotes a principal part integral. The $-i\pi$ comes from the small semicircle. We get $-1/2$ the residue because we have only a half circle, and that traversed in the "wrong" direction. (Remember that this fractional residue result is only true when we indent to avoid a *simple pole*—*i.e* one that is of order one.)

Reading off the real and imaginary parts, we conclude that

$$\int_{-\infty}^{\infty} \frac{\sin\alpha x}{x}\,dx = \pi, \quad P\int_{-\infty}^{\infty} \frac{\cos\alpha x}{x}\,dx = 0, \qquad \alpha > 0. \tag{9.33}$$

No "P" is needed in the sine integral, as the integrand is finite at $x = 0$.

If we relax the condition that $\alpha > 0$ and take into account that sine is an odd function of its argument, we have

$$\int_{-\infty}^{\infty} \frac{\sin \alpha x}{x}\, dx = \pi \operatorname{sgn} \alpha. \tag{9.34}$$

This identity is called *Dirichlet's discontinuous integral*.

We can interpret Dirichlet's integral as giving the Fourier transform of the principal part distribution $P(1/x)$ as

$$P \int_{-\infty}^{\infty} \frac{e^{i\omega x}}{x}\, dx = i\pi \operatorname{sgn} \omega. \tag{9.35}$$

This will be of use later in the chapter.
Example:

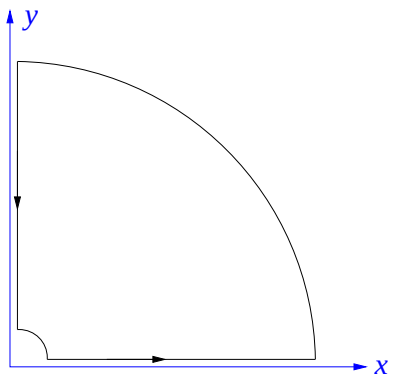

Figure 9.4: *Quadrant contour.*

We will evaluate the integral

$$\oint_C e^{iz} z^{a-1}\, dz \tag{9.36}$$

about the first-quadrant contour shown above. Observe that when $0 < a < 1$ neither the large nor the small arc makes a contribution, and that there are no poles. Hence, we deduce that

$$0 = \int_0^{\infty} e^{ix} x^{a-1}\, dx - i \int_0^{\infty} e^{-y} y^{a-1} e^{(a-1)\frac{\pi}{2} i}\, dy, \quad 0 < a < 1. \tag{9.37}$$

Taking real and imaginary parts, we find

$$\begin{aligned}\int_0^\infty x^{a-1}\cos x\,dx &= \Gamma(a)\cos\left(\frac{\pi}{2}a\right), \quad 0<a<1,\\ \int_0^\infty x^{a-1}\sin x\,dx &= \Gamma(a)\sin\left(\frac{\pi}{2}a\right), \quad 0<a<1,\end{aligned} \tag{9.38}$$

where

$$\Gamma(a)=\int_0^\infty y^{a-1}e^{-y}\,dy \tag{9.39}$$

is the Euler Gamma function.

Example: Fresnel integrals. Integrals of the form

$$C(t) = \int_0^t \cos(\pi x^2/2)\,dx, \tag{9.40}$$

$$S(t) = \int_0^t \sin(\pi x^2/2)\,dx, \tag{9.41}$$

occur in the theory of diffraction and are called *Fresnel integrals* after Augustin Fresnel. They are naturally combined as

$$C(t)+iS(t)=\int_0^t e^{i\pi x^2/2}\,dx. \tag{9.42}$$

The limit as $t\to\infty$ exists and is finite. Even though the integrand does not tend to zero at infinity, its rapid oscillation for large x is just sufficient to ensure convergence.[1]

As t varies, the complex function $C(t)+iS(t)$ traces out the *Cornu Spiral*, named after Marie Alfred Cornu, a 19th century French optical physicist.

[1]We can exhibit this convergence by setting $x^2=s$ and then integrating by parts to get

$$\int_0^t e^{i\pi x^2/2}\,dx=\frac{1}{2}\int_0^1 e^{i\pi s/2}\,\frac{ds}{s^{1/2}}+\left[\frac{e^{i\pi s/2}}{\pi i s^{1/2}}\right]_1^{t^2}+\frac{1}{2\pi i}\int_1^{t^2} e^{i\pi s/2}\,\frac{ds}{s^{3/2}}.$$

The right hand side is now manifestly convergent as $t\to\infty$.

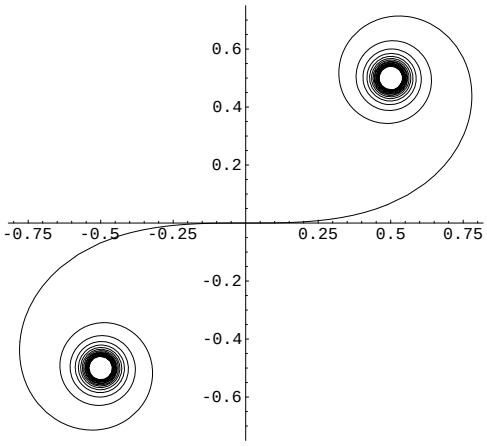

Figure 9.5: *The Cornu spiral $C(t)+iS(t)$ for t in the range $-8 < t < 8$. The spiral in the first quadrant corresponds to positive values of t.*

We can evaluate the limiting value

$$C(\infty) + iS(\infty) = \int_0^\infty e^{i\pi x^2/2}\, dx \tag{9.43}$$

by deforming the contour off the real axis and onto a line of length L running into the first quadrant at 45°, this being the direction of most rapid decrease of the integrand.

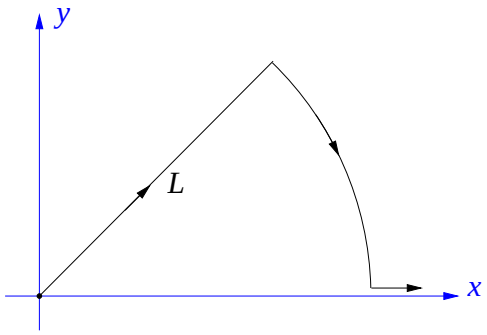

Figure 9.6: *Fresnel contour.*

A circular arc returns the contour to the axis whence it continues to ∞, but an estimate similar to that in Jordan's lemma shows that the arc and the

subsequent segment on the real axis make a negligeable contribution when L is large. To evaluate the integral on the radial line we set $z = e^{i\pi/4}s$, and so

$$\int_0^{e^{i\pi/4}\infty} e^{i\pi z^2/2}\,dz = e^{i\pi/4}\int_0^{\infty} e^{-\pi s^2/2}\,ds = \frac{1}{\sqrt{2}}e^{i\pi/4} = \frac{1}{2}(1+i). \tag{9.44}$$

Figure 9.5 shows how $C(t) + iS(t)$ orbits the limiting point $0.5 + 0.5i$ and slowly spirals in towards it. Taking real and imaginary parts we have

$$\int_0^{\infty} \cos\left(\frac{\pi x^2}{2}\right)\,dx = \int_0^{\infty} \sin\left(\frac{\pi x^2}{2}\right)\,dx = \frac{1}{2}. \tag{9.45}$$

9.2 The Schwarz Reflection Principle

Theorem (Schwarz): Let $f(z)$ be analytic in a domain D where ∂D includes a segment of the real axis. Assume that $f(z)$ is real when z is real. Then there is a unique *analytic continuation of f into the region $\overline{D}$ (the mirror image of D in the real axis) given by*

$$g(z) = \begin{cases} f(z), & z \in D, \\ \overline{f(\overline{z})}, & z \in \overline{D}, \\ \text{either}, & z \in \mathbb{R}. \end{cases} \tag{9.46}$$

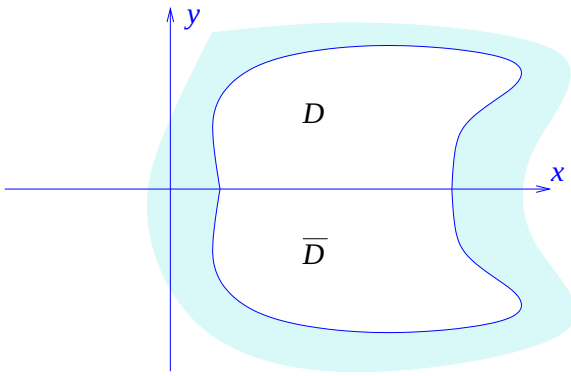

Figure 9.7: The domain D and its mirror image $\overline{D}$.

The proof invokes Morera's theorem to show analyticity, and then appeals to the uniqueness of analytic continuations. Begin by looking at a closed

contour lying only in $\overline{D}$:

$$\oint_C \overline{f(\overline{z})}\, dz, \tag{9.47}$$

where $C = \{\overline{\eta(t)}\}$ is the image of $\overline{C} = \{\eta(t)\} \subset D$ under reflection in the real axis. We can rewrite this as

$$\oint_C \overline{f(\overline{z})}\, dz = \oint \overline{f(\eta)}\, \frac{d\bar{\eta}}{dt} dt = \overline{\oint f(\eta) \frac{d\eta}{dt} dt} = \overline{\oint_{\overline{C}} f(\eta)\, dz} = 0. \tag{9.48}$$

At the last step we have used Cauchy and the analyticity of f in D. Morera's theorem therefore confirms that $g(z)$ is analytic in $\overline{D}$. By breaking a general contour up into parts in D and parts in $\overline{D}$, we can similarly show that $g(z)$ is analytic in $D \cup \overline{D}$.

The important corollary is that if $f(z)$ is analytic, and real on some segment of the real axis, but has a cut along some other part of the real axis, then $f(x + i\epsilon) = \overline{f(x - i\epsilon)}$ as we go over the cut. The discontinuity disc f is therefore $2\mathrm{Im}\, f(x + i\epsilon)$.

Suppose $f(z)$ is real on the negative real axis, and goes to zero as $|z| \to \infty$, then applying Cauchy to the contour Γ depicted in figure 9.8.

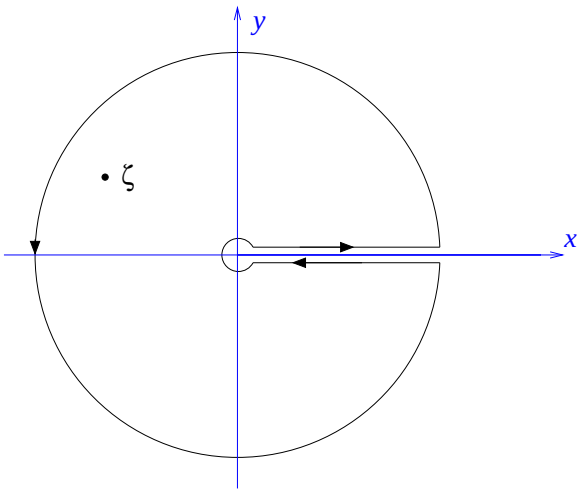

Figure 9.8: *The contour* Γ *for the dispersion relation.* .

we find

$$f(\zeta) = \frac{1}{\pi} \int_0^\infty \frac{\mathrm{Im}\, f(x + i\epsilon)}{x - \zeta}\, dx, \tag{9.49}$$

for ζ within the contour. This is an example of a *dispersion relation*. The name comes from the prototypical application of this technology to optical dispersion, *i.e.* the variation of the refractive index with frequency.

If $f(z)$ does not tend to zero at infinity then we cannot ignore the contribution to Cauchy's formula from the large circle. We can, however, still write

$$f(\zeta) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z-\zeta}\, dz, \tag{9.50}$$

and

$$f(b) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z-b}\, dz, \tag{9.51}$$

for some convenient point b within the contour. We then subtract to get

$$f(\zeta) = f(b) + \frac{(\zeta-b)}{2\pi i} \int_\Gamma \frac{f(z)}{(z-b)(z-\zeta)}\, dz. \tag{9.52}$$

Because of the extra power of z downstairs in the integrand, we only need f to be bounded at infinity for the contribution of the large circle to tend to zero. If this is the case, we have

$$f(\zeta) = f(b) + \frac{(\zeta-b)}{\pi} \int_0^\infty \frac{\mathrm{Im}\, f(x+i\epsilon)}{(x-b)(x-\zeta)}\, dx. \tag{9.53}$$

This is called a *once-subtracted* dispersion relation.

The dispersion relations derived above apply when ζ lies within the contour. In physics applications we often need $f(\zeta)$ for ζ real and positive. What happens as ζ approaches the axis, and we attempt to divide by zero in such an integral, is summarized by the *Plemelj formulæ*: If $f(\zeta)$ is defined by

$$f(\zeta) = \frac{1}{\pi} \int_\Gamma \frac{\rho(z)}{z-\zeta}\, dz, \tag{9.54}$$

where Γ has a segment lying on the real axis, then, if x lies in this segment,

$$\begin{aligned} \frac{1}{2}(f(x+i\epsilon) - f(x-i\epsilon)) &= i\rho(x) \\ \frac{1}{2}(f(x+i\epsilon) + f(x-i\epsilon)) &= \frac{P}{\pi} \int_\Gamma \frac{\rho(x')}{x'-x}\, dx'. \end{aligned} \tag{9.55}$$

As always, the "P" means that we are to delete an infinitesimal segment of the contour lying symmetrically about the pole.

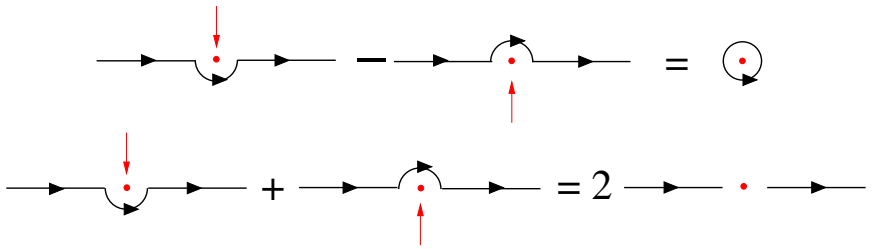

Figure 9.9: *Origin of the Plemelj formulae.*

The Plemelj formulæ hold under relatively mild conditions on the function $\rho(x)$. We won't try to give a general proof, but in the case that ρ is analytic the result is easy to understand: we can push the contour out of the way and let $\zeta \to x$ on the real axis from either above or below. In that case the drawing above shows how the the sum of these two limits gives the the principal-part integral and how their difference gives an integral round a small circle, and hence the residue $\rho(x)$.

The Plemelj equations usually appear in physics papers as the "$i\epsilon$" cabala

$$\frac{1}{x' - x \pm i\epsilon} = P\left(\frac{1}{x' - x}\right) \mp i\pi\delta(x' - x). \tag{9.56}$$

A limit $\epsilon \to 0$ is always to be understood in this formula.

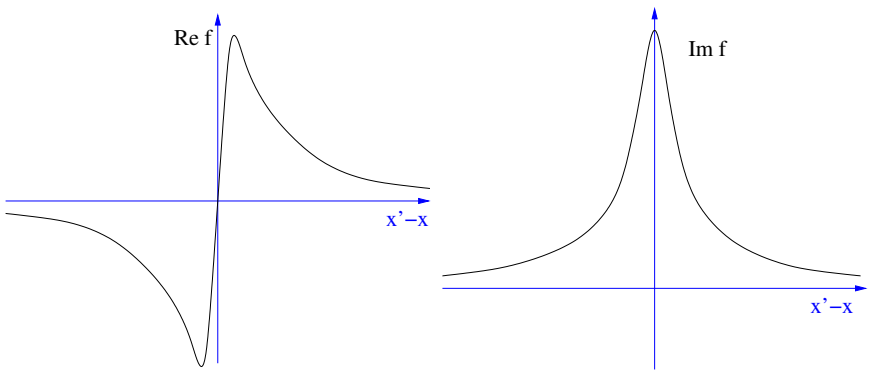

Figure 9.10: *Sketch of the real and imaginary parts of* $f(x') = 1/(x' - x - i\epsilon)$.

We can also appreciate the origin of the $i\epsilon$ rule by examining the following identity:

$$\frac{1}{x' - (x \pm i\epsilon)} = \frac{x - x'}{(x' - x)^2 + \epsilon^2} \pm \frac{i\epsilon}{(x' - x)^2 + \epsilon^2}. \tag{9.57}$$

The first term is a symmetrically cut-off version of $1/(x'-x)$ and provides the principal-part integral. The second term sharpens and tends to the delta function $\pm i\pi\delta(x'-x)$ as $\epsilon \to 0$.

Exercise 9.3: The Legendre function of the second kind $Q_n(z)$ may be defined for positive integer n by the integral

$$Q_n(z) = \frac{1}{2}\int_{-1}^{1} \frac{(1-t^2)^n}{2^n(z-t)^{n+1}}\,dt, \quad z \notin [-1,1].$$

Show that for $x \in [-1,1]$ we have

$$Q_n(x+i\epsilon) - Q_n(x-i\epsilon) = -i\pi P_n(x),$$

where $P_n(x)$ is the Legendre Polynomial. Deduce *Neumann 's formula*

$$Q_n(z) = \frac{1}{2}\int_{-1}^{1} \frac{P_n(t)}{z-t}\,dt, \quad z \notin [-1,1].$$

9.2.1 Kramers-Kronig Relations

Causality is the usual source of analyticity in physical applications. If $G(t)$ is a response function

$$\phi_{\text{response}}(t) = \int_{-\infty}^{\infty} G(t-t')f_{\text{cause}}(t')\,dt' \tag{9.58}$$

then for no effect to anticipate its cause we must have $G(t) = 0$ for $t < 0$. The Fourier transform

$$G(\omega) = \int_{-\infty}^{\infty} e^{i\omega t}G(t)\,dt, \tag{9.59}$$

is then automatically analytic everywhere in the upper half plane. Suppose, for example, we look at a forced, damped, harmonic oscillator whose displacement $x(t)$ obeys

$$\ddot{x} + 2\gamma\dot{x} + (\Omega^2+\gamma^2)x = F(t), \tag{9.60}$$

where the friction coefficient γ is positive. As we saw earlier, the solution is of the form

$$x(t) = \int_{-\infty}^{\infty} G(t,t')F(t')dt',$$

where the Green function $G(t,t') = 0$ if $t < t'$. In this case

$$G(t,t') = \begin{cases} \Omega^{-1} e^{-\gamma(t-t')} \sin \Omega(t-t') & t > t' \\ 0, & t < t' \end{cases} \tag{9.61}$$

and so

$$x(t) = \frac{1}{\Omega} \int_{-\infty}^{t} e^{-\gamma(t-t')} \sin \Omega(t-t')\, F(t')\, dt'. \tag{9.62}$$

Because the integral extends only from 0 to $+\infty$, the Fourier transform of $G(t,0)$,

$$\tilde{G}(\omega) \equiv \frac{1}{\Omega} \int_0^{\infty} e^{i\omega t} e^{-\gamma t} \sin \Omega t\, dt, \tag{9.63}$$

is nicely convergent when $\operatorname{Im} \omega > 0$, as evidenced by

$$\tilde{G}(\omega) = -\frac{1}{(\omega + i\gamma)^2 - \Omega^2} \tag{9.64}$$

having no singularities in the upper half-plane.[2]

Another example of such a causal function is provided by the complex, frequency-dependent, *refractive index* of a material $n(\omega)$. This is defined so that a travelling wave takes the form

$$\varphi(\mathbf{x}, t) = e^{in(\omega)\mathbf{k}\cdot\mathbf{x} - i\omega t}. \tag{9.65}$$

We can decompose n into its real and imaginary parts

$$\begin{aligned} n(\omega) &= n_R(\omega) + i n_I(\omega) \\ &= n_R(\omega) + \frac{i}{2|k|}\gamma(\omega) \end{aligned} \tag{9.66}$$

where γ is the extinction coefficient, defined so that the intensity falls off as $I \propto \exp(-\gamma \mathbf{n} \cdot \mathbf{x})$, where $\mathbf{n} = \mathbf{k}/|k|$ is the direction of propapagation. A non-zero γ can arise from either energy absorption or scattering out of the forward direction

[2] If a pole in a response function manages to sneak into the upper half plane, then the system will be unstable to exponentially growing oscillations. This may happen, for example, when we design an electronic circuit containing a feedback loop. Such poles, and the resultant instabilities, can be detected by applying the principle of the argument from the last chapter. This method leads to the *Nyquist stability criterion.*

Being a causal response, the refractive index extends to a function analytic in the upper half plane and $n(\omega)$ for real ω is the boundary value

$$n(\omega)_{\text{physical}} = \lim_{\epsilon \to 0} n(\omega + i\epsilon) \tag{9.67}$$

of this analytic function. Because a real ($\mathbf{E} = \mathbf{E}^*$) incident wave must give rise to a real wave in the material, and because the wave must decay in the direction in which it is propagating, we have the reality conditions

$$\begin{aligned} \gamma(-\omega + i\epsilon) &= -\gamma(\omega + i\epsilon), \\ n_R(-\omega + i\epsilon) &= +n_R(\omega + i\epsilon) \end{aligned} \tag{9.68}$$

with γ positive for positive frequency.

