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

Vectors and Tensors

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

1.1 Covariant and Contravariant Vectors

Suppose that we have a vector space V over $\mathbf{R}$, and that $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ and $\{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n\}$ are both bases for V. We can therefore expand each of the basis vectors $\mathbf{e}_i$ in terms of the $\mathbf{e}'_i$ as

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

(we are, as usual, using the Einstein summation convention that repeated indices are to be summed over.) Alternatively we could have expanded the $\mathbf{e}'_i$ in terms of the $\mathbf{e}_i$ as

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

The matrices of coefficients A^μ_ν and $(A^{-1})^\mu_\nu$ must be inverses of each other:

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

Now the components x'^μ of $\mathbf{x}$ in the new basis are found from

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

as $x'^\mu = A^\mu_\nu x^\nu$. Observe how the $\mathbf{e}_\mu$ and the x^μ map 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 $\mathbf{f} : V \to \mathbf{R}$. If $\mathbf{f} \in V^*$ and $\mathbf{x} = x^\mu \mathbf{e}_\mu$ we can use the linearity to evaluate $\mathbf{f}(\mathbf{x})$ as

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

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

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

Thus $f_\nu = A^\mu_\nu f'_\mu$. We see that the f_μ components transform in the same way 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. Physicists, however, like to give priority to the space in which we live and breathe. In typical physics applications, therefore, a displacement vector $\mathbf{x}$ will be a contravariant vector, and a Fourier-space wavenumber $\mathbf{k}$ will be a covariant vector. The "dot" in expressions such as

$$\psi(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}} \tag{1.4}$$

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

$$(\mathbf{k}, \mathbf{x}) \equiv \mathbf{k}(\mathbf{x}) = k_\mu x^\mu. \tag{1.5}$$

The physical units of $\mathbf{x}$ and $\mathbf{k}$ being different (meters *versus* meters^{-1}) should make it clear that they are not elements of the same vector space. There is no meaning to $\mathbf{x} + \mathbf{k}$.

Often our vector space will come equipped with a *metric*, which is derived from a non-degenerate inner product $\mathbf{g} : V \times V \to \mathbf{R}$. The length $\|\mathbf{x}\|$ of a vector $\mathbf{x}$ is is then given by $\sqrt{\mathbf{g}(\mathbf{x}, \mathbf{x})}$. The set of numbers

$$g_{\mu\nu} = \mathbf{g}(\mathbf{e}_\mu, \mathbf{e}_\nu) \tag{1.6}$$

are said to be the components of the *metric tensor*. The inner of product of any pair of vectors $\mathbf{x} = x^\mu \mathbf{e}_\mu$ and $\mathbf{y} = y^\mu \mathbf{e}_\mu$ is then

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) = g_{\mu\nu} x^\mu y^\nu. \tag{1.7}$$

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

The additional structure provided by the metric permits us to identify V with V^*. For any $\mathbf{f} \in V^*$ we can find a vector $\tilde{\mathbf{f}} \in V$ such that

$$\mathbf{f}(\mathbf{x}) = \mathbf{g}(\tilde{\mathbf{f}}, \mathbf{x}). \tag{1.8}$$

We simply solve the equation

$$f_\mu = g_{\mu\nu} \tilde{f}^\nu \tag{1.9}$$

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

$$f^\mu = g^{\mu\nu} f_\nu, \tag{1.10}$$

or *lowering*

$$f_\mu = g_{\mu\nu} f^\nu, \tag{1.11}$$

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

We sometimes play this game in $\mathbf{R}^n$ equipped with its Euclidean metric and associated "dot" inner product. Given a vector $\mathbf{x}$ and a non-orthogonal basis $\mathbf{e}_\mu$ with $g_{\mu\nu} = \mathbf{e}_\mu \cdot \mathbf{e}_\nu$, we can define two sets of components for the same vector. Firstly the coefficients x^μ appearing in the basis expansion

$$\mathbf{x} = x^\mu \mathbf{e}_\mu, \tag{1.12}$$

and secondly the "components"