Many materials have a frequency range $|\omega| < |\omega_{\min}|$ where $\gamma = 0$, so the material is transparent. For any such material $n(\omega)$ obeys the Schwarz reflection principle and so there is an analytic continuation into the lower half-plane. At frequencies ω where the material is not perfectly transparent, the refractive index has an imaginary part even when ω is real. By Schwarz, n must be discontinuous across the real axis at these frequencies: $n(\omega + i\epsilon) = n_R + in_I \neq n(\omega - i\epsilon) = n_R - in_I$. These discontinuities of $2in_I$ usually correspond to branch cuts.

No substance is able to respond to infinitely high frequency disturbances, so $n \to 1$ as $|\omega| \to \infty$, and we can apply our dispersion relation technology to the function $n - 1$. We will need the contour shown below, which has cuts for both positive and negative frequencies.

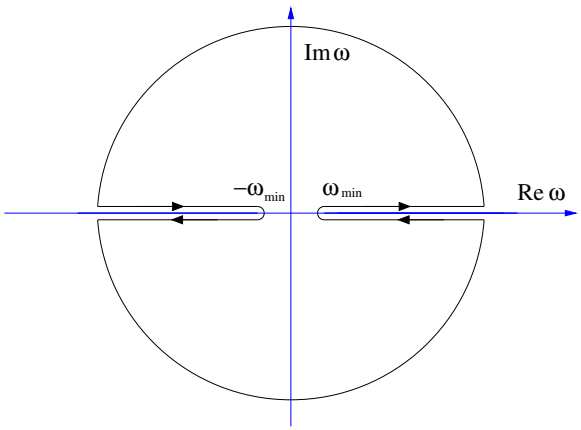

Figure 9.11: *Contour for the $n - 1$ dispersion relation.*

By applying the dispersion-relation strategy, we find

$$n(\omega) = 1 + \frac{1}{\pi}\int_{-\infty}^{\omega_{\min}} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' + \frac{1}{\pi}\int_{\omega_{\min}}^{\infty} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' \tag{9.69}$$

for ω within the contour. Using Plemelj we can now take ω onto the real axis to get

$$\begin{aligned} n_R(\omega) &= 1 + \frac{P}{\pi}\int_{-\infty}^{\omega_{\min}} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' + \frac{P}{\pi}\int_{\omega_{\min}}^{\infty} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' \\ &= 1 + \frac{P}{\pi}\int_{\omega^2_{\min}}^{\infty} \frac{n_I(\omega')}{\omega'^2 - \omega^2}\, d\omega'^2, \\ &= 1 + \frac{c}{\pi} P \int_{\omega_{\min}}^{\infty} \frac{\gamma(\omega')}{\omega'^2 - \omega^2}\, d\omega'. \end{aligned} \tag{9.70}$$

In the second line we have used the anti-symmetry of $n_I(\omega)$ to combine the positive and negative frequency range integrals. In the last line we have used the relation $\omega/k = c$ to make connection with the way this equation is written in R. G. Newton's authoritative *Scattering Theory of Waves and Particles*. This relation, between the real and absorptive parts of the refractive index, is called a *Kramers-Kronig* dispersion relation, after the original authors.[3]

If $n \to 1$ fast enough that $\omega^2(n-1) \to 0$ as $|\omega| \to \infty$, we can take the f in the dispersion relation to be $\omega^2(n-1)$ and deduce that

$$n_R = 1 + \frac{c}{\pi} P \int_{\omega^2_{min}}^{\infty} \left(\frac{\omega'^2}{\omega^2}\right) \frac{\gamma(\omega')}{\omega'^2 - \omega^2}\, d\omega', \tag{9.71}$$

another popular form of Kramers-Kronig. This second relation implies the first, but not *vice-versa*, because the second demands more restrictive behavior for $n(\omega)$.

Similar equations can be derived for other causal functions. A quantity closely related to the refractive index is the frequency-dependent dielectric "constant"

$$\epsilon(\omega) = \epsilon_1 + i\epsilon_2. \tag{9.72}$$

Again $\epsilon \to 1$ as $|\omega| \to \infty$, and, proceeding as before, we deduce that

$$\epsilon_1(\omega) = 1 + \frac{P}{\pi}\int_{\omega^2_{min}}^{\infty} \frac{\epsilon_2(\omega')}{\omega'^2 - \omega^2}\, d\omega'^2. \tag{9.73}$$

[3]H. A. Kramers, *Nature*, **117** (1926) 775; R. de L. Kronig, J. Opt. Soc. Am. **12** (1926) 547

9.2.2 Hilbert transforms

Suppose that $f(x)$ is the boundary value on the real axis of a function everywhere analytic in the upper half-plane, and suppose further that $f(z) \to 0$ as $|z| \to \infty$ there. Then we have

$$f(z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{f(x)}{x - z}\, dx \tag{9.74}$$

for z in the upper half-plane. This is because may close the contour with an upper semicircle without changing the value of the integral. For the same reason the integral must give zero when z is taken in the lower half-plane. Using the Plemelj formulæ we deduce that on the real axis,

$$f(x) = \frac{P}{\pi i} \int_{-\infty}^{\infty} \frac{f(x')}{x' - x}\, dx'. \tag{9.75}$$

We can use this strategy to derive the Kramers-Kronig relations even if n_I never vanishes, and so we cannot use the Schwarz reflection principle.

The relation (9.75) suggests the definition of the *Hilbert transform*, $\mathcal{H}\psi$, of a function $\psi(x)$, as

$$(\mathcal{H}\psi)(x) = \frac{P}{\pi} \int_{-\infty}^{\infty} \frac{\psi(x')}{x - x'}\, dx'. \tag{9.76}$$

Note the interchange of x, x' in the denominator of (9.76) when compared with (9.75). This switch is to make the Hilbert transform into a convolution integral. Equation (9.75) shows that a function that is the boundary value of a function analytic and tending to zero at infinity in the upper half-plane is automatically an eigenvector of $\mathcal{H}$ with eigenvalue $-i$. Similarly a function that is the boundary value of a function analytic and tending to zero at infinity in the lower half-plane will be an eigenvector with eigenvalue $+i$. (A function analytic in the *entire* complex plane and tending to zero at infinity must vanish identically by Liouville's theorem.)

Returning now to our original f, which had eigenvalue $-i$, and decomposing it as $f(x) = f_R(x) + if_I(x)$ we find that (9.75) becomes

$$\begin{aligned} f_I(x) &= (\mathcal{H}f_R)(x), \\ f_R(x) &= -(\mathcal{H}f_I)(x). \end{aligned} \tag{9.77}$$

Conversely, if we are given a real function $u(x)$ and set $v(x) = (\mathcal{H}u)(x)$, then, under some mild restrictions on u (that it lie in some $L^p(\mathbb{R})$, $p > 1$, for example, in which case $v(x)$ is also in $L^p(\mathbb{R})$.) the function

$$f(z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{u(x) + iv(x)}{x - z}\, dx \tag{9.78}$$

will be analytic in the upper half plane, tend to zero at infinity there, and have $u(x) + iv(x)$ as its boundary value as z approaches the real axis from above. The last line of (9.77) therefore shows that we may recover $u(x)$ from $v(x)$ as $u(x) = -(\mathcal{H}v)(x)$. The Hilbert transform $\mathcal{H} : L^p(\mathbb{R}) \to L^p(\mathbb{R})$ is therefore invertible, and its inverse is given by $\mathcal{H}^{-1} = -\mathcal{H}$. (Note that the Hilbert transform of a constant is zero, but the $L^p(\mathbb{R})$ condition excludes constants from the domain of $\mathcal{H}$, and so this fact does not conflict with invertibility.)

Hilbert transforms are useful in signal processing. Given a real signal $X_R(t)$ we can take its Hilbert transform so as to find the corresponding imaginary part, $X_I(t)$, which serves to make the sum

$$Z(t) = X_R(t) + iX_I(t) = A(t)e^{i\phi(t)} \tag{9.79}$$

analytic in the upper half-plane. This complex function is the *analytic signal*.[4] The real quantity $A(t)$ is then known as the *instantaneous amplitude*, or *envelope*, while $\phi(t)$ is the *instantaneous phase* and

$$\omega_{\rm IF}(t) = \dot{\phi}(t) \tag{9.80}$$

is called the *instantaneous frequency* (IF). These quantities are used, for example, in narrow band FM radio, in NMR, in geophysics, and in image processing.

Exercise 9.4: Let $\widetilde{f}(\omega) = \int_{-\infty}^{\infty} e^{i\omega t} f(t)\, dt$ denote the Fourier transform of $f(t)$. Use the formula (9.35) for the Fourier transform of $P(1/t)$, combined with the convolution theorem for Fourier transforms, to show that the Fourier transform of the Hilbert transform of $f(t)$ is

$$\widetilde{(\mathcal{H}f)}(\omega) = i\,\mathrm{sgn}(\omega)\widetilde{f}(\omega).$$

Deduce that the analytic signal is derived from the original real signal by suppressing all positive frequency components (those proportional to $e^{-i\omega t}$ with $\omega > 0$) and multiplying the remaining negative-frequency amplitudes by two.

[4] D. Gabor, *J. Inst. Elec. Eng. (Part 3)*, **93** (1946) 429-457.

Exercise 9.5: Suppose that $\varphi_1(x)$ and $\varphi_2(x)$ are real functions with finite $L^2(\mathbb{R})$ norms.

a) Use the Fourier transform result from the previous exercise to show that

$$\langle \varphi_1, \varphi_2 \rangle = \langle \mathcal{H}\varphi_1, \mathcal{H}\varphi_2 \rangle.$$

Thus, $\mathcal{H}$ is a unitary transformation from $L^2(\mathbb{R}) \to L^2(\mathbb{R})$.

b) Use the fact that $\mathcal{H}^2 = -I$ to deduce that

$$\langle \mathcal{H}\varphi_1, \varphi_2 \rangle = -\langle \varphi_1, \mathcal{H}\varphi_2 \rangle$$

and so $\mathcal{H}^\dagger = -\mathcal{H}$.

c) Conclude from part b) that

$$\int_{-\infty}^{\infty} \varphi_1(x) \left(P \int_{-\infty}^{\infty} \frac{\varphi_2(y)}{x-y}\, dy \right) dx = \int_{-\infty}^{\infty} \varphi_2(y) \left(P \int_{-\infty}^{\infty} \frac{\varphi_1(x)}{x-y}\, dx \right) dy,$$

i.e., for $L^2(\mathbb{R})$, functions, it is legitimate to interchange the order of "P" integration with ordinary integration.

d) By replacing $\varphi_1(x)$ by a constant, and $\varphi_2(x)$ by the Hilbert transform of a function f with $\int f\, dx \neq 0$, show that it is not *always* safe to interchange the order of "P" integration with ordinary integration

Exercise 9.6: Suppose that are given real functions $u_1(x)$ and $u_2(x)$ and substitute their Hilbert transforms $v_1 = \mathcal{H}u_1$, $v_2 = \mathcal{H}u_2$ into (9.78) to construct analytic functions $f_1(z)$ and $f_2(z)$. Then the product $f_1(z)f_2(z) = F(z)$ has boundary value

$$F_R(x) + iF_I(x) = (u_1u_2 - v_1v_2) + i(u_1v_2 + u_2v_1).$$

By assuming that $F(z)$ satisfies the conditions for (9.77) to be applicable to this boundary value, deduce that

$$\mathcal{H}\left((\mathcal{H}u_1)u_2\right) + \mathcal{H}((\mathcal{H}u_2)u_1) - (\mathcal{H}u_1)(\mathcal{H}u_2) = -u_1u_2. \quad \star$$

This result[5] sometimes appears in the physics literature[6] in the guise of the distributional identity

$$\frac{P}{x-y}\frac{P}{y-z} + \frac{P}{y-z}\frac{P}{z-x} + \frac{P}{z-x}\frac{P}{x-y} = -\pi^2\delta(x-y)\delta(x-z), \quad \star\star$$

[5]F. G. Tricomi, *Quart. J. Math. (Oxford)*, (2) **2**, (1951) 199.

[6]For example, in R. Jackiw, A. Strominger, *Phys. Lett.* **99B** (1981) 133.

where $P/(x-y)$ denotes the principal-part distribution $P\Big(1/(x-y)\Big)$. This attractively symmetric form conceals the fact that a specific order of integration is to be understood. As the next exercise shows, were we to freely re-arrange the integration order we could use the identity

$$\frac{1}{x-y}\frac{1}{y-z}+\frac{1}{y-z}\frac{1}{z-x}+\frac{1}{z-x}\frac{1}{x-y}=0$$

to wrongly conclude that the right-hand side is zero.

Exercise 9.7: Show that the identity $\star$ from exercise 9.6 can be written as

$$\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{\varphi_1(y)\varphi_2(z)}{(z-y)(y-x)}dz\right)dy=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{\varphi_1(y)\varphi_2(z)}{(z-y)(y-x)}dy\right)dz-\pi^2\varphi_1(x)\varphi_2(x),$$

principal-part integrals being understood where necessary. This is a special case of a more general change-of-integration-order formula

$$\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{f(x,y,z)}{(z-y)(y-x)}dz\right)dy=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{f(x,y,z)}{(z-y)(y-x)}dy\right)dz-\pi^2 f(x,x,x),$$

which is due to G. H. Hardy (1908). Show that Hardy's formula is equivalent to the distributional identity $\star\star$.

Exercise 9.8: Use the licit interchange of "P" integration with ordinary integration to show that

$$\int_{-\infty}^{\infty}\varphi(x)\left(P\int_{-\infty}^{\infty}\frac{\varphi(y)}{x-y}\,dy\right)^2dx=\frac{\pi^2}{3}\int_{-\infty}^{\infty}\varphi^3(x)\,dx.$$

Exercise 9.9: Let $f(z)$ be analytic within the unit circle, and let $u(\theta)$ and $v(\theta)$ be the boundary values of its real and imaginary parts, respectively, at $z=e^{i\theta}$. Use Plemelj to show that

$$\begin{aligned}u(\theta) &= -\frac{1}{2\pi}P\int_0^{2\pi}v(\theta')\cot\left(\frac{\theta-\theta'}{2}\right)d\theta'+\frac{1}{2\pi}\int_0^{2\pi}u(\theta')\,d\theta',\\ v(\theta) &= \frac{1}{2\pi}P\int_0^{2\pi}u(\theta')\cot\left(\frac{\theta-\theta'}{2}\right)d\theta'+\frac{1}{2\pi}\int_0^{2\pi}v(\theta')\,d\theta'.\end{aligned}$$

9.3 Partial-Fraction and Product Expansions

In this section we will study other useful representations of functions which devolve from their analyticity properties.

9.3.1 Mittag-Leffler Partial-Fraction Expansion

Let $f(z)$ be a meromorphic function with poles (perhaps infinitely many) at $z = z_j$, $(j = 1, 2, 3, \ldots)$, where $|z_1| < |z_2| < \ldots$. Let Γ_n be a contour enclosing the first n poles. Suppose further (for ease of description) that the poles are simple and have residue r_n. Then, for z inside Γ_n, we have

$$\frac{1}{2\pi i}\oint_{\Gamma_n}\frac{f(z')}{z'-z}\,dz' = f(z) + \sum_{j=1}^{n}\frac{r_j}{z_j - z}. \tag{9.81}$$

We often want to to apply this formula to trigonometric functions whose periodicity means that they do not tend to zero at infinity. We therefore employ the same *subtraction* strategy that we used for dispersion relations. We subtract

$$f(z) - f(0) = \frac{z}{2\pi i}\oint_{\Gamma_n}\frac{f(z')}{z'(z'-z)}\,dz' + \sum_{j=1}^{n} r_j\left(\frac{1}{z-z_j} + \frac{1}{z_j}\right). \tag{9.82}$$

If we now assume that $f(z)$ is uniformly bounded on the Γ_n — this meaning that $|f(z)| < A$ on Γ_n, with the same constant A working for all n — then the integral tends to zero as n becomes large, yielding the partial fraction, or *Mittag-Leffler*, decomposition

$$f(z) = f(0) + \sum_{j=1}^{\infty} r_j\left(\frac{1}{z-z_j} + \frac{1}{z_j}\right) \tag{9.83}$$

Example 1): Look at $\operatorname{cosec} z$. The residues of $1/(\sin z)$ at its poles at $z = n\pi$ are $r_n = (-1)^n$. We can take the Γ_n to be squares with corners $(n+1/2)(\pm 1 \pm i)\pi$. A bit of effort shows that cosec is uniformly bounded on them. To use the formula as given, we first need subtract the pole at $z = 0$, then

$$\operatorname{cosec} z - \frac{1}{z} = {\sum_{n=-\infty}^{\infty}}' (-1)^n\left(\frac{1}{z-n\pi} + \frac{1}{n\pi}\right). \tag{9.84}$$

The prime on the summation symbol indicates that we are omit the $n = 0$ term. The positive and negative n series converge separately, so we can add them, and write the more compact expression

$$\operatorname{cosec} z = \frac{1}{z} + 2z\sum_{1}^{\infty}(-1)^n\frac{1}{z^2 - n^2\pi^2}. \tag{9.85}$$

Example 2): A similar method gives

$$\cot z = \frac{1}{z} + {\sum_{n=-\infty}^{\infty}}' \left(\frac{1}{z - n\pi} + \frac{1}{n\pi} \right). \tag{9.86}$$

We can pair terms together to writen this as

$$\begin{aligned} \cot z &= \frac{1}{z} + \sum_{n=1}^{\infty} \left(\frac{1}{z - n\pi} + \frac{1}{z + n\pi} \right), \\ &= \frac{1}{z} + \sum_{n=1}^{\infty} \frac{2z}{z^2 - n^2\pi^2} \end{aligned} \tag{9.87}$$

or

$$\cot z = \lim_{N\to\infty} \sum_{n=-N}^{N} \frac{1}{z - n\pi}. \tag{9.88}$$

In the last formula it is important that the upper and lower limits of summation be the same. Neither the sum over positive n nor the sum over negative n converges separately. By taking asymmetric upper and lower limits we could therefore obtain any desired number as the limit of the sum.

Exercise 9.10: Use Mittag-Leffler to show that

$$\mathrm{cosec}^2 z = \sum_{n=-\infty}^{\infty} \frac{1}{(z + n\pi)^2}.$$

Now use this infinite series to give a one-line proof of the trigonometric identity

$$\sum_{m=0}^{N-1} \mathrm{cosec}^2 \left(z + \frac{m\pi}{N} \right) = N^2 \mathrm{cosec}^2(Nz).$$

(Is there a comparably easy *elementary* derivation of this finite sum?) Take a limit to conclude that

$$\sum_{m=1}^{N-1} \mathrm{cosec}^2 \left(\frac{m\pi}{N} \right) = \frac{1}{3}(N^2 - 1).$$

Exercise 9.11: From the partial fraction expansion for $\cot z$, deduce that

$$\frac{d}{dz} \ln[(\sin z)/z] = \frac{d}{dz} \sum_{n=1}^{\infty} \ln(z^2 - n^2\pi^2).$$

Integrate this along a suitable path from $z = 0$, and so conclude that that

$$\sin z = z \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2\pi^2}\right).$$

Exercise 9.12: By differentiating the partial fraction expansion for $\cot z$, show that, for k an integer ≥ 1, and $\operatorname{Im} z > 0$, we have

$$\sum_{n=-\infty}^{\infty} \frac{1}{(z+n)^{k+1}} = \frac{(-2\pi i)^{k+1}}{k!} \sum_{n=1}^{\infty} n^k e^{2\pi i n z}.$$

This is called *Lipshitz' formula*.

Exercise 9.13: The *Bernoulli numbers* are defined by

$$\frac{x}{e^x - 1} = 1 + B_1 x + \sum_{k=1}^{\infty} B_{2k} \frac{x^{2k}}{(2k)!}.$$

The first few are $B_1 = -1/2$, $B_2 = 1/6$, $B_4 = -1/30$. Except for B_1, the B_n are zero for n odd. Show that

$$x \cot x = ix + \frac{2ix}{e^{2ix} - 1} = 1 - \sum_{k=1}^{\infty} (-1)^{k+1} B_{2k} \frac{2^{2k} x^{2k}}{(2k)!}.$$

By expanding $1/(x^2 - n^2\pi^2)$ as a power series in x and comparing coefficients, deduce that, for positive integer k,

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = (-1)^{k+1} \pi^{2k} \frac{2^{2k-1}}{(2k)!} B_{2k}.$$

Exercise 9.14: *Euler-Maclaurin sum formula.* Use the formal expansion

$$\frac{D}{e^D - 1} = \sum_k B_k \frac{D^k}{k!} = 1 - \frac{1}{2}D + \frac{1}{6}\frac{D^2}{2!} - \frac{1}{30}\frac{D^4}{4!} + \cdots,$$

with D interpreted as d/dx, to obtain

$$(-f'(x) - f'(x+1) - f'(x+2) + \cdots) = f(x) - \frac{1}{2}f'(x) + \frac{1}{6}\frac{f''(x)}{2!} - \frac{1}{30}\frac{f^{(4)}}{4!} + \cdots.$$

By integrating this from a to $b \equiv a + m$, motivate the Euler-Maclaurin formula

$$\sum_{k=0}^{m-1} f(a+k) \quad = \quad \int_a^b f(x)\,dx + \frac{1}{2}(f(a) - f(b)) + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!}(f^{(2k-1)}(a) - f^{(2k-1)}(b)).$$

This "derivation," while suggestive, is only heuristic. It gives no insight into whether the series converges (it usually does not) or what the error might be if we truncate after a finite number of terms.

9.3.2 Infinite Product Expansions

We can play a variant of the Mittag-Leffler game with suitable entire functions $g(z)$ and derive for them a representation as an infinite product. Suppose that $g(z)$ has simple zeros at z_i. Then $(\ln g)' = g'(z)/g(z)$ is meromorphic with poles at z_i, all with unit residues. Assuming that it satisfies the uniform boundedness condition, we now use Mittag Leffler to write

$$\frac{d}{dz}\ln g(z) = \left.\frac{g'(z)}{g(z)}\right|_{z=0} + \sum_{j=1}^{\infty}\left(\frac{1}{z-z_j}+\frac{1}{z_j}\right). \tag{9.89}$$

Integrating up we have

$$\ln g(z) = \ln g(0) + cz + \sum_{j=1}^{\infty}\left(\ln(1-z/z_j)+\frac{z}{z_j}\right), \tag{9.90}$$

where $c = g'(0)/g(0)$. We now re-exponentiate to get

$$g(z) = g(0)e^{cz}\prod_{j=1}^{\infty}\left(1-\frac{z}{z_j}\right)e^{z/z_j}. \tag{9.91}$$

Example: Let $g(z) = \sin z/z$, then $g(0) = 1$, while the constant c, which is the logarithmic derivative of g at $z = 0$, is zero, and

$$\frac{\sin z}{z} = \prod_{n=1}^{\infty}\left(1-\frac{z}{n\pi}\right)e^{z/n\pi}\left(1+\frac{z}{n\pi}\right)e^{-z/n\pi}. \tag{9.92}$$

Thus

$$\sin z = z\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2\pi^2}\right). \tag{9.93}$$

Convergence of Infinite Products

We have derived several infinite problem formulæ without discussing the issue of their convergence. For products of terms of the form $(1+a_n)$ with positive a_n we can reduce the question of convergence to that of $\sum_{n=1}^{\infty} a_n$.

To see why this is so, let

$$p_N = \prod_{n=1}^{N}(1+a_n), \quad a_n > 0. \tag{9.94}$$

Then we have the inequalities

$$1 + \sum_{n=1}^{N} a_n < p_N < \exp\left\{\sum_{n=1}^{N} a_n\right\}. \tag{9.95}$$

The infinite sum and product therefore converge or diverge together. If

$$P = \prod_{n=1}^{\infty}(1 + |a_n|), \tag{9.96}$$

converges, we say that

$$p = \prod_{n=1}^{\infty}(1 + a_n), \tag{9.97}$$

converges absolutely. As with infinite sums, absolute convergence implies convergence, but not vice-versa. Unlike infinite sums, however, an infinite product containing negative a_n can diverge to *zero*. If $(1 + a_n) > 0$ then $\prod(1 + a_n)$ converges if $\sum \ln(1 + a_n)$ does, and we will say that $\prod(1 + a_n)$ diverges to zero if $\sum \ln(1 + a_n)$ diverges to $-\infty$.

Exercise 9.15: Show that

$$\begin{aligned} \prod_{n=1}^{N}\left(1 + \frac{1}{n}\right) &= N + 1, \\ \prod_{n=2}^{N}\left(1 - \frac{1}{n}\right) &= \frac{1}{N}. \end{aligned}$$

From these deduce that

$$\prod_{n=2}^{\infty}\left(1 - \frac{1}{n^2}\right) = \frac{1}{2}.$$

Exercise 9.16: For $|z| < 1$, show that

$$\prod_{n=0}^{\infty}\left(1 + z^{2^n}\right) = \frac{1}{1 - z}.$$

(Hint: think binary)

Exercise 9.17: For $|z| < 1$, show that

$$\prod_{n=1}^{\infty}(1 + z^n) = \prod_{n=1}^{\infty}\frac{1}{1 - z^{2n-1}}.$$

(Hint: $1 - x^{2n} = (1 - x^n)(1 + x^n)$.)

9.4 Wiener-Hopf Equations II

The theory of Hilbert transforms has shown us some the consequences of functions being analytic in the upper or lower half-plane. Another application of these ideas is to *Wiener-Hopf equations*. Although we have discussed Wiener-Hopf integral equations in chapter **??**, it is only now that we possess the tools to appreciate the general theory. We begin, however, with the slightly simpler Wiener-Hopf *sum* equations, which are their discrete analogue. Here, analyticity in the upper or lower half-plane is replaced by analyticity within or without the unit circle.

9.4.1 Wiener-Hopf Sum Equations

Consider the infinite system of equations

$$y_n = \sum_{m=-\infty}^{\infty} a_{n-m} x_m, \quad -\infty < n < \infty \tag{9.98}$$

where we are given the y_n and are seeking the x_n.

If the a_n, y_n are the Fourier coefficients of smooth complex-valued functions

$$\begin{aligned} A(\theta) &= \sum_{n=-\infty}^{\infty} a_n e^{in\theta}, \\ Y(\theta) &= \sum_{n=-\infty}^{\infty} y_n e^{in\theta}, \end{aligned} \tag{9.99}$$

then the systems of equations is, in principle at least, easy to solve. We introduce the function

$$X(\theta) = \sum_{n=-\infty}^{\infty} x_n e^{in\theta}, \tag{9.100}$$

and (9.98) becomes

$$Y(\theta) = A(\theta) X(\theta). \tag{9.101}$$

From this, the desired x_n may be read off as the Fourier expansion coefficients of $Y(\theta)/A(\theta)$. We see that $A(\theta)$ must be nowhere zero or else the operator A represented by the infinite matrix a_{n-m} will not be invertible. This technique

is a discrete version of the Fourier transform method for solving the integral equation

$$y(s) = \int_{-\infty}^{\infty} A(s-t)x(t)\,dt, \quad -\infty < s < \infty. \tag{9.102}$$

The connection with complex analysis is made by regarding $A(\theta), X(\theta), Y(\theta)$ as being functions on the unit circle in the z plane. If they are smooth enough we can extend their definition to an annulus about the unit circle, so that

$$\begin{aligned} A(z) &= \sum_{n=-\infty}^{\infty} a_n z^n, \\ X(z) &= \sum_{n=-\infty}^{\infty} x_n z^n, \\ Y(z) &= \sum_{n=-\infty}^{\infty} y_n z^n. \end{aligned} \tag{9.103}$$

The x_n may now be read off as the Laurent expansion coefficients of $Y(z)/A(z)$.

The discrete analogue of the *Wiener-Hopf integral equation*

$$y(s) = \int_{0}^{\infty} A(s-t)x(t)\,dt, \quad 0 \le s < \infty \tag{9.104}$$

is the *Wiener-Hopf sum equation*

$$y_n = \sum_{m=0}^{\infty} a_{n-m}x_m, \quad 0 \le n < \infty. \tag{9.105}$$

This requires a more sophisticated approach. If you look back at our earlier discussion of Wiener-Hopf integral equations in chapter **??**, you will see that the trick for solving them is to extend the definition $y(s)$ to negative s (analogously, the y_n to negative n) and find these values at the same time as we find $x(s)$ for positive s (analogously, the x_n for positive n.)

We proceed by introducing the same functions $A(z), X(z), Y(z)$ as before, but now keep careful track of whether their power-series expansions contain positive or negative powers of z. In doing so, we will discover that the Fredholm alternative governing the existence and uniqueness of the solutions will depend on the winding number $N = n(\Gamma, 0)$ where Γ is the image of the unit circle under the map $z \mapsto A(z)$ — in other words, on how many times $A(z)$ wraps around the origin as z goes once round the unit circle.

Suppose that $A(z)$ is smooth enough that it is analytic in an annulus including the unit circle, and that we can factorize $A(z)$ so that

$$A(z) = \lambda q_+(z) z^N [q_-(z)]^{-1}, \tag{9.106}$$

where

$$\begin{aligned} q_+(z) &= 1 + \sum_{n=1}^{\infty} q_n^+ z^n, \\ q_-(z) &= 1 + \sum_{n=1}^{\infty} q_{-n}^- z^{-n}. \end{aligned} \tag{9.107}$$

Here we demand that $q_+(z)$ be analytic and non-zero for $|z| < 1+\epsilon$, and that $q_-(z)$ be analytic and non-zero for $|1/z| < 1+\epsilon$. These no pole, no zero, conditions ensure, *via* the principle of the argument, that the winding numbers of $q_\pm(z)$ about the origin are zero, and so all the winding of $A(z)$ is accounted for by the N-fold winding of the z^N factor. The non-zero condition also ensures that the reciprocals $[q_\pm(z)]^{-1}$ have same class of expansions (*i.e.* in positive or negative powers of z only) as the direct functions.

We now introduce the notation $[F(z)]_+$ and $[F(z)]_-$, meaning that we expand $F(z)$ as a Laurent series and retain only the positive powers of z (including z^0), or only the negative powers (starting from z^{-1}), respectively. Thus $F(z) = [F(z)]_+ + [F(z)]_-$. We will write $Y_\pm(z) = [Y(z)]_\pm$, and similarly for $X(z)$. We can therefore rewrite (9.105) in the form

$$\lambda z^N q_+(z) X_+ = [Y_+(z) + Y_-(z)] q_-(z). \tag{9.108}$$

If $N \geq 0$, and we break this equation into its positive and negative powers, we find

$$\begin{aligned} [Y_+ q_-]_+ &= \lambda z^N q_+(z) X_+, \\ [Y_+ q_-]_- &= -Y_- q_-(z). \end{aligned} \tag{9.109}$$

From the first of these equations we can read off the desired x_n as the positive-power Laurent coefficients of

$$X_+(z) = [Y_+ q_-]_+ (\lambda z^N q_+(z))^{-1}. \tag{9.110}$$

As a byproduct, the second alows us to find the coefficient y_{-n} of $Y_-(z)$. Observe that there is a condition on Y_+ for this to work: the power series

expansion of $\lambda z^N q_+(z) X_+$ starts with z^N, and so for a solution to exist the first N terms of $(Y_+ q_-)_+$ as a power series in z must be zero. The given vector y_n must therefore satisfy N consistency conditions. A formal way of expressing this constraint begins by observing that it means that the range of the operator A represented by the matrix a_{n-m} falls short, by N dimensions, of the being the entire space of possible y_n. This is exactly the situation that the notion of a "cokernel" is intended to capture. Recall that if $A : V \to V$, then $\mathrm{Coker}\, A = V/\mathrm{Im}\, A$. We therefore have

$$\dim\left[\mathrm{Coker}\, A\right] = N.$$

When $N < 0$, on the other hand, we have

$$\begin{aligned} [Y_+(z)q_-(z)]_+ &= [\lambda z^{-|N|} q_+(z) X_+(z)]_+ \\ [Y_+(z)q_-(z)]_- &= -Y_-(z)q_-(z) + [\lambda z^{-|N|} q_+(z) X_+(z)]_- . \end{aligned} \tag{9.111}$$

Here the last term in the second equation contains no more than N terms. Because of the $z^{-|N|}$, we can add any to X_+ any multiple of $Z_+(x) = z^n [q_+(z)]^{-1}$ for $n = 0, \ldots, N-1$, and still have a solution. Thus the solution is not unique. Instead, we have $\dim\left[\mathrm{Ker}\,(A)\right] = |N|$.

We have therefore shown that

$$\boxed{\mathrm{Index}\,(A) \stackrel{\mathrm{def}}{=} \dim\left(\mathrm{Ker}\, A\right) - \dim\left(\mathrm{Coker}\, A\right) = -N}$$

This connection between a topological quantity – in the present case the winding number — and the difference in dimension of the kernel and cokernel is an example of an index theorem.

We now need to show that we can indeed factorize $A(z)$ in the desired manner. When $A(z)$ is a rational function, the factorization is straightforward: if

$$A(z) = C \frac{\prod_n (z - a_n)}{\prod_m (z - b_m)}, \tag{9.112}$$

we simply take

$$q_+(z) = \frac{\prod_{|a_n|>0} (1 - z/a_n)}{\prod_{|b_m|>0} (1 - z/b_m)}, \tag{9.113}$$

where the products are over the linear factors corresponding to poles and zeros outside the unit circle, and

$$q_-(z) = \frac{\prod_{|b_m|<0} (1 - b_m/z)}{\prod_{|a_n|<0} (1 - a_n/z)}, \tag{9.114}$$

containing the linear factors corresponding to poles and zeros inside the unit circle. The constant λ and the power z^N in equation (9.106) are the factors that we have extracted from the right-hand sides of (9.113) and (9.114), respectively, in order to leave 1's as the first term in each linear factor.

More generally, we take the logarithm of

$$z^{-N}A(z) = \lambda q_+(z)(q_-(z))^{-1} \tag{9.115}$$

to get

$$\ln[z^{-N}A(z)] = \ln[\lambda q_+(z)] - \ln[q_-(z)], \tag{9.116}$$

where we desire $\ln[\lambda q_+(z)]$ to be the boundary value of a function analytic within the unit circle, and $\ln[q_-(z)]$ the boundary value of function analytic outside the unit circle and with $q_-(z)$ tending to unity as $|z| \to \infty$. The factor of z^{-N} in the logarithm serves to undo the winding of the argument of $A(z)$, and results in a single-valued logarithm on the unit circle. Plemelj now shows that

$$Q(z) = \frac{1}{2\pi i}\oint_{|z|=1} \frac{\ln[\zeta^{-N}A(\zeta)]}{\zeta - z}\, d\zeta \tag{9.117}$$

provides us with the desired factorization. This function $Q(z)$ is everywhere analytic except for a branch cut along the unit circle, and its branches, Q_+ within and Q_- without the circle, differ by $\ln[z^{-N}A(z)]$. We therefore have

$$\begin{aligned} \lambda q_+(z) &= e^{Q_+(z)}, \\ q_-(z) &= e^{Q_-(z)}. \end{aligned} \tag{9.118}$$

The expression for Q as an integral shows that $Q(z) \sim const./z$ as $|z|$ goes to infinity and so guarantees that $q_-(z)$ has the desired limit of unity there.

The task of finding this factorization is known as the *scalar Riemann-Hilbert problem*. In effect, we are decomposing the infinite matrix

$$\mathbf{A} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & a_0 & a_1 & a_2 & \cdots \\ \cdots & a_{-1} & a_0 & a_1 & \cdots \\ \cdots & a_{-2} & a_{-1} & a_0 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{9.119}$$

into the product of an upper triangular matrix

$$\mathbf{U} = \lambda \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & 1 & q_1^+ & q_2^+ & \cdots \\ \cdots & 0 & 1 & q_1^+ & \cdots \\ \cdots & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{9.120}$$

a lower triangular matrix $\mathbf{L}$, where

$$\mathbf{L}^{-1} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & \cdots \\ \cdots & q_{-1}^- & 1 & 0 & \cdots \\ \cdots & q_{-2}^- & q_{-1}^- & 1 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{9.121}$$

has 1's on the diagonal, and a matrix $\mathbf{\Lambda}^N$ which which is zero everywhere except for a line of 1's located N steps above the main diagonal. The set of triangular matrices with unit diagonal form a group, so the inversion required to obtain $\mathbf{L}$ results in a matrix of the same form. The resulting *Birkhoff factorization*

$$\mathbf{A} = \mathbf{L}\mathbf{\Lambda}^N\mathbf{U}, \tag{9.122}$$

is an infinite-dimensional extension of the Gauss-Bruhat (or generalized LU) decomposition of a matrix. The finite-dimensional Gauss-Bruhat decomposition provides a factorization of a matrix $\mathbf{A} \in \mathrm{GL}(n)$ as

$$\mathbf{A} = \mathbf{L}\mathbf{\Pi}\mathbf{U}, \tag{9.123}$$

where $\mathbf{L}$ is a lower triangular matrix with 1's on the diagonal, $\mathbf{U}$ is an upper triangular matrix with no zero's on the diagonal, and $\mathbf{\Pi}$ is a permutation matrix, *i.e.* a matrix that permutes the basis vectors by having one entry of 1 in each row and in each column, and all other entries zero. Our present $\mathbf{\Lambda}^N$ is playing the role of such a matrix. The matrix $\mathbf{\Pi}$ is uniquely determined by $\mathbf{A}$. The $\mathbf{L}$ and $\mathbf{U}$ matrices become unique if $\mathbf{L}$ is chosen so that $\mathbf{\Pi}^T\mathbf{L}\mathbf{\Pi}$ is also lower triangular.

9.4.2 Wiener-Hopf Integral Equations

We now carry over our insights from the simpler sum equations to Weiner-Hopf integral equations

$$\int_0^\infty K(x-y)\phi(y)\,dy = f(x), \quad x > 0, \tag{9.124}$$

by imagining replacing the unit circle by a circle of radius R, and then taking $R \to \infty$ in such a way that the sums go over to integrals. In this way many features are retained: the problem is still solved by factorizing the Fourier transform

$$\widetilde{K}(k) = \int_{-\infty}^{\infty} K(x)e^{ikx}\,dx \tag{9.125}$$

of the kernel, and there remains an index theorem

$$\dim(\mathrm{Ker}\,K) - \dim(\mathrm{Coker}\,K) = -N, \tag{9.126}$$

but N now counts the winding of the phase of $\widetilde{K}(k)$ as k ranges over the real axis:

$$N = \frac{1}{2\pi}\arg\widetilde{K}\Big|_{k=-\infty}^{k=+\infty}. \tag{9.127}$$

One restriction arises though: we will require K to be of the form

$$K(x-y) = \delta(x-y) + g(x-y) \tag{9.128}$$

for some continuous function $g(x)$. Our discussion is therefore being restricted to Wiener-Hopf Integral equations of the *second kind*.