$$x_\mu = \mathbf{x} \cdot \mathbf{e}_\mu = \mathbf{g}(\mathbf{x}, \mathbf{e}_\mu) = g_{\mu\nu} x^\nu, \tag{1.13}$$

of $\mathbf{x}$ along the basis vectors. These two set of numbers are then called the contravariant and covariant components, respectively, of the vector $\mathbf{x}$. If the $\mathbf{e}_\mu$ constitute an orthonormal basis, then $g_{\mu\nu} = \delta_{\mu\nu}$, and the two sets of components are numerically coincident. When using non-orthogonal bases we must never to add contravariant components to covariant ones, and we must always be careful with units.

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 tensor products of the spaces V and V^*.

1.2.1 Transformation Rules

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

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

These are be related to the old components as

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

Equivalently

$$g'_{\mu\nu} = (A^{-1})^\rho_\mu (A^{-1})^\sigma_\nu g_{\rho\sigma}.$$

Both indices transform as the downstairs indices of a covariant vector. We ntherefore say that $g_{\mu\nu}$ transforms as a *doubly covariant tensor*.

A set of numbers such as $Q^{ij}{}_{klm}$ with transformation rule

$$Q^{ij}{}_{klm} = (A^{-1})^i_{i'} (A^{-1})^j_{j'} A^{k'}_k \, A^{l'}_l \, A^{m'}_m Q'^{i'j'}{}_{k'l'm'}$$

or, equivalently

$$Q'^{ij}{}_{klm} = A^i_{i'} A^j_{j'} (A^{-1})^{k'}_k (A^{-1})^{l'}_l (A^{-1})^{m'}_m Q^{i'j'}{}_{k'l'm'}$$

are the components of a *doubly contravariant and triply covariant* tensor. More compactly, they are a the components of a tensor of type $(2, 3)$. Tensors of type (p, q) are defined analogously.

Notice how the indices are wired up: free (not summed over) upstairs indices on the left hand side of the equation match to free upstairs indices on the right hand side, similarly downstairs indices. Also upstairs indices are summed only with downstairs ones.

Similar conditions apply to equations relating tensors in any particular frame. 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

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

is fine, but

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

has something wrong in each term.

Incidentally, although not illegal, it is a good idea not to write 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.

Although often associated with general relativity, tensors occur in many places in physics. Perhaps the most obvious, and the source of the name "tensor", is elasticity theory. The deformation of an object is described by the *strain tensor* e_{ij}, which is a symmetric tensor of type (0,2). The forces to which the strain gives rise are described by the *stress tensor*, σ^{ij}, usually also symmetric, and these are linked via a tensor of elastic constants c^{ijkl} as $\sigma^{ij} = c^{ijkl} e_{kl}$. We will study stress and strain later in this chapter.

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

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

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

An important operation is *contraction*, which consists of setting a contravariant index index equal to a covariant index and summing over them This reduces the type of tensor, so

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

is a tensor of type $(0,3)$. The reason for this is that setting an upper index and a lower index to a common value μ, and summing over μ, leads to the factor $\cdots (A^{-1})^{\mu}_{\alpha} A^{\beta}_{\mu} \cdots$ appearing in the transformation rule, but

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

and the Kroneker delta effects a summation over the corresponding pair of indices in the transformed tensor. For example, the combination $x^{\mu} f_{\mu}$ takes the same value in all bases — as it should since it is equal to $\mathbf{f}(\mathbf{x})$, and both $\mathbf{f}(\ )$ and $\mathbf{x}$ are basis-independent objects.

Remember that upper indices can only be contracted with lower indices, and vice-versa.