The restriction comes about about because we will seek to obtain a factorization of $\widetilde{K}$ as

$$\tau(\kappa)\widetilde{K}(k) = \exp\{Q_+(k) - Q_-(k)\} = q_+(k)(q_-(k))^{-1} \tag{9.129}$$

where $q_+(k) \equiv \exp\{Q_+(k)\}$ is analytic and non-zero in the upper half k-plane and $q_-(k) \equiv \exp\{Q_-(k)\}$ analytic and non-zero in the lower half-plane. The factor $\tau(\kappa)$ is a phase such as

$$\tau(k) = \left(\frac{k+i}{k-i}\right)^N, \tag{9.130}$$

which winds $-N$ times and serves serves to undo the $+N$ phase winding in $\widetilde{K}$. The $Q_\pm(k)$ will be the boundary values from above and below the real axis, respectively, of

$$Q(k) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{\ln[\tau(\kappa)\widetilde{K}(\kappa)]}{\kappa - k}\, d\kappa \tag{9.131}$$

The convergence of this infinite integral requires that $\ln[\tau(\kappa)\widetilde{K}(k)]$ go to zero at infinity, or, in other words,

$$\lim_{k\to\infty} \widetilde{K}(k) = 1. \tag{9.132}$$

This, in turn, requires that the original $K(x)$ contain a delta function.
Example: We will solve the problem

$$\phi(x) - \lambda\int_0^\infty e^{-|x-y|-\alpha(x-y)}\phi(y)\, dy = f(x), \quad x > 0. \tag{9.133}$$

We require that $0 < \alpha < 1$. The upper bound on α is necessary for the integral kernel to be bounded. We will also assume for simplicity that $\lambda < 1/2$. Following the same strategy as in the sum case, we extend the integral equation to the entire range of x by writing

$$\phi(x) - \lambda\int_0^\infty e^{-|x-y|-\alpha(x-y)}\phi(y)\, dy = f(x) + g(x), \tag{9.134}$$

where $f(x)$ is nonzero only for $x > 0$ and $g(x)$ is non-zero only for $x < 0$. The Fourier transform of this equation is

$$\left(\frac{(k+i\alpha)^2 + a^2}{(k+i\alpha)^2 + 1}\right)\widetilde{\phi}_+(k) = \widetilde{f}_+(k) + \widetilde{g}_-(k), \tag{9.135}$$

where $a^2 = 1 - 2\lambda$ and the $\pm$ subscripts are to remind us that $\widetilde{\phi}(k)$ and $\widetilde{f}(k)$ are analytic in the upper half-plane, and $\widetilde{g}(k)$ in the lower. We will use the notation H_+ for the space of functions analytic in the upper half plane, and H_- for functions analytic in the lower half plane, and so

$$\widetilde{\phi}_+(k),\ \widetilde{f}(_+k) \in H_+, \quad \widetilde{g}_-(k) \in H_- \tag{9.136}$$

We can factorize

$$\widetilde{K}(k) = \frac{(k+i\alpha)^2 + a^2}{(k+i\alpha)^2 + 1} = \frac{[k + i(\alpha - a)]}{[k + i(\alpha - 1)]}\frac{[k + i(\alpha + a)]}{[k + i(\alpha + 1)]} \tag{9.137}$$

Now suppose that a is small enough that $\alpha \pm a > 0$ and so the numerator has two zeros in the lower half plane, and the numerator a one zero in each of the upper and lower half-planes. The change of phase in $\widetilde{K}(k)$ as we go from minus to plus infinity is therefore -2π, and so the index is $N = -1$. We should therefore multiply $\widetilde{K}$ by

$$\tau(k) = \left(\frac{k+i}{k-i}\right)^{-1} \tag{9.138}$$

before seeking to break it into its $q_\pm$ factors. We can however equally well take

$$\tau(k) = \left(\frac{k+i(\alpha-1)}{k+i(\alpha-a)}\right) \tag{9.139}$$

as this also undoes the winding and allows us to factorize with

$$q_-(k) = 1, \quad q_+(k) = \left(\frac{k+i(\alpha+a)}{k+i(\alpha+1)}\right). \tag{9.140}$$

The resultant equation analagous to (9.108) is therefore

$$\begin{aligned} \left(\frac{k+i(\alpha+a)}{k+i(\alpha+1)}\right)\widetilde{\phi}_+ &= \left(\frac{k+i(\alpha-1)}{k+i(\alpha-a)}\right)\widetilde{f}_+ + \left(\frac{k+i(\alpha-1)}{k+i(\alpha-a)}\right)\widetilde{g}_- \\ q_+\widetilde{\phi}_+ &= (\tau q_-)\widetilde{f}_+ \quad + \quad \tau q_-\widetilde{g}_- \end{aligned} \tag{9.141}$$

The second line of this equation shows the interpretation of the first line in terms of the objects in the general theory. The left hand side is in H_+ — *i.e.* analytic in the upper half-plane. The first term on the right is also in H_+. (We are lucky. More generally it would have to be decomposed into its $H_\pm$ parts.) If it were not for the $\tau(\kappa)$, the last term would be in H_-, but it has a potential pole at $k = -i(\alpha - a)$. We therefore remove this pole by substracting a term

$$-\frac{\beta}{k+i(\alpha-a)}$$

(an element of H_+) from each side of the equation before projecting onto the $H^\pm$ parts. After projecting, we find that

$$\begin{aligned} H_+ : &\qquad \left(\frac{k+i(\alpha+a)}{k+i(\alpha+1)}\right)\widetilde{\phi}_+ - \left(\frac{k+i(\alpha-1)}{k+i(\alpha-a)}\right)\widetilde{f}_+ - \frac{\beta}{k+i(\alpha-a)} = 0, \\ H_- : &\qquad \left(\frac{k+i(\alpha-1)}{k+i(\alpha-a)}\right)\widetilde{g}_- - \frac{\beta}{k+i(\alpha-a)} = 0. \end{aligned} \tag{9.142}$$

We solve for $\widetilde{\phi}_+(k)$ and $\widetilde{g}_-(k)$

$$\begin{aligned}\widetilde{\phi}_+(k) &= \left(\frac{(k+i\alpha)^2+1}{(k+i\alpha)^2+a^2}\right)\widetilde{f}_- - \beta\left(\frac{k+i(\alpha+1)}{(k+i\alpha)^2+a^2}\right)\\ \widetilde{g}_-(k) &= \frac{\beta}{k+i(\alpha-1)}.\end{aligned} \tag{9.143}$$

Observe $g_-(k)$ is always in H_- because its only singularity is in the upper half-plane for any β. The constant β is therefore arbitrary. Finally, we invert the Fourier transform, using

$$\mathcal{F}\left(\theta(x)e^{-\alpha x}\sinh ax\right) = -\frac{a}{(k+i\alpha)^2+a^2}, \quad (\alpha \pm a) > 0, \tag{9.144}$$

to find that

$$\begin{aligned}\phi(x) &= f(x) - \frac{2\lambda}{a}\int_0^x e^{-\alpha(x-y)}\sinh a(x-y)f(y)\,dy\\ &\quad +\beta'\left\{(a-1)e^{-(\alpha+a)x} + (a+1)e^{-(\alpha-a)x}\right\},\end{aligned} \tag{9.145}$$

where β' (proportional to β) is an arbitrary constant.

By taking α in the range $-1 < \alpha < 0$ with $(\alpha \pm a) < 0$, we make index to be $N = +1$. We will then find there is condition on $f(x)$ for the solution to exist. This condition is, of course, that $f(x)$ be orthogonal to the solution

$$\phi_0(x) = \left\{(a-1)e^{-(\alpha+a)x} + (a+1)e^{-(\alpha-a)x}\right\} \tag{9.146}$$

of the homogenous adjoint problem, this being the $f(x) = 0$ case of the $\alpha > 0$ problem that we have just solved.

9.5 Further Exercises and Problems

Exercise 9.18: *Contour Integration*: Use the calculus of residues to evaluate the following integrals:

$$\begin{aligned}I_1 &= \int_0^{2\pi}\frac{d\theta}{(a+b\cos\theta)^2}, \qquad 0<b<a.\\ I_2 &= \int_0^{2\pi}\frac{\cos^2 3\theta}{1-2a\cos 2\theta+a^2}d\theta, \qquad 0<a<1.\\ I_3 &= \int_0^{\infty}\frac{x^\alpha}{(1+x^2)^2}dx, \qquad -1<\alpha<2.\end{aligned}$$

These are not meant to be easy! You will have to dig for the residues.

Answers:

$$\begin{aligned} I_1 &= \frac{2\pi a}{(a^2-b^2)^{3/2}}, \\ I_2 &= \frac{\pi(a^3+1)}{a^2-1} = \frac{\pi(1-a+a^2)}{a-1}, \\ I_3 &= \frac{\pi(1-\alpha)}{4\cos(\pi\alpha/2)}. \end{aligned}$$

Exercise 9.19: By considering the integral of

$$f(z) = \ln(1-e^{2iz}) = \ln(-2ie^{iz}\sin z)$$

around the indented rectangle

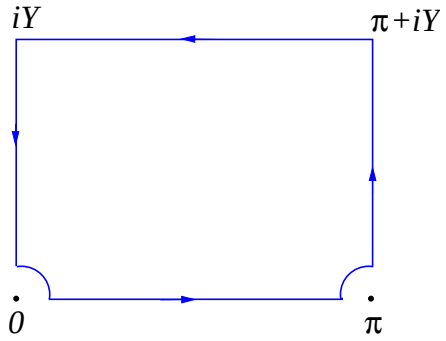

Figure 9.12: *Indented rectangle.*

with vertices 0, π, $\pi+iY$, iY, and letting Y become large, evaluate the integral

$$I = \int_0^\pi \ln(\sin x)\,dx.$$

Explain how the fact that $\epsilon\ln\epsilon \to 0$ as $\epsilon \to 0$ allows us to ignore contributions from the small indentations. You should also provide justification for any other discarded contributions. Take care to make consistent choices of the branch of the logarithm, especially if expanding $\ln(-2ie^{ix}\sin x) = ix + \ln 2 + \ln(\sin x) + \ln(-i)$. The value of I is a real number.

Exercise 9.20: By integrating a suitable function around the quadrant containing the point $z_0 = e^{i\pi/4}$, evaluate the integral

$$I(\alpha) = \int_0^\infty \frac{x^{\alpha-1}}{1+x^4} dx \qquad 0 < \alpha < 4.$$

(It should only be necessary to consider the residue at z_0.)

Exercise 9.21: In section **??** we considered the causal Green function for the damped harmonic oscillator

$$G(t) = \begin{cases} \frac{1}{\Omega} e^{-\gamma t} \sin(\Omega t), & t > 0, \\ 0, & t < 0, \end{cases}$$

and showed that its Fourier transform

$$\int_{-\infty}^{\infty} e^{i\omega t} G(t)\, dt = \frac{1}{\Omega^2 - (\omega + i\gamma)^2}, \tag{9.147}$$

had no singularities in the upper half-plane. Use Jordan's lemma to compute the inverse Fourier transform

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-i\omega t}}{\Omega^2 - (\omega + i\gamma)^2}\, d\omega,$$

and verify that it reproduces $G(t)$.

Problem 9.22: *Jordan's Lemma and one-dimensional scattering theory*. In problem **??**.**??** we considered the one-dimensional scattering problem solutions

$$\begin{aligned} \psi_k(x) &= \begin{cases} e^{ikx} + r_L(k) e^{-ikx}, & x \in L, \\ t_L(k) e^{ikx}, & x \in R, \end{cases} \qquad k > 0. \\ &= \begin{cases} t_R(k) e^{ikx}, & x \in L, \\ e^{ikx} + r_R(k) e^{-ikx}, & x \in R. \end{cases} \qquad k < 0. \end{aligned}$$

and claimed that the bound-state contributions to the completeness relation were given in terms of the reflection and transmission coefficients as

$$\begin{aligned} \sum_{\text{bound}} \psi_n^*(x) \psi_n(x') &= -\int_{-\infty}^{\infty} \frac{dk}{2\pi} r_L(k) e^{-ik(x+x')}, \qquad x, x' \in L, \\ &= -\int_{-\infty}^{\infty} \frac{dk}{2\pi} t_L(k) e^{-ik(x-x')}, \qquad x \in L,\ x' \in R, \\ &= -\int_{-\infty}^{\infty} \frac{dk}{2\pi} t_R(k) e^{-ik(x-x')}, \qquad x \in R,\ x' \in L, \\ &= -\int_{-\infty}^{\infty} \frac{dk}{2\pi} r_R(k) e^{-ik(x+x')}, \qquad x, x' \in R. \end{aligned}$$

The eigenfunctions

$$\psi_k^{(+)}(x) = \begin{cases} e^{ikx} + r_L(k)e^{-ikx}, & x \in L, \\ t_L(k)e^{ikx}, & x \in R, \end{cases}$$

and

$$\psi_k^{(-)}(x) = \begin{cases} t_R(k)e^{ikx}, & x \in L, \\ e^{ikx} + r_R(k)e^{-ikx}, & x \in R. \end{cases}$$

are initially refined for k real and positive ($\psi_k^{(+)}$) or for k real and negative ($\psi_k^{(-)}$), but they separately have analytic continuations to all of $k \in \mathbb{C}$. The reflection and transmission coefficients $r_{L,R}(k)$ and $t_{L,R}(k)$ are also analytic functions of k, and obey $r_{L,R}(k) = r_{L,R}^*(-k^*)$, $t_{L,R}(k) = t_{L,R}^*(-k^*)$.

a) By inspecting the formulæ for $\psi_k^{(+)}(x)$, show that the bound states $\psi_n(x)$, with $E_n = -\kappa_n^2$, are proportional to $\psi_k^{(+)}(x)$ evaluated at points $k = i\kappa_n$ on the positive imaginary axis at which $r_L(k)$ and $t_L(k)$ simultaneously have poles. Similarly show that these same bound states are proportional to $\psi_k^{(-)}(x)$ evaluated at points $-i\kappa_n$ on the *negative* imaginary axis at which $r_R(k)$ and $t_R(k)$ have poles. (All these functions $\psi_k^{(\pm)}(x)$, $r_{R,L}(k)$, $t_{R,L}(k)$, may have branch points and other singularities in the half-plane on the opposite side of the real axis from the bound-state poles.)

b) Use Jordan's lemma to evaluate the Fourier transforms given above in terms of the position and residues of the bound-state poles. Confirm that your answers are of the form

$$\sum_n A_n^*[\mathrm{sgn}(x)]e^{-\kappa_n|x|}A_n[\mathrm{sgn}(x')]e^{-\kappa_n|x'|},$$

as you would expect for the bound-state contribution to the completeness relation.

Exercise 9.23: *Lattice Matsubara sums*: Show that sums over the N-th roots of -1 can be written as an integral

$$\frac{1}{N}\sum_{\omega^N+1=0} f(\omega) = \frac{1}{2\pi i}\int_C \frac{dz}{z}\frac{z^N}{z^N+1}f(z),$$

where C consists of a pair of oppositely oriented concentric circles. The annulus formed by the circles should include all the roots of unity, but exclude all singularites of f. Use this trick to show that, for N even,

$$\frac{1}{N}\sum_{n=0}^{N-1}\frac{\sinh E}{\sinh^2 E + \sin^2\frac{(2n+1)\pi}{N}} = \frac{1}{\cosh E}\tanh\frac{NE}{2}.$$

Take the $N \to \infty$ limit in some suitable manner, and hence show that

$$\sum_{n=-\infty}^{\infty} \frac{a}{a^2 + [(2n+1)\pi]^2} = \frac{1}{2}\tanh\frac{a}{2}.$$

(Hint: If you are careless, you will end up differing by a factor of two from this last formula. There are *two* regions in the finite sum that tend to the infinite sum in the large N limit.)

Problem 9.24: If we define $\chi(h) = e^{\alpha x}\phi(x)$, and $F(x) = e^{\alpha x}f(x)$, then the Wiener-Hopf equation

$$\phi(x) - \lambda \int_0^\infty e^{-|x-y|-\alpha(x-y)}\phi(y)\,dy = f(x), \quad x > 0.$$

becomes

$$\chi(x) - \lambda \int_0^\infty e^{-|x-y|}\chi(y)\,dy = F(x), \quad x > 0,$$

all mention of α having disappeared! Why then does our answer, worked out in such detail, in section 9.4.2 depend on the parameter α? Show that if α small enough that $\alpha + a$ is positive and $\alpha - a$ is negative, then $\phi(x)$ really is independent of α. (Hint: What tacit assumptions about function spaces does our use of Fourier transforms entail? How does the inverse Fourier transform of $[(k + i\alpha)^2 + a^2]^{-1}$ vary with α?)

Chapter 10

Special Functions II

In this chapter we will apply complex analytic methods so as to obtain a wider view of some of the special functions of mathematical physics than can be obtained on the real axis. The standard text in this field remains the venerable *Course of Modern Analysis* of E. T. Whittaker and G. N. Watson.

10.1 The Gamma Function

We begin with Euler's "Gamma Function" $\Gamma(z)$. You probably have some acquaintance with this creature. The usual definition is

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \quad \operatorname{Re} z > 0, \quad \text{(definition A)}. \tag{10.1}$$

An integration by parts, based on

$$\frac{d}{dt}\left(t^z e^{-t}\right) = z t^{z-1} e^{-t} - t^z e^{-t}, \tag{10.2}$$

shows that

$$\left[t^z e^{-t}\right]_0^\infty = z \int_0^\infty t^{z-1} e^{-t}\, dt - \int_0^\infty t^z e^{-t}\, dt. \tag{10.3}$$

The integrated out part vanishes at both limits, provided the real part of z is greater than zero. Thus

$$\Gamma(z+1) = z\Gamma(z). \tag{10.4}$$

Since $\Gamma(1) = 1$, we deduce that

$$\Gamma(n) = (n-1)!, \quad n = 1, 2, 3, \cdots. \tag{10.5}$$

We can use the recurrence relation to extend the definition of $\Gamma(z)$ to the left half plane, where the real part of z is negative. Choosing an integer n such that the real part of $z + n$ is positive, we write

$$\Gamma(z) = \frac{\Gamma(z+n)}{z(z+1)\cdots(z+n-1)}. \tag{10.6}$$

We see that $\Gamma(z)$ has poles at zero, and at the negative integers. The residue of the pole at $z = -n$ is $(-1)^n/n!$.

We can also view the analytic continuation as an example of Taylor series subtraction. Let us recall how this works. Suppose that $-1 < \mathrm{Re}\, x < 0$. Then, from

$$\frac{d}{dt}(t^x e^{-t}) = x t^{x-1} e^{-t} - t^x e^{-t} \tag{10.7}$$

we have

$$\left[t^x e^{-t}\right]_\epsilon^\infty = x \int_\epsilon^\infty dt\, t^{x-1} e^{-t} - \int_\epsilon^\infty dt\, t^x e^{-t}. \tag{10.8}$$

Here we have cut off the integral at the lower limit so as to avoid the divergence near $t = 0$. Evaluating the left-hand side and dividing by x we find

$$-\frac{1}{x}\epsilon^x = \int_\epsilon^\infty dt\, t^{x-1} e^{-t} - \frac{1}{x}\int_\epsilon^\infty dt\, t^x e^{-t}. \tag{10.9}$$

Since, for this range of x,

$$-\frac{1}{x}\epsilon^x = \int_\epsilon^\infty dt\, t^{x-1}, \tag{10.10}$$

we can rewrite (10.9) as

$$\frac{1}{x}\int_\epsilon^\infty dt\, t^x e^{-t} = \int_\epsilon^\infty dt\, t^{x-1}\left(e^{-t} - 1\right). \tag{10.11}$$

The integral on the right-hand side of this last expression is convergent as $\epsilon \to 0$, so we may safely take the limit and find

$$\frac{1}{x}\Gamma(x+1) = \int_0^\infty dt\, t^{x-1}\left(e^{-t} - 1\right). \tag{10.12}$$

Since the left-hand side is equal to $\Gamma(x)$, we have shown that

$$\Gamma(x) = \int_0^\infty dt\, t^{x-1}\left(e^{-t} - 1\right), \qquad -1 < \mathrm{Re}\, x < 0. \tag{10.13}$$

Similarly, if $-2 < \mathrm{Re}\, x < -1$, we can show that

$$\Gamma(x) = \int_0^\infty dt\, t^{x-1}\left(e^{-t} - 1 + t\right). \tag{10.14}$$

Thus the analytic continuation of the original integral is given by a new integral in which we have subtracted exactly as many terms from the Taylor expansion of e^{-t} as are needed to just make the integral convergent.

Other useful identities, usually proved by elementary real-variable methods, include Euler's "Beta function" identity,

$$B(a,b) \stackrel{\mathrm{def}}{=} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 (1-t)^{a-1} t^{b-1}\, dt \tag{10.15}$$

(which, as the *Veneziano formula*, was the original inspiration for string theory) and

$$\Gamma(z)\Gamma(1-z) = \pi \mathrm{cosec}\, \pi z. \tag{10.16}$$

The proofs of both formulæ begin in the same way: set $t = y^2$, x^2, so that

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= 4\int_0^\infty y^{2a-1} e^{-y^2}\, dy \int_0^\infty x^{2b-1} e^{-x^2}\, dx \\
&= 4\int_0^\infty\int_0^\infty e^{-(x^2+y^2)} x^{2b-1} y^{2a-1}\, dx dy \\
&= 2\int_0^\infty e^{-r^2} (r^2)^{a+b-1}\, d(r^2) \int_0^{\pi/2} \sin^{2a-1}\theta \cos^{2b-1}\theta\, d\theta.
\end{aligned}$$

We have appealed to Fubini's theorem twice: once to turn a product of integrals into a double integral, and once (after setting $x = r\cos\theta$, $y = r\sin\theta$) to turn the double integral back into a product of decoupled integrals. In the second factor of the third line we can now change variables to $t = \sin^2\theta$ and obtain the Beta function identity. If, on the other hand, we put $a = 1-z$, $b = z$ we have

$$\Gamma(z)\Gamma(1-z) = 2\int_0^\infty e^{-r^2}\, d(r^2) \int_0^{\pi/2} \cot^{2z-1}\theta\, d\theta = 2\int_0^{\pi/2} \cot^{2z-1}\theta\, d\theta. \tag{10.17}$$

Now set $\cot\theta = \zeta$. The last integral then becomes (see exercise 9.1):

$$2\int_0^\infty \frac{\zeta^{2z-1}}{\zeta^2+1}\,d\zeta = \pi \operatorname{cosec}\pi z, \quad 0 < z < 1. \tag{10.18}$$

Although this integral has a restriction on the range of z, the result (10.16) can be analytically continued to so as to hold for all z . If we put $z = 1/2$ we find that $(\Gamma(1/2))^2 = \pi$. The positive square root is the correct one, and

$$\Gamma(1/2) = \sqrt{\pi}. \tag{10.19}$$

The integral in definition A is only convergent for $\operatorname{Re} z > 0$. A more powerful definition, involving an integral which converges for all z, is

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\int_C \frac{e^t}{t^z}\,dt. \quad \text{(definition B)} \tag{10.20}$$

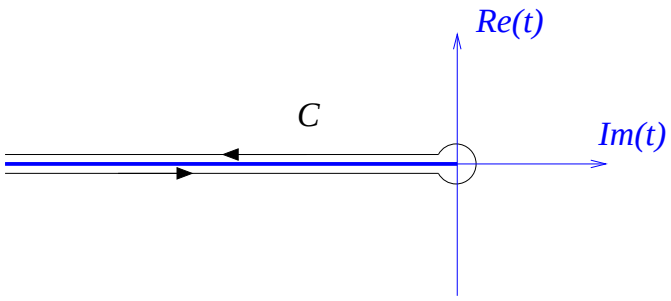

Figure 10.1: *Definition "B" contour for* $\Gamma(z)$.

Here C is a contour originating at $z = -\infty - i\epsilon$, below the negative real axis (on which a cut serves to make t^{-z} single valued) rounding the origin, and then heading back to $z = -\infty + i\epsilon$ — this time staying above the cut. We take $\arg t$ to be $+\pi$ immediately above the cut, and $-\pi$ immediately below it. This new definition is due to Hankel.

For z an integer, the cut is ineffective and we can close the contour to find

$$\frac{1}{\Gamma(0)} = 0; \qquad \frac{1}{\Gamma(n)} = \frac{1}{(n-1)!}, \quad n > 0. \tag{10.21}$$

Thus definitions A and B agree on the integers. It is less obvious that they agree for all z. A hint that this is true stems integrating by parts

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\left[\frac{e^t}{(z-1)t^{z-1}}\right]_{-\infty-i\epsilon}^{-\infty+i\epsilon} + \frac{1}{(z-1)2\pi i}\int_C \frac{e^t}{t^{z-1}}\,dt = \frac{1}{(z-1)\Gamma(z-1)}. \tag{10.22}$$

The integrated out part vanishes because e^t is zero at $-\infty$. Thus the "new" gamma function obeys the same functional relation as the "old" one.

To show the equivalence in general we will examine the definition B expression for $\Gamma(1-z)$

$$\frac{1}{\Gamma(1-z)} = \frac{1}{2\pi i}\int_C e^t t^{z-1}\,dt. \tag{10.23}$$

We will asume initially that $\mathrm{Re}\, z > 0$, so that there is no contribution from the small circle about the origin. We can therefore focus on contribution from the discontinuity across the cut

$$\begin{aligned}\frac{1}{\Gamma(1-z)} &= \frac{1}{2\pi i}\int_C e^t t^{z-1}\,dt = -\frac{1}{2\pi i}(2i\sin\pi(z-1))\int_0^\infty t^{z-1}e^{-t}\,dt \\ &= \frac{1}{\pi}\sin\pi z\int_0^\infty t^{z-1}e^{-t}\,dt. \end{aligned}\tag{10.24}$$

The proof is then completed by using $\Gamma(z)\Gamma(1-z) = \pi\mathrm{cosec}\,\pi z$, which we proved using definition A, to show that, under definition A, the right hand side is indeed equal to $1/\Gamma(1-z)$. We now use the uniqueness of analytic continuation, noting that if two analytic functions agree on the region $\mathrm{Re}\, z > 0$, then they agree everywhere.

Infinite Product for $\Gamma(z)$

The function $\Gamma(z)$ has poles at $z = 0, -1, -2, \ldots$ therefore $(z\Gamma(z))^{-1} = (\Gamma(z+1))^{-1}$ has zeros as $z = -1, -2, \ldots$. Furthermore the integral in "definition B" converges for all z, and so $1/\Gamma(z)$ has no singularities in the finite z plane *i.e.* it is an entire function. Thus means that we can use the infinite product formula

$$g(z) = g(0)e^{cz}\prod_1^\infty \left\{\left(1-\frac{z}{z_j}\right)e^{z/z_j}\right\} \tag{10.25}$$

for entire functions.

We need to recall the definition of Euler-Mascheroni constant $\gamma = -\Gamma'(1) = .5772157\ldots$, and that $\Gamma(1) = 1$. Then

$$\frac{1}{\Gamma(z)} = ze^{\gamma z} \prod_1^\infty \left\{ \left(1 + \frac{z}{n}\right) e^{-z/n} \right\}. \tag{10.26}$$

We can use this formula to compute

$$\begin{aligned} \frac{1}{\Gamma(z)\Gamma(1-z)} = \frac{1}{(-z)\Gamma(z)\Gamma(-z)} &= z \prod_1^\infty \left\{ \left(1 + \frac{z}{n}\right) e^{-z/n} \left(1 - \frac{z}{n}\right) e^{z/n} \right\} \\ &= z \prod_1^\infty \left(1 - \frac{z^2}{n^2}\right) \\ &= \frac{1}{\pi} \sin \pi z \end{aligned}$$

and so obtain another demonstration that $\Gamma(z)\Gamma(1-z) = \pi \mathrm{cosec}\, \pi z$.

Exercise 10.1: Starting from the infinite product formula for $\Gamma(z)$, show that

$$\frac{d^2}{dz^2} \ln \Gamma(z) = \sum_{n=0}^\infty \frac{1}{(z+n)^2}.$$

(Compare this "half series", with the expansion

$$\pi^2 \mathrm{cosec}^2 \pi z = \sum_{n=-\infty}^\infty \frac{1}{(z+n)^2}.)$$

10.2 Linear Differential Equations

When a linear differential equation has meromorphic coeffecients, its solutions can be extended off the real line and into the complex plane. The broader horizon then allows us to see much more of their structure.

10.2.1 Monodromy

Consider the linear differential equation

$$Ly \equiv y'' + p(z)y' + q(z)y = 0, \tag{10.27}$$

where p and q are meromorphic. Recall that the point $z = a$ is a *regular singular point* of the equation if p or q is singular there, but

$$(z-a)p(z), \quad (z-a)^2 q(z) \tag{10.28}$$

are both analytic at $z = a$. We know, from the explicit construction of power series solutions, that near a regular singular point y is a sum of functions of the form $y = (z-a)^\alpha \varphi(z)$ or $y = (z-a)^\alpha(\ln(z-a)\varphi(z) + \chi(z))$, where both $\varphi(z)$ and $\chi(z)$ are analytic near $z = a$. We now examine this fact in a more topological way.

Suppose that y_1 and y_2 are linearly independent solutions of $Ly = 0$. Start from some ordinary (non-singular) point of the equation and analytically continue the solutions round the singularity at $z = a$ and back to the starting point. The continued functions $\tilde{y}_1$ and $\tilde{y}_2$ will not in general coincide with the original solutions, but being still solutions of the equation, must be linear combinations of them. Therefore

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \tag{10.29}$$

for some constants a_{ij}. By a suitable redefinition of the y_i we may either diagonalise this *monodromy* matrix to find

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{10.30}$$

or, if the eigenvalues coincide and the matrix is not diagonalizable, reduce it to a Jordan form

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{10.31}$$

These equations are satisfied, in the diagonalizable case, by functions of the form

$$y_1 = (z-a)^{\alpha_1}\varphi_1(z), \quad y_2 = (z-a)^{\alpha_2}\varphi_2(z), \tag{10.32}$$

where $\lambda_k = e^{2\pi i \alpha_k}$, and $\varphi_k(z)$ is single valued near $z = a$. In the Jordan-form case we must have

$$y_1 = (z-a)^\alpha \left[\varphi_1(z) + \frac{1}{2\pi i \lambda}\ln(z-a)\varphi_2(z)\right], \quad y_2 = (z-a)^\alpha \varphi_2(z), \tag{10.33}$$

where again the $\varphi_k(z)$ are single valued. Notice that coincidence of the monodromy eigenvalues λ_1 and λ_2 does not require the exponents α_1 and α_2

to be the same, only that they differ by an integer. This is the same condition that signals the presence of a logarithm in the traditional series solution.

The occurrence of fractional powers and logarithms in solutions near a regular singular point is therefore quite natural.

10.2.2 Hypergeometric Functions

Most of the special functions of Mathematical Physics are special cases of the hypergeometric function $F(a,b;c;z)$, which may be defined by the series

$$\begin{aligned} F(a,b;c;z) &= 1+\frac{a.b}{1.c}z+\frac{a(a+1)b(b+1)}{2!c(c+1)}z^2+ \\ &\qquad +\frac{a(a+1)(a+2)b(b+1)(b+2)}{3!c(c+1)(c+2)}z^3+\cdots. \\ &= \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_0^\infty \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)\Gamma(1+n)}z^n. \end{aligned} \tag{10.34}$$

For general values of a,b,c, this series converges for $|z|<1$, the singularity restricting the convergence being a branch point at $z=1$.
Examples:

$$(1+z)^n = F(-n,b;b;-z), \tag{10.35}$$

$$\ln(1+z) = zF(1,1;2;-z), \tag{10.36}$$

$$z^{-1}\sin^{-1}z = F\left(\frac{1}{2},\frac{1}{2};\frac{3}{2};z^2\right), \tag{10.37}$$

$$e^z = \lim_{b\to\infty} F(1,b;1/b;z/b), \tag{10.38}$$

$$P_n(z) = F\left(-n,n+1;1;\frac{1-z}{2}\right), \tag{10.39}$$

where in the last line P_n is the Legendre polynomial.

For future reference, note that expanding the right hand side as a powers series in z and integrating term by term shows that

$$F(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 (1-tz)^{-a}t^{b-1}(1-t)^{c-b-1}dt. \tag{10.40}$$

If $\mathrm{Re}\,c > \mathrm{Re}\,(a+b)$, we may set $z=1$ in this integral to get

$$F(a,b;c;1)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}. \tag{10.41}$$

The hypergeometric function is a solution of the second-order differential equation

$$z(1-z)y'' + [c - (a+b+1)z]y' - aby = 0. \tag{10.42}$$

this equation has regular singular points at $z = 0, 1, \infty$. Provided that $1 - c$ is not an integer, the general solution is

$$y = AF(a, b; c; z) + Bz^{1-c}F(b - c + 1, a - c + 1; 2 - c; z). \tag{10.43}$$

The hypergeometric equation is a particular case of the general *Fuchsian equation* having three[1] regular singularities at $z = z_1, z_2, z_3$. This equation is

$$y'' + P(z)y' + Q(z)y = 0, \tag{10.44}$$

where

$$\begin{aligned} P(z) &= \left(\frac{1-\alpha-\alpha'}{z-z_1} + \frac{1-\beta-\beta'}{z-z_2} + \frac{1-\gamma-\gamma'}{z-z_3}\right) \\ Q(z) &= \frac{1}{(z-z_1)(z-z_2)(z-z_3)} \times \\ &\quad \left(\frac{(z_1-z_2)(z_1-z_3)\alpha\alpha'}{z-z_1} + \frac{(z_2-z_3)(z_2-z_1)\beta\beta'}{z-z_2} + \frac{(z_3-z_1)(z_3-z_2)\gamma\gamma'}{z-z_3}\right). \end{aligned} \tag{10.45}$$

The parameters are subject to the constraint $\alpha + \beta + \gamma + \alpha' + \beta' + \gamma' = 1$, which ensures that $z = \infty$ is not a singular point of the equation. This

[1]The Fuchsian equation with *two* regular singularities is

$$y'' + p(z)y' + q(z)y = 0$$

with

$$\begin{aligned} p(z) &= \left(\frac{1-\alpha-\alpha'}{z-z_1} + \frac{1+\alpha+\alpha'}{z-z_2}\right) \\ q(z) &= \frac{\alpha\alpha'(z_1-z_2)^2}{(z-z_1)^2(z-z_2)^2}. \end{aligned}$$

Its general solution is

$$y = A\left(\frac{z-z_1}{z-z_2}\right)^{\alpha} + B\left(\frac{z-z_1}{z-z_2}\right)^{\alpha'}.$$

equation is sometimes called *Riemann's P-equation*. The P probably stands for Papperitz, who discovered it.

The indicial equation relative to the regular singular point at z_1 is

$$r(r-1) + (1-\alpha-\alpha')r + \alpha\alpha' = 0, \tag{10.46}$$

and has roots $r = \alpha, \alpha'$. From this we deduce that Riemann's equation has solutions which behave like $(z-z_1)^\alpha$ and $(z-z_1)^{\alpha'}$ near z_1. Similarly, there are solutions that behave like $(z-z_2)^\beta$ and $(z-z_2)^{\beta'}$ near z_2, and like $(z-z_3)^\gamma$ and $(z-z_3)^{\gamma'}$ near z_3. The solution space of Riemann's equation is traditionally denoted by the Riemann "P" symbol

$$y = P\left\{\begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array}\right\} \tag{10.47}$$

where the six quantities $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$, are called the *exponents* of the solution. A particular solution is

$$y = \left(\frac{z-z_1}{z-z_2}\right)^\alpha \left(\frac{z-z_3}{z-z_2}\right)^\gamma F\left(\alpha+\beta+\gamma, \alpha+\beta'+\gamma; 1+\alpha-\alpha'; \frac{(z-z_1)(z_3-z_2)}{(z-z_2)(z_3-z_1)}\right). \tag{10.48}$$

By permuting the triples (z_1, α, α'), (z_2, β, β'), (z_3, γ, γ'), and within them interchanging the pairs $\alpha \leftrightarrow \alpha'$, $\gamma \leftrightarrow \gamma'$, we may find a total[2] of $6 \times 4 = 24$ solutions of this form. They are called the *Kummer* solutions. Only two of these can be linearly independent, and a large part of the theory of special functions is devoted to obtaining the linear relations between them.

It is straightforward, but a trifle tedious, to show that

$$(z-z_1)^r(z-z_2)^s(z-z_3)^t P\left\{\begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array}\right\} = P\left\{\begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha+r & \beta+s & \gamma+t & z \\ \alpha'+r & \beta'+s & \gamma'+t & \end{array}\right\} \tag{10.49}$$

provided $r+s+t=0$. Riemann's equation retains its form under Möbius maps, only the location of the singular points changing. We therefore deduce that

$$P\left\{\begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array}\right\} = P\left\{\begin{array}{cccc} z_1' & z_2' & z_3' & \\ \alpha & \beta & \gamma & z' \\ \alpha' & \beta' & \gamma' & \end{array}\right\} \tag{10.50}$$

[2]The interchange $\beta \leftrightarrow \beta'$ leaves the hypergeometric function invariant, and so does not give a new solution.

where

$$z' = \frac{az+b}{cz+d}, \quad z_1' = \frac{az_1+b}{cz_1+d}, \quad z_2' = \frac{az_2+b}{cz_2+d}, \quad z_3' = \frac{az_3+b}{cz_3+d}. \tag{10.51}$$

By using the Möbius map which takes $(z_1, z_2, z_3) \to (0, 1, \infty)$, and by extracting powers to shift the exponents, we can reduce the general eight-parameter Riemann equation to the three-parameter hypergeometric equation.

The P symbol for the hypergeometric equation is

$$F(a,b;c;z) = P\left\{\begin{array}{cccc} 0 & \infty & 1 & \\ 0 & a & 0 & z \\ 1-c & b & c-a-b & \end{array}\right\}. \tag{10.52}$$

Using this observation and a suitable Möbius map we see that

$$F(a,b;a+b-c;1-z)$$

and

$$(1-z)^{c-a-b}F(c-b,c-a;c-a-b+1;1-z)$$

are also solutions of the Hypergeometric equation, each having a pure (as opposed to a linear combination of) power-law behaviors near $z = 1$. (The previous solutions had pure power-law behaviours near z=0.) These new solutions must be linear combinations of the old, and we may use

$$F(a,b;c;1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}, \quad \mathrm{Re}\,(c-a-b) > 0, \tag{10.53}$$

together with the trick of substituting $z = 0$ and $z = 1$, to determine the coefficients and show that

$$\begin{aligned} F(a,b;c;z) &= \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}F(a,b;a+b-c;1-z) \\ &\quad + \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}(1-z)^{c-a-b}F(c-b,c-a;c-a-b+1;1-z). \end{aligned} \tag{10.54}$$

This last equation holds for all values of a, b, c such that the gamma functions make sense.

A complete set of pure-power solutions can be taken to be

$$\begin{aligned}
\phi_0^{(0)}(z) &= F(a,b;c;z) \\
\phi_0^{(1)}(z) &= z^{1-c}F(a+1-c,b+1-c;2-c;z) \\
\phi_1^{(0)}(z) &= F(a,b;1-c+a+b;1-z) \\
\phi_1^{(1)}(z) &= (1-z)^{c-a-b}F(c-a,c-b;1+c-a-b;1-z) \\
\phi_\infty^{(0)}(z) &= z^{-a}F(a,a+1-c;1+a-b;z^{-1}) \\
\phi_\infty^{(1)}(z) &= z^{-b}F(a,b+1-c;1-a+b;z^{-1}),
\end{aligned} \tag{10.55}$$

The connection coefficients are then

$$\begin{aligned}
\phi_0^{(0)} &= \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}\phi_1^{(0)} + \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}\phi_1^{(1)}, \\
\phi_0^{(1)} &= \frac{\Gamma(2-c)\Gamma(c-a-b)}{\Gamma(1-a)\Gamma(1-b)}\phi_1^{(0)}\frac{\Gamma(2-c)\Gamma(a+b-c)}{\Gamma(a+1-c)\Gamma(b+1-c)}\phi_1^{(1)},
\end{aligned} \tag{10.56}$$

and

$$\begin{aligned}
\phi_0^{(0)} &= e^{-i\pi a}\frac{\Gamma(c)\Gamma(b-a)}{\Gamma(c-a)\Gamma(b)}\phi_\infty^{(0)} + e^{-i\pi b}\frac{\Gamma(2-c)\Gamma(a-b)}{\Gamma(a+1-c)\Gamma(1-b)}\phi_\infty^{(1)}, \\
\phi_0^{(1)} &= e^{-i\pi(a+1-c)}\frac{\Gamma(2-c)\Gamma(b-a)}{\Gamma(b+1-c)\Gamma(1-a)}\phi_\infty^{(0)} + e^{-i\pi(b+1-c)}\frac{\Gamma(2-c)\Gamma(a-b)}{\Gamma(a+1-c)\Gamma(1-b)}\phi_\infty^{(1)}.
\end{aligned} \tag{10.57}$$

These relations assume that $\mathrm{Im} z > 0$. The signs in the exponential factors must be reversed when $\mathrm{Im}\, z < 0$.

Example: The Pöschel-Teller problem for general positive l. A substitution $z = (1+e^{2x})^{-1}$ shows that the Pöschel-Teller Schrodinger equation

$$\left(-\frac{d^2}{dx^2} - l(l+1)\mathrm{sech}^2 x\right)\psi = E\psi \tag{10.58}$$

has solution

$$\psi(x) = (1+e^{2x})^{-\kappa/2}(1+e^{-2x})^{-\kappa/2}F\left(\kappa+l+1, \kappa-l; \kappa+1; \frac{1}{1+e^{2x}}\right), \tag{10.59}$$

where $E = -\kappa^2$. This solution behaves near $x = \infty$ as

$$\psi \sim e^{-\kappa x} F(\kappa + l + 1, \kappa - l; \kappa +; 0) = e^{-\kappa x}. \tag{10.60}$$

We use the connection formula (10.54) to see that it behaves in the vicinity of $x = -\infty$ as

$$\begin{aligned} \psi &\sim e^{\kappa x} F(\kappa + l + 1, \kappa - l; \kappa + 1; 1 - e^{2x}) \\ &\to e^{\kappa x} \frac{\Gamma(\kappa + 1)\Gamma(-\kappa)}{\Gamma(-l)\Gamma(1 + l)} + e^{-\kappa x} \frac{\Gamma(\kappa + 1)\Gamma(\kappa)}{\Gamma(\kappa + l + 1)\Gamma(\kappa - l)}. \end{aligned} \tag{10.61}$$

To find the bound-state spectrum, assume that κ is positive. Then $E = -\kappa^2$ will be an eigenvalue provided that coefficient of $e^{-\kappa x}$ near $x = -\infty$ vanishes. In other words, when

$$\frac{\Gamma(\kappa + 1)\Gamma(\kappa)}{\Gamma(\kappa + l + 1)\Gamma(\kappa - l)} = 0. \tag{10.62}$$

This condition is satisfied for a finite set κ_n, $n = 1, \ldots, [l]$ (where $[l]$ denotes the integer part of l) at which κ is positive but $\kappa - l$ is zero or a negative integer.

On setting $\kappa = -ik$, we find the scattering solution

$$\psi(x) = \begin{cases} e^{ikx} + r(k)e^{-ikx} & x \ll 0, \\ t(k)e^{ikx} & x \gg 0, \end{cases} \tag{10.63}$$

where

$$\begin{aligned} r(k) &= \frac{\Gamma(l + 1 - ik)\Gamma(-ik - l)\Gamma(ik)}{\Gamma(-l)\Gamma(1 + l)\Gamma(ik)}, \\ &= -\frac{\sin \pi l}{\pi} \frac{\Gamma(l + 1 - ik)\Gamma(-ik - l)\Gamma(ik)}{\Gamma(-ik)}, \end{aligned} \tag{10.64}$$

and

$$t(k) = \frac{\Gamma(l + 1 - ik)\Gamma(-ik - l)}{\Gamma(1 - ik)\Gamma(-ik)}. \tag{10.65}$$

Whenever l is a (positive) integer, the divergent factor of $\Gamma(-l)$ in the denominator of $r(k)$ causes the the reflected wave to vanish. This is something we had discovered in earlier chapters. In this particular case the transmission coefficient $t(k)$ reduces to a phase

$$t(k) = \frac{(-ik + 1)(-ik + 2) \cdots (-ik + l)}{(-ik - 1)(-ik - 2) \cdots (-ik - l)}. \tag{10.66}$$

10.3 Solving ODE's via Contour integrals

Our task in this section is to understand the origin of contour integral solutions such as the expression

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 (1-tz)^{-a} t^{b-1}(1-t)^{c-b-1} dt, \tag{10.67}$$

we have previously seen for the hypergeometric equation.

We are given a differential operator

$$L_z = \partial_{zz}^2 + p(z)\partial_z + q(z) \tag{10.68}$$

and seek a solution of $L_z u = 0$ as an integral

$$u(z) = \int_\Gamma F(z,t)\, dt. \tag{10.69}$$

If we can find an F such that

$$L_z F = \frac{\partial Q}{\partial t}, \tag{10.70}$$

for some function $Q(z,t)$ then

$$L_z u = \int_\Gamma L_z F(z,t)\, dt = \int_\Gamma \left(\frac{\partial Q}{\partial t}\right) dt = [Q]_\Gamma . \tag{10.71}$$

Thus, if Q vanishes at both ends of the contour, if it takes the same value at the two ends, or if the contour is closed and has no ends, we have succeeded in our quest.

Example: Consider Legendre's equation

$$L_z u \equiv (1-z^2)\frac{d^2u}{dz^2} - 2z\frac{du}{dz} + \nu(\nu+1)u = 0. \tag{10.72}$$

The identity

$$L_z\left\{\frac{(t^2-1)^\nu}{(t-z)^{\nu+1}}\right\} = (\nu+1)\frac{d}{dt}\left\{\frac{(t^2-1)^{\nu+1}}{(t-z)^{\nu+2}}\right\} \tag{10.73}$$

shows that

$$P_\nu(z) = \frac{1}{2\pi i}\int_\Gamma \left\{\frac{(t^2-1)^\nu}{2^\nu(t-z)^{\nu+1}}\right\} dt \tag{10.74}$$

will be a solution of Legendre's equation provided that

$$[Q]_\Gamma \equiv \left[\frac{(t^2-1)^{\nu+1}}{(t-z)^{\nu+2}}\right]_\Gamma = 0. \tag{10.75}$$

We could, for example, take a contour that circles the points $t = z$ and $t = 1$, but excludes the point $t = -1$. On going round this contour, the numerator aquires a phase of $e^{2\pi i(\nu+1)}$, while the denominator of $[Q]_\Gamma$ aquires a phase of $e^{2\pi i(\nu+2)}$. The net phase change is therefore $e^{-2\pi i} = 1$. The function in the integrated-out part is therefore single-valued, and so the integrated-out part vanishes. When ν is an integer, Cauchy's formula shows that

$$P_n(z) = \frac{1}{2^n n!}\frac{d^n}{dz^n}(z^2-1)^n, \tag{10.76}$$

which is Rodriguez' formula for the Legendre polynomials.

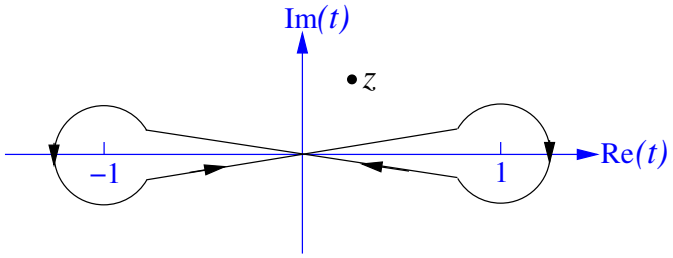

Figure 10.2: *Figure-of-eight contour for* $Q_\nu(Z)$.

The figure-of-eight contour shown in figure 10.2 gives us another solution

$$Q_\nu(z) = \frac{1}{4i\sin\pi\nu}\int_\Gamma \left\{\frac{(t^2-1)^\nu}{2^\nu(z-t)^{\nu+1}}\right\} dt, \quad \nu \notin \mathbb{Z}. \tag{10.77}$$

Here we define $\arg(t-1)$ and $\arg(t-1)$ to be zero for $t > 1$. The integrated out part vanishes because the phase gained by the $(t^2-1)^{\nu+1}$ in the numerator of $[Q]_\Gamma$ during the clockwise winding about $t = 1$ is undone during the anti-clockwise winding about $t = -1$, and, provided that z is outside the contour, there is no phase change in the $(z-t)^{-(\nu+2)}$ in the denominator.

When ν is real and positive the contributions from the circular arcs surrounding $t = \pm 1$ become negligeable as we shrink this new contour down onto the real axis. After this manouvre the integral (10.77) becomes

$$Q_\nu(z) = \frac{1}{2}\int_{-1}^{1} \left\{\frac{(1-t^2)^\nu}{2^\nu(z-t)^{\nu+1}}\right\} dt, \quad \nu > 0. \tag{10.78}$$

In contrast to (10.77), this last formula continues to make sense when ν is a positive integer, and so provides a convenient definition of $Q_n(z)$, the Legendre function of the second kind (See exercise 9.3).

It is hard to find a suitable $F(z,t)$ in one fell swoop. (The identity (10.73) exploited in the example is not exactly obvious!) An easier strategy is to seek solution in the form of an integral operator with kernel K acting on function $v(t)$. Thus we set

$$u(z) = \int_a^b K(z,t)v(t)\,dt. \tag{10.79}$$

Suppose that $L_z\,K(z,t) = M_t\,K(z,t)$, where M_t is differential operator in t that does not involve z. The operator M_t will have have a formal adjoint $M_t^\dagger$ such that

$$\int_a^b v(M_tK)\,dt - \int_a^b K(M_t^\dagger v)\,dt = \left[Q(K,v)\right]_a^b. \tag{10.80}$$

(This is Lagrange's identity.) Now

$$\begin{aligned} L_z u &= \int_a^b L_z K(z,t)v\,dt \\ &= \int_a^b (M_t K(z,t))v\,dt \\ &= \int_a^b K(z,t)(M_t^\dagger v)\,dt + \left[Q(K,v)\right]_a^b. \end{aligned}$$

We can therefore solve the original equation, $L_z u = 0$, by finding a v such that $(M_t^\dagger v) = 0$, and a contour with endpoints such that $\left[Q(K,v)\right]_a^b = 0$. This may sound complicated, but an artful choice of K can make it much simpler than solving the original problem.

Example: We will solve

$$L_z u = \frac{d^2u}{dz^2} - z\frac{du}{dz} + \nu u = 0, \tag{10.81}$$

by using the kernel $K(z,t) = e^{-zt}$. We have $L_z K(z,t) = M_t K(z,t)$ where

$$M_t = t^2 - t\frac{\partial}{\partial t} + \nu, \tag{10.82}$$

so

$$M_t^\dagger = t^2 + \frac{\partial}{\partial t}t + \nu = t^2 + (\nu+1) + t\frac{\partial}{\partial t}. \tag{10.83}$$

The equation $M_t^\dagger v = 0$ has solution

$$v(t) = t^{-(\nu+1)} e^{-\frac{1}{2}t^2}, \tag{10.84}$$

and so

$$u = \int_\Gamma t^{-(1+\nu)} e^{-(zt+\frac{1}{2}t^2)}\, dt, \tag{10.85}$$

for some suitable Γ.

10.3.1 Bessel Functions

As an illustration of the general method we will explore the theory of Bessel functions. Bessel functions are member of the family of *confluent hypergeometric functions*, obtained by letting the two regular singular points z_2, z_3 of the Riemann-Papperitz equation coalesce at infinity. The resulting singular point is no longer regular, and confluent hypergeometric functions have an essential singularity at infinity. The confluent hypergeometric equation is

$$zy'' + (c - z)y' - ay = 0, \tag{10.86}$$

with solution

$$\Phi(a, c; z) = \frac{\Gamma(c)}{\Gamma(a)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)}{\Gamma(c+n)\Gamma(n+1)} z^n. \tag{10.87}$$

The second solution, when c is not an integer, is

$$z^{1-c}\Phi(a - c + 1, 2 - c; z). \tag{10.88}$$

We see that

$$\Phi(a, c; z) = \lim_{b\to\infty} F(a, b; c; z/b). \tag{10.89}$$

Other functions of this family are the *parabolic cylinder functions*, which in special cases reduce to $e^{-z^2/4}$ times the *Hermite polynomials*, the *error function*

$$\operatorname{erf}(z) = \int_0^z e^{-t^2}\, dt = z\Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right) \tag{10.90}$$

and the *Laguerre polynomials*

$$L_n^m = \frac{\Gamma(n + m + 1)}{\Gamma(n + 1)\Gamma(m + 1)} \Phi(-n, m + 1; z). \tag{10.91}$$

Bessel's equation involves

$$L_z = \partial_{zz}^2 + \frac{1}{z}\partial_z + \left(1 - \frac{\nu^2}{z^2}\right). \tag{10.92}$$

Experience shows that a useful kernel is

$$K(z,t) = \left(\frac{z}{2}\right)^\nu \exp\left(t - \frac{z^2}{4t}\right). \tag{10.93}$$

Then

$$L_z K(z,t) = \left(\partial_t - \frac{\nu+1}{t}\right) K(z,t) \tag{10.94}$$

so M is a first order operator, which is simpler to deal with than the original second order L_z. In this case

$$M^\dagger = \left(-\partial_t - \frac{\nu+1}{t}\right) \tag{10.95}$$

and we need a v such that

$$M^\dagger v = -\left(\partial_t + \frac{\nu+1}{t}\right) v = 0. \tag{10.96}$$

Clearly $v = t^{-\nu-1}$ will work. The integrated out part is

$$[Q(K,v)]_a^b = \left[t^{-\nu-1} \exp\left(t - \frac{z^2}{4t}\right)\right]_a^b. \tag{10.97}$$

We see that

$$J_\nu(z) = \frac{1}{2\pi i}\left(\frac{z}{2}\right)^\nu \int_C t^{-\nu-1} e^{\left(t - \frac{z^2}{4t}\right)}\, dt. \tag{10.98}$$

solves Bessel's equation provided we use a suitable contour.

We can take for C a contour starting at $-\infty - i\epsilon$ and ending at $-\infty + i\epsilon$, and surrounding the branch cut of $t^{-\nu-1}$, which we take as the negative t axis.

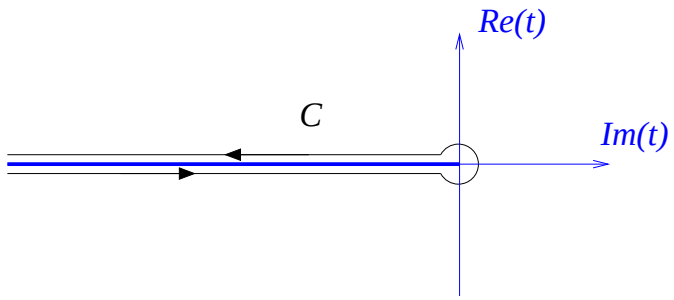

Figure 10.3: *Contour for solving Bessel equation.*

This contour works because Q is zero at both ends of the contour.

A cosmetic rewrite $t = uz/2$ gives

$$J_\nu(z) = \frac{1}{2\pi i} \int_C u^{-\nu-1} e^{\frac{z}{2}\left(u-\frac{1}{u}\right)}\, du. \tag{10.99}$$

For ν an integer, there is no discontinuity across the cut, so we can ignore it and take C to be the unit circle. Then, recognizing the resulting

$$J_n(z) = \frac{1}{2\pi i} \int_{|z|=1} u^{-n-1} e^{\frac{z}{2}\left(u-\frac{1}{u}\right)}\, du. \tag{10.100}$$

to be a Laurent coefficient, we obtain the familiar generating function

$$e^{\frac{z}{2}\left(u-\frac{1}{u}\right)} = \sum_{-\infty}^{\infty} J_n(z) u^n. \tag{10.101}$$

When ν is not an integer, we see why we need a branch cut integral.

If we set $u = e^w$ we get

$$J_\nu(z) = \frac{1}{2\pi i} \int_{C'} dw\, e^{z\sinh w - \nu w}, \tag{10.102}$$

where C' starts goes from $\infty - i\pi$ to $-i\pi$, to $+i\pi$ to $\infty + i\pi$.

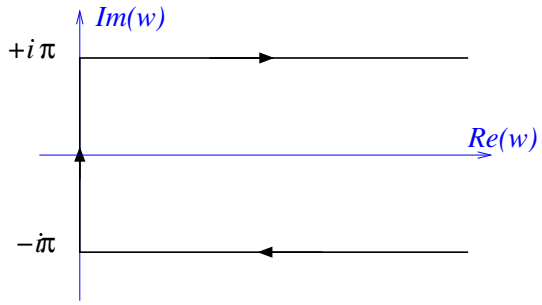

Figure 10.4: *Bessel contour after change of variables.*

If we set $w = t \pm i\pi$ on the horizontals and $w = i\theta$ on the vertical part, we can rewrite this as

$$J_\nu(z) = \frac{1}{\pi}\int_0^\pi \cos(\nu\theta - z\sin\theta)\,d\theta - \frac{\sin\nu\pi}{\pi}\int_0^\infty e^{-\nu t - z\sinh t}\,dt. \qquad (10.103)$$

All these are standard formulae for the Bessel function whose origin would be hard to understand without the contour solutions trick.

When ν becomes an integer, the functions $J_\nu(z)$ and $J_{-\nu}(z)$ are no longer independent. In order to have a Bessel equation solution that retains its independence from $J_\nu(z)$, even as ν becomes a whole number, we define the Neumann function

$$\begin{aligned} N_\nu(z) &\stackrel{\text{def}}{=} \frac{J_\nu(z)\cos\nu\pi - J_{-\nu}(z)}{\sin\nu\pi} \\ &= \frac{\cot\nu\pi}{\pi}\int_0^\pi \cos(\nu\theta - z\sin\theta)\,d\theta - \operatorname{cosec}\nu\pi\,\pi\int_0^\pi \cos(\nu\theta + z\sin\theta)\,d\theta \\ &\quad -\frac{\cos\nu\pi}{\pi}\int_0^\infty e^{-\nu t - z\sinh t}\,dt - \frac{1}{\pi}\int_0^\infty e^{\nu t - z\sinh t}\,dt. \end{aligned} \qquad (10.104)$$

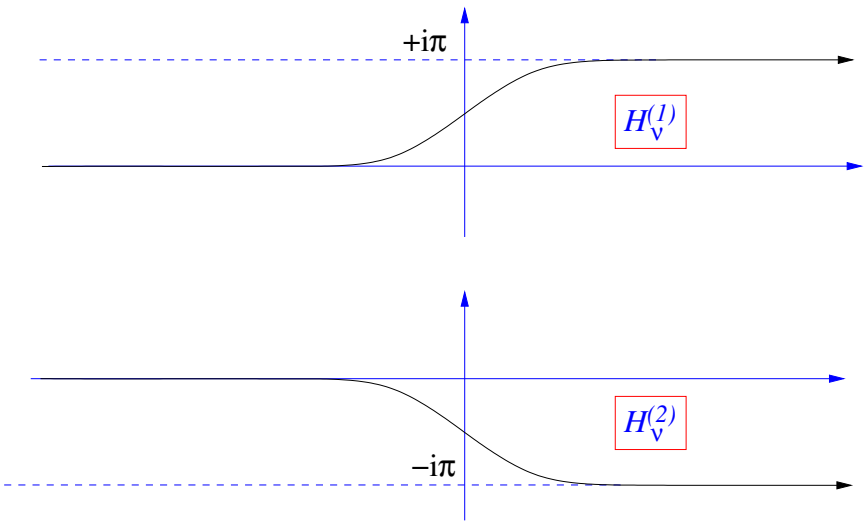

Figure 10.5: *Contours defining* $H_\nu^{(1)}(z)$ *and* $H_\nu^{(2)}(z)$.

Both Bessel and Neumann functions are real for positive real x. As x becomes large they oscillate as slowly decaying sines and cosines. It is sometimes convenient to decompose these real functions into solutions that behave as $e^{\pm ix}$. We therefore define the *Hankel functions* by

$$\begin{aligned} H_\nu^{(1)}(z) &= \frac{1}{i\pi}\int_{-\infty}^{\infty+i\pi} e^{z\sinh w-\nu w}\,dw, \qquad |\arg z|<\pi/2 \\ H_\nu^{(2)}(z) &= -\frac{1}{i\pi}\int_{-\infty}^{\infty-i\pi} e^{z\sinh w-\nu w}\,dw, \qquad |\arg z|<\pi/2. \end{aligned} \tag{10.105}$$

Then

$$\begin{aligned} \frac{1}{2}(H_\nu^{(1)}(z)+H_\nu^{(2)}(z)) &= J_\nu(z), \\ \frac{1}{2}(H_\nu^{(1)}(z)-H_\nu^{(2)}(z)) &= N_\nu(z). \end{aligned} \tag{10.106}$$

10.4 Asymptotic Expansions

We often need the understand the behaviour of solutions of differential equations and functions, such as $J_\nu(x)$, when x takes values that are very large, or very small. This is the subject of *asymptotics*.

As an introduction to this art, consider the function

$$Z(\lambda) = \int_{-\infty}^{\infty} e^{-x^2-\lambda x^4}\, dx. \tag{10.107}$$

Those of you who have taken a course quantum field theory based on path integrals will recognize that this is a "toy," 0-dimensional, version of the path integral for the $\lambda\varphi^4$ model of a self-interacting scalar field. Suppose we wish to obtain the perturbation expansion for $Z(\lambda)$ as a power series in λ. We naturally proceed as follows

$$\begin{aligned} Z(\lambda) &= \int_{-\infty}^{\infty} e^{-x^2-\lambda x^4}\, dx \\ &= \int_{-\infty}^{\infty} e^{-x^2} \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n x^{4n}}{n!}\, dx \\ &\stackrel{?}{=} \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!} \int_{-\infty}^{\infty} e^{-x^2} x^{4n}\, dx \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!}\, \Gamma(2n+1/2). \end{aligned} \tag{10.108}$$

Something has clearly gone wrong here! The gamma function $\Gamma(2n+1/2) \sim (2n)! \sim 4^n (n!)^2$ overwhelms the $n!$ in the denominator and the radius of convergence of the final power series is zero.

The invalid, but popular, manoeuvre is the interchange of the order of performing the integral and the sum. This interchange cannot be justified because the sum inside the integral does not converge uniformly on the domain of integration. Does this mean that the series is useless? It had better not! All quantum field theory (and most quantum mechanics) perturbation theory relies on versions of this manoeuvre.

We are saved to some (often adequate) degree because, while the interchange of integral and sum does not lead to a convergent series, it does lead to a valid *asymptotic expansion*. We write

$$Z(\lambda) \sim \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!} \Gamma(2n+1/2) \tag{10.109}$$

where

$$Z(\lambda) \sim \sum_{n=0}^{\infty} a_n \lambda^n \tag{10.110}$$

is shorthand for the more explicit

$$Z(\lambda) = \sum_{n=0}^{N} a_n \lambda^n + O\left(\lambda^{N+1}\right), \quad N = 1, 2, 3, \ldots. \tag{10.111}$$

The "big O" notation

$$Z(\lambda) - \sum_{n=0}^{N} a_n \lambda^n = O(\lambda^{N+1}) \tag{10.112}$$

as $\lambda \to 0$, means that

$$\lim_{\lambda \to 0} \left\{ \frac{|Z(\lambda) - \sum_0^N a_n \lambda^n|}{|\lambda^{N+1}|} \right\} = K < \infty. \tag{10.113}$$

The basic idea is that, given a convergent power series $\sum_n a_n \lambda^n$ for the function $f(\lambda)$, we fix the value of λ and take more and more terms. The sum then gets closer to $f(\lambda)$. Given an asymptotic expansion, on the other hand, we select a *fixed number of terms* in the series and then make λ smaller and smaller. The graph of $f(\lambda)$ and the graph of our polynomial approximation then approach each other. The more terms we take the sooner they get close, but for any non-zero λ we can never get exacty $f(\lambda)$—no matter how many terms we take.

We often consider asymptotic expansions where the independent variable becomes *large*. Here we have expansions in inverse powers of x:

$$F(x) = \sum_{n=0}^{N} b_n x^{-n} + O\left(x^{-N-1}\right), \quad N = 1, 2, 3 \ldots. \tag{10.114}$$

In this case

$$F(x) - \sum_{n=0}^{N} b_n x^{-n} = O\left(x^{-N-1}\right) \tag{10.115}$$

means that

$$\lim_{x \to \infty} \left\{ \frac{|F(x) - \sum_0^N b_n x^{-n}|}{|x^{-N-1}|} \right\} = K < \infty. \tag{10.116}$$

Again we take a fixed number of terms, and as x becomes large the function and its approximation get closer.

Observations:

i) Knowledge of the asymptotic expansion gives us useful knowledge about the function, but does not give us everything. In particular, two distinct functions may have the *same* asymptotic expansion. For example, for small positive λ, the functions $F(\lambda)$ and $F(\lambda)+ae^{-b/\lambda}$ have exactly the same asymptotic expansions as series in positive powers of λ. This is because $e^{-b/\lambda}$ goes to zero faster than any power of λ, and so its asymptotic expansion $\sum_n a_n\lambda^n$ has every coefficient a_n being zero. Physicists commonly say that $e^{-b/\lambda}$ is a *non-perturbative* function, meaning that it will not be visible to a perturbation expansion in powers of λ.

ii) An asymptotic expansion is usually valid only in a sector $a < \arg z < b$. Different sectors have different expansions. This is called the *Stokes' phenomenon*.

The most useful methods for obtaining asymptotic expansions require that the function to be expanded be given in terms of an integral. This is the reason why we have stressed the contour integral method of solving differential equations. If the integral can be approximated by a Gaussian, we are lead to the *method of steepest descents*. This technique is best explained by means of examples.

10.4.1 Stirling's Approximation for $n!$

We start from the integral representation of the Gamma function

$$\Gamma(z+1) = \int_0^\infty e^{-t}t^z\,dt \tag{10.117}$$

Set $t = z\zeta$, so

$$\Gamma(z+1) = z^{z+1}\int_0^\infty e^{zf(\zeta)}\,d\zeta, \tag{10.118}$$

where

$$f(\zeta) = \ln\zeta - \zeta. \tag{10.119}$$

We are going to be interested in evaluating this integral in the limit that $|z| \to \infty$ and finding the first term in the asymptotic expansion of $\Gamma(z+1)$ in powers of $1/z$. In this limit, the exponential will be dominated by the part of the integration region near the absolute maximum of $f(\zeta)$ Now $f(\zeta)$ is a maximum at $\zeta = 1$ and

$$f(\zeta) = -1 - \frac{1}{2}(\zeta-1)^2 + \cdots. \tag{10.120}$$

So

$$\begin{aligned}\Gamma(z+1) &= z^{z+1}e^{-z}\int_0^\infty e^{-\frac{z}{2}(\zeta-1)^2+\cdots}\,d\zeta \\ &\approx z^{z+1}e^{-z}\int_{-\infty}^\infty e^{-\frac{z}{2}(\zeta-1)^2}\,d\zeta \\ &= z^{z+1}e^{-z}\sqrt{\frac{2\pi}{z}} \\ &= \sqrt{2\pi}z^{z+1/2}e^{-z}. \end{aligned} \tag{10.121}$$

By keeping more of the terms represented by the dots, and expanding them as

$$e^{-\frac{z}{2}(\zeta-1)^2+\cdots} = e^{-\frac{z}{2}(\zeta-1)^2}\left[1+a_1(\zeta-1)+a_2(\zeta-1)^2+\cdots\right], \tag{10.122}$$

we would find, on doing the integral, that

$$\Gamma(z+1) \approx \sqrt{2\pi}z^{z+1/2}e^{-z}\left[1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{24888320z^4}+O\left(\frac{1}{z^5}\right)\right]. \tag{10.123}$$

Since $\Gamma(n+1)=n!$ we also have

$$n! \approx \sqrt{2\pi}n^{n+1/2}e^{-n}\left[1+\frac{1}{12n}+\cdots\right]. \tag{10.124}$$

We make contact with our discusion of asymptotic series by rewriting the expansion as

$$\frac{\Gamma(z+1)}{\sqrt{2\pi}z^{z+1/2}e^{-z}} \sim 1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{24888320z^4}+\ldots \tag{10.125}$$

This typical. We usually have to pull out a leading factor from the function whose asymptotic behaviour we are studying, before we are left with a plain asymptotic power series.

10.4.2 Airy Functions

The Airy functions $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$ are closely related to Bessel functions, and are named after the mathematician and astronomer George Biddell Airy. They occur widely in physics. We will investigate the behaviour of $\mathrm{Ai}(x)$ for

large values of $|x|$. A more sophisticated treatment is needed for this problem, and we will meet with Stokes' phenomenon. Airy's differential equation is

$$\frac{d^2y}{dz^2} - zy = 0. \tag{10.126}$$

On the real axis Airy's equation becomes

$$-\frac{d^2y}{dx^2} + xy = 0, \tag{10.127}$$

and we we can think of this as the Schrodinger equation for a particle running up a linear potential. A classical particle incident from the left with total energy $E = 0$ will come to rest at $x = 0$, and then retrace its path. The point $x = 0$ is therefore called a *classical turning point*.The corresponding quantum wavefunction, $\mathrm{Ai}\,(x)$, contains a travelling wave incident from the left and becoming evanescent as it tunnels into the classically forbidden region, $x > 0$, together with a reflected wave returning to $-\infty$. The sum of the incident and reflected waves is a real-valued standing wave.

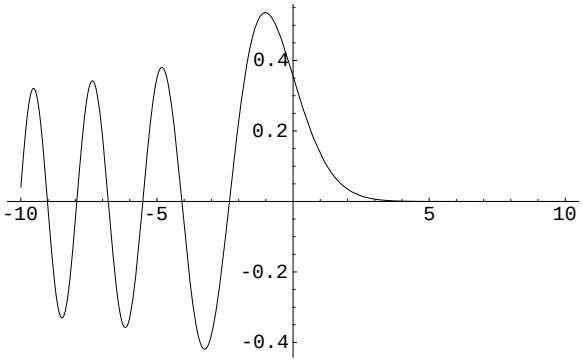

Figure 10.6: *The Airy function,* $\mathrm{Ai}\,(x)$.

We will look for contour integral solutions to Airy's equation of the form

$$y(x) = \int_C e^{xt} f(t)\, dt. \tag{10.128}$$

Denoting the Airy differential operator by $L_x \equiv \partial_x^2 - x$, we have

$$\begin{aligned} L_x\, y &= \int_C (t^2 - x) e^{xt} f(t)\, dt = \int_a^b f(t) \left\{ t^2 - \frac{d}{dt} \right\} e^{xt}\, dt. \\ &= \left[-e^{xt} f(t) \right]_C + \int_C \left(\left\{ t^2 + \frac{d}{dt} \right\} f(t) \right) e^{xt}\, dt. \end{aligned} \tag{10.129}$$

Thus $f(t) = e^{-\frac{1}{3}t^3}$ and

$$y(x) = \int_a^b e^{xt-\frac{1}{3}t^3}\, dt. \tag{10.130}$$

The contour must end at points where the integrated-out term, $\left[e^{xt-\frac{1}{3}t^3}\right]_C$, vanishes. There are therefore three possible contours, which end at any two of

$$+\infty, \quad \infty\, e^{2\pi i/3}, \quad \infty\, e^{-2\pi i/3}.$$

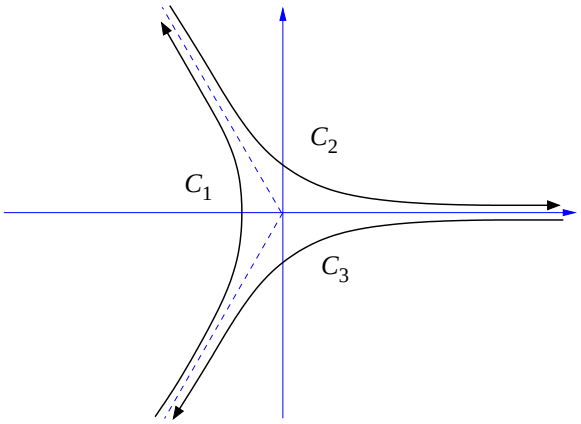

Figure 10.7: *Contours providing solutions of Airy's equation.*

Since the integrand is an entire function, the sum $y_{C_1} + y_{C_2} + y_{C_3}$ is zero, so only two of the three solutions are linearly independent. The Airy function itself is defined by

$$\mathrm{Ai}\,(z) = \frac{1}{2\pi i}\int_{C_1} e^{xt-\frac{1}{3}t^3}\, dt = \frac{1}{\pi}\int_0^\infty \cos\left(xs + \frac{1}{3}s^3\right)\, ds \tag{10.131}$$

In obtaining last equality, we have deformed the contour of integration, C_1, that ran from $\infty\, e^{-2\pi i/3}$ to $\infty\, e^{2\pi i/3}$ so that it lies on the imaginary axis, and there we have written $t = is$. You may check (*à la* Jordan) that this deformation does not alter the value of the integral.

To study the asymptotics of this function we need to examine separately two cases $x \gg 0$ and $x \ll 0$. For both ranges of x, the principal contribution to the integral will come from the neighbourhood of the stationary points of $f(t) = xt - t^3/3$. These stationary points are never pure maxima or

minima of the real part of f (the real part alone determines the magnitude of the integrand) but are always *saddle points*. We must deform the contour so that on the integration path the stationary point is the highest point in a mountain pass. We must also ensure that everywhere on the contour the difference between f and its maximum value stays *real*. Because of the orthogonality of the real and imaginary part contours, this means that we must take a path of *steepest descent* from the pass — hence the name of the method. If we stray from the steepest descent path, the phase of the exponent will be changing. This means that the integrand will oscillate and we can no longer be sure that the result is dominated by the contributions near the saddle point.

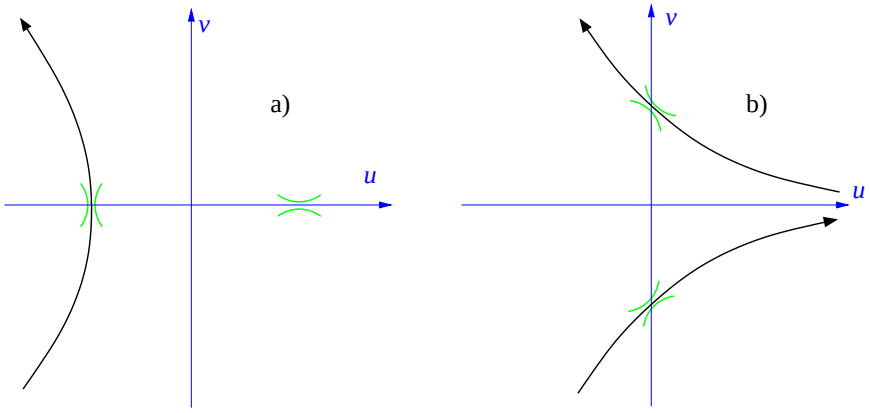

Figure 10.8: *Steepest descent contours and location and orientation of the saddle passes for a) $x \gg 0$, b) $x \ll 0$.*

i) $x \gg 0$: The stationary points are at $t = \pm\sqrt{x}$. Writing $t = \xi - \sqrt{x}$ have

$$f(\xi) = -\frac{2}{3}x^{3/2} + \xi^2\sqrt{x} - \frac{1}{3}\xi^3 \tag{10.132}$$

while near $t = +\sqrt{x}$ we write $t = \zeta + \sqrt{x}$ and find

$$f(\zeta) = -\frac{2}{3}x^{3/2} - \zeta^2\sqrt{x} - \frac{1}{3}\zeta^3 \tag{10.133}$$

We see that the saddle point near $-\sqrt{x}$ is a local maximum when we route the contour vertically, while the saddle point near $+\sqrt{x}$ is a local maximum as we go down the real axis. Since the contour in $\mathrm{Ai}\,(x)$ is

aimed vertically we can distort it to pass through the saddle point near $-\sqrt{x}$, but cannot find a route through the point at $+\sqrt{x}$ without the integrand oscillating wildly. At the saddle point the exponent, $xt-t^3/3$, is real. If we write $t = u + iv$ we have

$$\mathrm{Im}\,(xt - t^3/3) = v(x - u^2 + v^3/3), \tag{10.134}$$

so the exact steepest descent path, on which the imaginary part remains zero is given by the union of real axis ($v = 0$) and the curve

$$u^2 - \frac{1}{3}v^2 = x. \tag{10.135}$$

This is a hyperbola, and the branch passing through the saddle point at $-\sqrt{x}$ is plotted in a).
Now setting $\xi = is$, we find

$$\mathrm{Ai}\,(x) = \frac{1}{2\pi}e^{-\frac{2}{3}x^{3/2}}\int_{-\infty}^{\infty} e^{-\sqrt{x}s^2+\cdots}\,ds \sim \frac{1}{2\sqrt{\pi}}x^{-1/4}e^{-\frac{2}{3}x^{3/2}}. \tag{10.136}$$

ii) $x \ll 0$: The stationary points are now at $\pm i\sqrt{|x|}$. Setting $t = \xi \pm i\sqrt{|x|}$ find that

$$f(x) = \mp i\frac{2}{3}|x|^{3/2} \mp i\xi^2\sqrt{|x|}. \tag{10.137}$$

The exponent is no longer real, but the imaginary part will be constant and the integrand non-oscillatory provided we deform the contour so that it becomes the disconnected pair of curves shown in b). The new contour passes through both saddle points and we must sum their contributions. Near $t = i\sqrt{|x|}$ we set $\xi = e^{3\pi i/4}s$ and get

$$\begin{aligned}\frac{1}{2\pi i}e^{3\pi i/4}e^{-i\frac{2}{3}|x|^{3/2}}\int_{-\infty}^{\infty} e^{-\sqrt{|x|}s^2}\,ds &= \frac{1}{2i\sqrt{\pi}}e^{3\pi i/4}|x|^{-1/4}e^{-i\frac{2}{3}|x|^{3/2}} \\ &= -\frac{1}{2i\sqrt{\pi}}e^{-i\pi/4}|x|^{-1/4}e^{-i\frac{2}{3}|x|^{3/2}}\end{aligned} \tag{10.138}$$

Near $t = -i\sqrt{|x|}$we set $\xi = e^{2\pi i/3}s$ and get

$$\frac{1}{2i\pi}e^{\pi i/4}e^{i\frac{2}{3}|x|^{3/2}}\int_{-\infty}^{\infty} e^{-\sqrt{|x|}s^2}\,ds = \frac{1}{2i\sqrt{\pi}}e^{\pi i/4}|x|^{-1/4}e^{i\frac{2}{3}|x|^{3/2}} \tag{10.139}$$

The sum of these two contributions is

$$\mathrm{Ai}\,(x) \sim \frac{1}{\sqrt{\pi}|x|^{1/4}} \sin\left(\frac{2}{3}|x|^{3/2} + \frac{\pi}{4}\right). \tag{10.140}$$

The fruit of our labours is therefore

$$\begin{aligned} \mathrm{Ai}\,(x) &\sim \frac{1}{2\sqrt{\pi}} x^{-1/4} e^{-\frac{2}{3}x^{3/2}} \left[1 + O\left(\frac{1}{x}\right)\right], \quad x > 0, \\ &\sim \frac{1}{\sqrt{\pi}|x|^{1/4}} \sin\left(\frac{2}{3}|x|^{3/2} + \frac{\pi}{4}\right) \left[1 + O\left(\frac{1}{x}\right)\right], \quad x < 0. \end{aligned} \tag{10.141}$$

Suppose that we allow x to become complex $x \to z = |z|e^{i\theta}$, with $-\pi < \theta < \pi$. Then figure 10.9 shows how the steepest contour evolves and leads the two quite different expansion for positive and negative x. We see that for $0 < \theta < 2\pi/3$ the steepest descent path continues to be routed through the single stationary point at $-\sqrt{|z|}e^{i\theta/2}$. Once θ reaches $2\pi/3$, though, it passes through both stationary points. The contribution to the integral from the newly aquired stationary point is, however, exponentially smaller as $|z| \to \infty$ than that of $t = -\sqrt{|z|}e^{i\theta/2}$. The new term is therefore said to be *subdominant*, and makes an insignificant contribution to the asymptotic behaviour of $\mathrm{Ai}\,(z)$. The two saddle points only make contributions of the same magnitude when θ reaches π. If we analytically continue beyond $\theta = \pi$, the new saddlepoint will now dominate over the old, and only its contribtion is significant at large $|z|$. The *Stokes line*, at which we must change the form of the asymptotic expansion is therefore at $\theta = \pi$.

If we try to systematically keep higher order terms we will find, for the oscillating $\mathrm{Ai}\,(-z)$, a double series

$$\begin{aligned} \mathrm{Ai}\,(-z) \sim \pi^{-1/2} z^{-1/4} \Bigg[&\sin(\rho + \pi/4) \sum_{n=0}^{\infty} (-1)^n c_{2n} \rho^{-2n} \\ &- \cos(\rho + \pi/4) \sum_{n=0}^{\infty} (-1)^n c_{2n+1} \rho^{-2n-1} \Bigg] \end{aligned} \tag{10.142}$$

where $\rho = 2z^{3/2}/3$. In this case, therefore we need to extract two leading coefficients before we have asymptotic power series.

The subject of asymptotics contains many subtleties, and the reader in search of a more detailed discussion is recommened to read Bender and Orszags *Advanced Mathematical methods for Scientists and Engineers*.

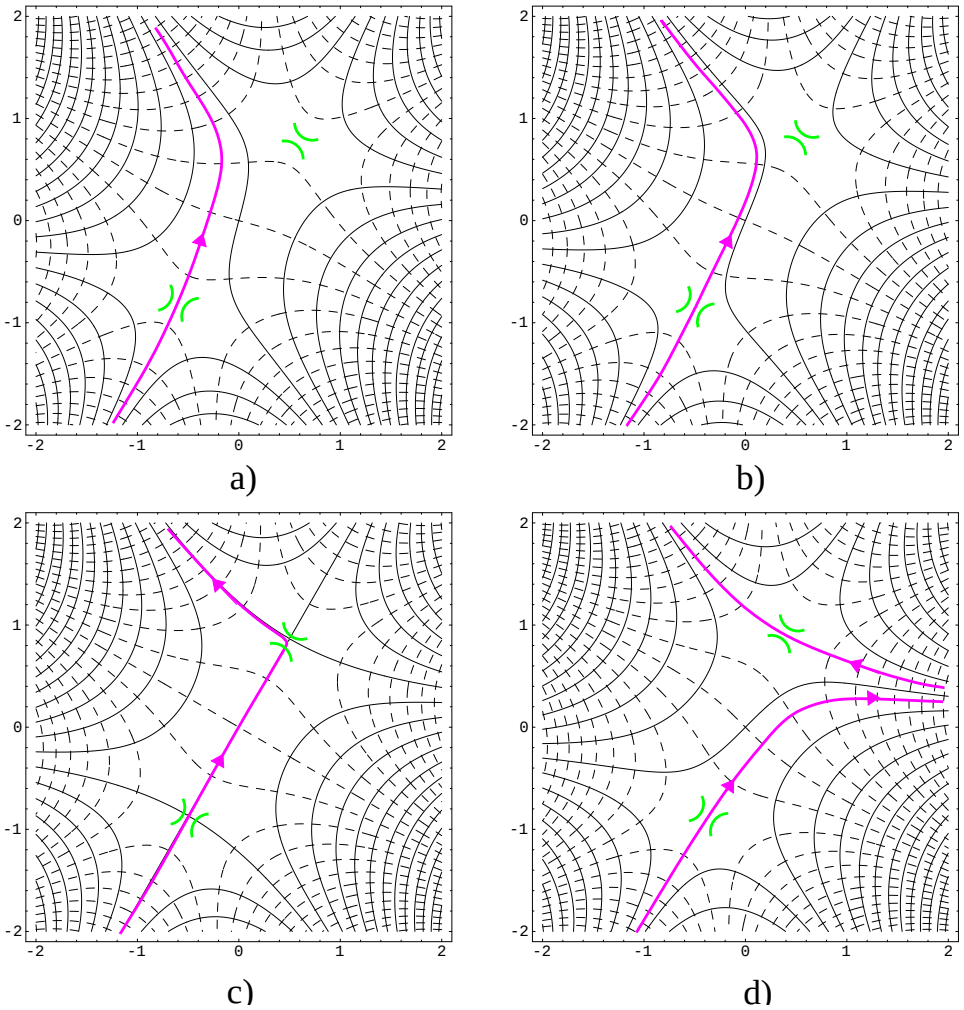

Figure 10.9: *Evolution of the steepest-descent contour from passing through only one saddle point to passing through both. The dashed and solid lines are contours of the real and imaginary parts, repectively, of* $(zt - t^3/3)$. $\theta = \mathrm{Arg}\, z$ *takes the values a)* $7\pi/12$, *b)* $15\pi/24$, *c)* $2\pi/3$, *d)* $9\pi/12$.

Exercise 10.2: Consider the behaviour of Bessel functions when x is large. By applying the method of steepest descent to the Hankel function contours show that

$$\begin{aligned} H_\nu^{(1)}(x) &\sim \sqrt{\frac{2}{\pi x}} e^{i(x-\nu\pi/2-\pi/4)} \left[1 - \frac{4\nu^2-1}{8\pi x} + \cdots\right] \\ H_\nu^{(2)}(x) &\sim \sqrt{\frac{2}{\pi x}} e^{-i(x-\nu\pi/2-\pi/4)} \left[1 + \frac{4\nu^2-1}{8\pi x} + \cdots\right], \end{aligned}$$

and hence

$$\begin{aligned} J_\nu(x) &\sim \sqrt{\frac{2}{\pi x}} \left[\cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) - \frac{4\nu^2-1}{8x} \sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + \cdots\right], \\ N_\nu(x) &\sim \sqrt{\frac{2}{\pi x}} \left[\sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + \frac{4\nu^2-1}{8x} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + \cdots\right]. \end{aligned}$$

10.5 Elliptic Functions

The subject of elliptic functions goes back to remarkable identities of Guilio Fagnano (1750) and Leonhard Euler (1761). Euler's formula is

$$\int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_0^v \frac{dy}{\sqrt{1-y^4}} = \int_0^r \frac{dz}{\sqrt{1-z^4}}, \tag{10.143}$$

where $0 \le u, v \le 1$, and

$$r = \frac{u\sqrt{1-v^4} + v\sqrt{1-u^4}}{1+u^2v^2}. \tag{10.144}$$

This looks mysterious, but perhaps so does

$$\int_0^u \frac{dx}{\sqrt{1-x^2}} + \int_0^v \frac{dy}{\sqrt{1-y^2}} = \int_0^r \frac{dz}{\sqrt{1-z^2}}, \tag{10.145}$$

where

$$r = u\sqrt{1-v^2} + v\sqrt{1-u^2}, \tag{10.146}$$

until you realize that the latter formula is merely

$$\sin(a+b) = \sin a \cos b + \cos a \sin b \tag{10.147}$$

in disguise. To see this set

$$u = \sin a, \quad v = \sin b \tag{10.148}$$

and remember the integral formula for the inverse trig function

$$a = \sin^{-1} u = \int_0^u \frac{dx}{\sqrt{1-x^2}}. \tag{10.149}$$

The Fagnano-Euler formula is a similarly disguised addition formula for an *elliptic function*. Just as we use the substitution $x = \sin y$ in the $1/\sqrt{1-x^2}$ integral, we can use an elliptic function substitution to evaluate *elliptic integrals* such as

$$I_4 = \int_0^x \frac{dt}{\sqrt{(t-a_1)(t-a_2)(t-a_3)(t-a_4)}} \tag{10.150}$$

$$I_3 = \int_0^x \frac{dt}{\sqrt{(t-a_1)(t-a_2)(t-a_3)}}. \tag{10.151}$$

The integral I_3 is a special case of I_4, where a_4 has been sent to infinity by use of a Möbius map

$$t \to t' = \frac{at+b}{ct+d}, \quad dt' = (ad-bc)\frac{dt}{(ct+d)^2}. \tag{10.152}$$

Indeed, we can use a suitable Möbius map to send any three of the four points a_n to $0, 1, \infty$.

The idea of elliptic functions (as opposed to the integrals, which are their functional inverse) was known to Gauss, but Abel and Jacobi were the first to publish (1827). For the general theory, the simplest elliptic function is the Weierstrass $\wp$. This is defined by first selecting two linearly independent *periods* ω_1, ω_2, and setting

$$\wp(z) = \frac{1}{z^2} + \sum_{(m,n)\neq 0} \left\{ \frac{1}{(z - m\omega_1 - n\omega_2)^2} - \frac{1}{(m\omega_1 + n\omega_2)^2} \right\}. \tag{10.153}$$

The sum is over integers m, n, positive and negative, but not both 0. Helped by the counterterm, the sum is absolutely convergent, so we can rearrange the terms to prove double periodicity

$$\wp(z + m\omega_1 + n\omega_2) = \wp(z). \tag{10.154}$$

The function is thus determined everywhere by its values in the period parallelogram $P = \{\lambda\omega_1 + \mu\omega_2 : 0 \leq \lambda, \mu < 1\}$. Double periodicity is the defining characteristic of elliptic functions.

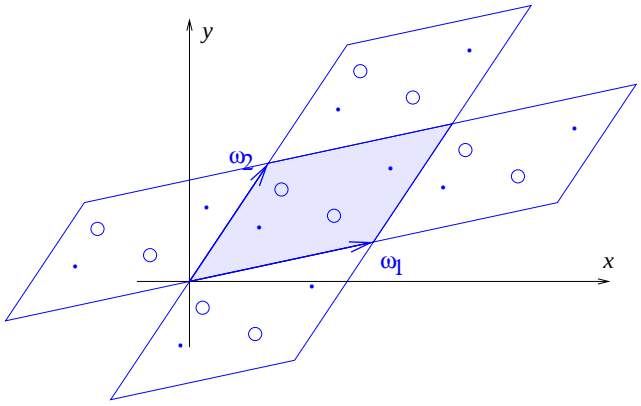

Figure 10.10: *Unit cell and double-periodicity.*

Any non-constant meromorphic function, $f(z)$, which is doubly periodic has four basic properties:

a) The function must have at least one pole in its unit cell. Otherwise it would be holomorphic and bounded, and therefore a constant by Liouville.

b) The sum of the residues at the poles must add to zero. This follows from integrating $f(z)$ around the boundary of the period parallelogram and observing that the contributions from opposite edges cancel.

c) The number of poles in each unit cell must equal the number of zeros. This follows from integrating f'/f round the boundary of the period parallelogram.

d) If f has zeros at the N points z_i and poles at the N points p_i then

$$\sum_{i=1}^{N} z_i - \sum_{i=1}^{N} p_i = n\omega_1 + m\omega_2$$

where m, n are integers. This follows from integrating zf'/f round the boundary of the period parallelogram.

The Weierstass $\wp$ has a second-order pole at the origin. It also obeys

$$\lim_{|z|\to 0} \left(\wp(z) - \frac{1}{z^2} \right) \quad = \quad 0,$$

$$\begin{aligned} \wp(z) &= \wp(-z), \\ \wp'(z) &= -\wp'(-z). \end{aligned} \tag{10.155}$$

The property that makes $\wp(z)$ useful for evaluating integrals is

$$(\wp'(z))^2 = 4\wp^3(z) - g_2\wp(z) - g_3, \tag{10.156}$$

where

$$g_2 = 60 \sum_{(m,n)\neq 0} \frac{1}{(m\omega_1 + n\omega_2)^4}, \quad g_3 = 140 \sum_{(m,n)\neq 0} \frac{1}{(m\omega_1 + n\omega_2)^6}. \tag{10.157}$$

Equation (10.156) is proved by examining the first few terms in the Laurent expansion in z of the difference of the left hand and right hand sides. All negative powers cancel, as does the constant term. The difference is zero at $z = 0$, has no poles or other singularities, and being continuous and periodic is automatically bounded. It is therefore identically zero by Liouville's theorem.

From the symmetry and periodicity of $\wp$ we see that $\wp'(z) = 0$ at $\omega_1/2$, $\omega_2/2$ and $(\omega_1 + \omega_2)/2$ where $\wp(z)$ takes values $e_1 = \wp(\omega_1/2)$, $e_2 = \wp(\omega_2/2)$, and $e_3 = \mathcal{P}((\omega_1+\omega_2)/2)$. Now $\wp'$ must have exactly three zeros since it has a pole of order three at the origin and, by property c), the number of zeros in the unit cell is equal to the number of poles. We therefore know the location of all three zeros and can factorize

$$4\wp^3(z) - g_2\wp(z) - g_3 = 4(\wp - e_1)(\wp - e_2)(\wp - e_3). \tag{10.158}$$

We note that the coefficient of $\wp^2$ in the polynomial on the left side is zero, implying that $e_1 + e_2 + e_3 = 0$. This is consistent with property d).

The roots e_i can never coincide. For example, $(\wp(z) - e_1)$ has a double zero at $\omega_1/2$, but two zeros is all it is allowed because the number of poles per unit cell equals the number of zeros, and $(\wp(z) - e_1)$ has a double pole at 0 as its only singularity. Thus $(\wp - e_1)$ cannot be zero at another point, but it would be if e_1 coincided with e_2 or e_3. As a consequence, the *discriminant*

$$\Delta = 16(e_1 - e_2)^2(e_2 - e_3)^2(e_1 - e_3)^2 = g_2^3 - 27g_3^2, \tag{10.159}$$

is never zero.

We use $\wp$ to write

$$z = \wp^{-1}(u) = \int_\infty^u \frac{dt}{2\sqrt{(t - e_1)(t - e_2)(t - e_3)}} = \int_\infty^u \frac{dt}{\sqrt{4t^3 - g_2t - g_3}}. \tag{10.160}$$

This maps the u plane cut from e_1 to e_2 and e_3 to ∞ one-to-one onto the 2-torus, regarded the unit cell of the $\omega_{n,m} = n\omega_1 + m\omega_2$ lattice.

As z sweeps over the torus, the points $x = \wp(z)$, $y = \wp'(z)$ move on the *elliptic curve*

$$y^2 = 4x^3 - g_2 x - g_3 \tag{10.161}$$

which should be thought of as a set in $\mathbb{C}P^2$. These curves, and the finite fields of rational points that lie on them, are exploited in modern cryptography.

The magic which leads to addition formula, such as the Euler-Fagnano relation with which we began this section, lies in the (not immediatley obvious) fact that any elliptic function having the same periods as $\wp(z)$ can be expressed as a rational function of $\wp(z)$ and $\wp'(z)$. From this it follows (after some thought) that any two such elliptic functions, $f_1(z)$ and $f_2(z)$, obey a relation $F(f_1, f_2) = 0$, where

$$F(x, y) = \sum a_{n,m} x^n y^m \tag{10.162}$$

is a polynomial in x and y. We can eliminate $\wp'(z)$ in these relations at the expense of introducing square roots.

modular invariance

If ω_1 and ω_2 are periods and define a unit cell, so are

$$\begin{aligned} \omega_1' &= a\omega_1 + b\omega_2 \\ \omega_2' &= c\omega_1 + d\omega_2 \end{aligned}$$

where a, b, c, d are integers with $ad - bc = \pm 1$. This condition on the determinant ensures that the matrix inverse also has integer entries, and so the ω_i can be expressed in terms of the ω_i' with integer coefficients. Consequently the set of integer linear combinations of the ω_i' generate the same lattice as the integer linear combinations of the original ω_i. This notion of redefining the unit cell should be familiar to your from solid state physics. If we wish to preserve the orientation of the basis vectors, we must restrict ourselves to maps whose determinant $ad - bc$ is unity. The set of such transforms constitute the the *modular* group $\mathrm{SL}(2, \mathbb{Z})$. Clearly $\wp$ is invariant under this group, as are g_2 and g_3 and Δ. Now define $\omega_2/\omega_1 = \tau$, and write

$$g_2(\omega_1, \omega_2) = \frac{1}{\omega_1^4}, \tilde{g}_2(\tau), \quad g_3(\omega_1, \omega_2) = \frac{1}{\omega_1^6}, \tilde{g}_3(\tau). \quad \Delta(\omega_1, \omega_2) = \frac{1}{\omega_1^{12}} \tilde{\Delta}(\tau), \tag{10.163}$$

and also

$$J(\tau) = \frac{\tilde{g}_2^3}{\tilde{g}_2^3 - 27\tilde{g}_3^2} = \frac{\tilde{g}_2^3}{\tilde{\Delta}}. \tag{10.164}$$

Because the denominator is never zero when $\mathrm{Im}\,\tau > 0$, the function $J(\tau)$ is holomorphic in the upper half-plane — but not on the real axis. The function $J(\tau)$ is called the *elliptic modular function*.

Except for the prefactors ω_1^n, the functions $\tilde{g}_i(\tau)$, $\tilde{\Delta}(\tau)$ and $J(\tau)$ are invariant under the Möbius transformation

$$\tau \to \frac{a\tau + b}{c\tau + d}. \tag{10.165}$$

with

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}). \tag{10.166}$$

This Möbius transformation does not change if the entries in the matrix are multiplied by a common factor of ± 1, and so the transformation is an element of the modular group $\mathrm{PSL}(2,\mathbb{Z}) \equiv \mathrm{SL}(2,\mathbb{Z})/\{I, -I\}$.

Taking into account the change in the ω_1^α prefactors we have

$$\begin{aligned} \tilde{g}_2\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^4 \tilde{g}_3(\tau), \\ \tilde{g}_3\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^6 \tilde{g}_3(\tau), \\ \tilde{\Delta}\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^{12} \tilde{\Delta}(\tau). \end{aligned} \tag{10.167}$$

Because $c = 0$ and $d = 1$ for the special case $\tau \to \tau + 1$, these three functions obey $f(\tau+1) - f(\tau)$ and so depend on τ only via the combination $q^2 = e^{2\pi i \tau}$. For example, it is not hard to prove that

$$\tilde{\Delta}(\tau) = (2\pi)^{12} q^2 \prod_{n=1}^{\infty} \left(1 - q^{2n}\right)^{24}. \tag{10.168}$$

We can also expand them as power series in q^2 — and here things get interesting because the coefficients have number-theoretic properties. For example

$$\begin{aligned} \tilde{g}_2(\tau) &= (2\pi)^4 \left[\frac{1}{12} + 20 \sum_{n=1}^{\infty} \sigma_3(n) q^{2n}\right], \\ \tilde{g}_3(\tau) &= (2\pi)^6 \left[\frac{1}{216} - \frac{7}{3} \sum_{n=1}^{\infty} \sigma_5(n) q^{2n}\right]. \end{aligned} \tag{10.169}$$

The symbol $\sigma_k(n)$ is defined by $\sigma_k(n) = \sum d^k$ where d runs over all positive divisors of the number n.

In the case of the function $J(\tau)$, the prefactors cancel and

$$J\left(\frac{a\tau + b}{c\tau + d}\right) = J(\tau), \tag{10.170}$$

so $J(\tau)$ is a *modular invariant*. One can show that if $J(\tau_1) = J(\tau_2)$, then

$$\tau_2 = \frac{a\tau_1 + b}{c\tau_1 + d} \tag{10.171}$$

for some modular transformation with integer a, b, c, d, where $ad - bc = 1$, and further, that any modular invariant function is a rational function of $J(\tau)$. It seems clear that $J(\tau)$ is rather a special object.

This $J(\tau)$ is the function referred to on page 174 in connection with the Monster group. As with the $\tilde{g}_i$, $J(\tau)$ depends on τ only through q^2. The first few terms in the power series expansion of $J(\tau)$ in terms of q^2 turn out to be

$$1728J(\tau) = q^{-2} + 744 + 196884q^2 + 21493760q^4 + 864299970q^6 + \cdots. \tag{10.172}$$

Since $AJ(\tau) + B$ has all the same modular invariance properties as $J(\tau)$, the numbers $1728 = 12^3$ and 744 are just conventional normalizations. Once we set the coefficient of q^{-2} to unity, however, the remaining integer coefficients are completely determined by the modular properties. A number-theory interpretation of these integers seemed lacking until John McKay and others observed that that

$$\begin{aligned} 1 &= 1 \\ 196884 &= 1 + 196883 \\ 21493760 &= 1 + 196883 + 21296786 \\ 864299970 &= 2 \times 1 + 2 \times 196883 + 21296786 + 842609326, \end{aligned} \tag{10.173}$$

where "1" and the large integers on the right-hand side are the dimensions of the smallest irreducible representations of the Monster. This "Monstrous Moonshine" was originally mysterious and almost unbelievable, ("moonshine" = "fantastic nonsense") but it was explained by Richard Borcherds by the use of techniques borrowed from string theory.[3] Borcherds received the 1998 Fields Medal for this work.

[3] "I was in Kashmir. I had been traveling around northern India, and there was one

10.6 Further Exercises and Problems

Exercise 10.3: Show that the binomial series expansion of $(1+x)^{-\nu}$ can be written as

$$(1+x)^{-\nu} = \sum_{m=0}^{\infty} (-x)^m \frac{\Gamma(m+\nu)}{\Gamma(\nu)\, m!}, \quad |x| < 1.$$

Exercise 10.4: *A Mellin transform and its inverse.* Combine the Beta-function identity (10.15) with a suitable change of variables to evaluate the Mellin transform

$$\int_0^{\infty} x^{s-1}(1+x)^{-\nu}\, dx, \quad \nu > 0,$$

of $(1+x)^{-\nu}$ as a product of Gamma functions. Now consider the integral

$$\frac{1}{2\pi i\, \Gamma(\nu)} \int_{c-i\infty}^{c+i\infty} x^{-s}\Gamma(\nu-s)\Gamma(s)\, ds.$$

Here $\operatorname{Re} c \in (0, \nu)$. The contour therefore runs parallel to the imaginary axis with the poles of $\Gamma(s)$ to its left and the poles of $\Gamma(\nu - s)$ to its right. Use the identity

$$\Gamma(s)\Gamma(1-s) = \pi \operatorname{cosec} \pi s$$

to show that when $|x| < 1$ the contour can be closed by a large semicircle lying to the left of the imaginary axis. By using the preceding exercise to sum the contributions from the enclosed poles at $s = -n$, evaluate the integral.

Exercise 10.5: *Mellin-Barnes integral.* Use the technique developed in the preceding exercise to show that

$$F(a,b,c;-x) = \frac{\Gamma(c)}{2\pi i\, \Gamma(a)\Gamma(b)} \int_{c-i\infty}^{c+i\infty} x^{-s} \frac{\Gamma(a-s)\Gamma(b-s)\Gamma(s)}{\Gamma(c-s)}\, ds,$$

for a suitable range of x. This integral representation of the hypergeometric function is due to the English mathematician Ernest Barnes (1908), later a controversial Bishop of Birmingham.

---

really long tiresome bus journey, which lasted about 24 hours. Then the bus had to stop because there was a landslide and we couldn't go any further. It was all pretty darn unpleasant. Anyway, I was just toying with some calculations on this bus journey and finally I found an idea which made everything work"- Richard Borcherds (Interview in *The Guardian*, August 1998).

Exercise 10.6: Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

Show that the matrix differential equation

$$\frac{d}{dx}Y = \frac{A}{z}Y + \frac{B}{1-z}Y,$$

where

$$A = \begin{pmatrix} 0 & a \\ 0 & 1-c \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ b & a+b-c+1 \end{pmatrix},$$

has a solution

$$Y(z) = F(a,b,;c,z)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{z}{a}F'(a,b;c;z)\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Exercise 10.7: *Kniznik-Zamolodchikov equation.* The monodromy properties of solutions of differential equations play an important role in conformal field theory. The Fuchsian equations studied in this exercise are obeyed by the correlation functions in the level-k Wess-Zumino-Witten model.

Let $V^{(a)}$, $a = 1, \ldots n$, be spin-j_a representation spaces for the group SU(2). Let $W(z_1, \ldots, z_n)$ be a function taking values in $V^{(1)} \otimes V^{(2)} \otimes \cdots \otimes V^{(n)}$. (In other words W is a function $W_{i_1,\ldots,i_n}(z_1, \ldots, z_n)$ where the index i_a labels states in the spin-j_a factor.) Suppose that W obeys the *Kniznik-Zamolodchikov (K-Z) equations*

$$(k+2)\frac{\partial}{\partial z_a}W = \sum_{b,b\neq a} \frac{\mathbf{J}^{(a)} \cdot \mathbf{J}^{(b)}}{z_a - z_b}W, \quad a = 1, \ldots, n,$$

where

$$\mathbf{J}^{(a)} \cdot \mathbf{J}^{(b)} \equiv J_1^{(a)}J_1^{(b)} + J_2^{(a)}J_2^{(b)} + J_3^{(a)}J_3^{(b)},$$

and $J_i^{(a)}$ indicates the $\mathfrak{su}(2)$ generator J_i acting on the $V^{(a)}$ factor in the tensor product. If we set $z_1 = z$, for example and fix the position of $z_2, \ldots z_n$, then the differential equation in z has regular singular points at the $n-1$ remaining z_b.

a) By diagonalizing the operator $\mathbf{J}^{(a)} \cdot \mathbf{J}^{(b)}$ show that there are solutions $W(z)$ that behave for z_a close to z_b as

$$W(z) \sim (z_a - z_b)^{\Delta_j - \Delta_{j_a} - \Delta_{j_b}},$$

where

$$\Delta_j = \frac{j(j+1)}{k+2}, \quad \Delta_{j_a} = \frac{j_a(j_a+1)}{k+2},$$

and j is one of the spins $|j_a - j_b| \leq j \leq j_1 + j_a$ occuring in the decomposition of $j_a \otimes j_b$.

b) Define covariant derivatives

$$\nabla_a = \frac{\partial}{\partial z_a} - \sum_{b, b \neq a} \frac{\mathbf{J}^{(a)} \cdot \mathbf{J}^{(b)}}{z_a - z_b}$$

and show that $[\nabla_a, \nabla_b] = 0$. Conclude that the effect of parallel transport of the solutions of the K-Z equations provides a representation of the braid group of the world lines of the z_a.

Index