1.2.2 Tensor Product Spaces

We may regard the set of numbers $Q^{ij}{}_{klm}$ as being the components of an an object $\mathbf{Q}$ which is element of the vector space of type $(2,3)$ tensors. We will 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 $\{\mathbf{e}_i\}$ for V, and the dual basis $\{e^{*i}\}$ for V^*, we expand it as

$$\mathbf{Q} = Q^{ij}{}_{klm}\, \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}^{*k} \otimes \mathbf{e}^{*l} \otimes \mathbf{e}^{*m}.$$

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

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

associative,

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

but is not commutative,

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

Everthing commutes with the field however,

$$\lambda(\mathbf{a} \otimes \mathbf{b}) = (\lambda \mathbf{a}) \otimes \mathbf{b} = \mathbf{a} \otimes (\lambda \mathbf{b}),$$

so, if

$$\mathbf{e}_i = A_i^j \mathbf{e}'_j,$$

then

$$\mathbf{e}_i \otimes \mathbf{e}_j = A_i^k A_j^l \, \mathbf{e}'_k \otimes \mathbf{e}'_l.$$

From the analogous formula for $\mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}^{*k} \otimes \mathbf{e}^{*l} \otimes \mathbf{e}^{*m}$ we can reproduce the transformation rule for the components of $\mathbf{Q}$

The meaning of the tensor product of a set of vector spaces should now be clear: 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 $V \times V \times V :\rightarrow \mathbf{R}$. We define

$$\mathbf{e}^{*\lambda} \otimes \mathbf{e}^{*\mu} \otimes \mathbf{e}^{*\nu} \, (\mathbf{e}_\alpha, \mathbf{e}_\beta, \mathbf{e}_\gamma) = \delta_\alpha^\lambda \, \delta_\beta^\mu \, \delta_\gamma^\nu.$$

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

$$\mathbf{e}^{*\lambda} \otimes \mathbf{e}^{*\mu} \otimes \mathbf{e}^{*\nu} \, (\mathbf{e}_\alpha \otimes \mathbf{e}_\beta \otimes \mathbf{e}_\gamma) = \delta_\alpha^\lambda \, \delta_\beta^\mu \, \delta_\gamma^\nu,$$

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

$$V^* \otimes V^* \otimes V^* \simeq (V \otimes V \otimes V)^*.$$

[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 Cartesian product of two vector spaces can be given the structure of a vector space by defining $\lambda(\mathbf{x}_1, \mathbf{y}_1) + \mu(\mathbf{x}_2, \mathbf{y}_2) = (\lambda \mathbf{x}_1 + \mu \mathbf{x}_2, \lambda \mathbf{y}_1 + \mu \mathbf{y}_2)$, but this construction does not lead to the tensor-product. Instead it is the *direct sum* $V \oplus W$.

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 $\mathbf{g} : V \to V^*$.

Tensor Products and Quantum Mechanics

If we have two quantum mechanical systems with Hilbert spaces $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$, the Hilbert space for the combined system is $\mathcal{H}^{(1)} \otimes \mathcal{H}^{(1)}$. 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 therefore a linear combination

$$\mathbf{\Psi} = \sum_{n_1,n_2} \psi_{n_1,n_2} |n_1, n_2\rangle,$$

where

$$\psi_{n_1,n_2} = \langle n_1, n_2 | \Psi \rangle,$$

regarded as a function of n_1, n_2, is the wavefunction. This is the tensor product construction in disguise. To unmask it, we simply make the notational translation

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

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

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

If we can find vectors

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

such that

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

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

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 $\mathbf{C}^n$, 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 $\mathbf{CP}^{n-1}$ 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 $\mathbf{CP}^{nm-1}$, 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: 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 parametrs 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) Using your result 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.

1.2.3 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 $Q^{ij}{}_{kjl} = Q^{ji}{}_{kjl}$) or *skew-symmetric* ($Q^{ij}{}_{kjl} = -Q^{ji}{}_{kjl}$) in one basis, it remains so after the bases have 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

$$\mathbf{A} = \frac{1}{2} A^{ij}\, \mathbf{e}_i \wedge \mathbf{e}_j,$$

where the basis elements obey $\mathbf{e}_i \wedge \mathbf{e}_j = -\mathbf{e}_j \wedge \mathbf{e}_i$ and the coefficients are skew-symmetric, $A^{ij} = -A^{ji}$. The half (replaced by $1/p!$ when there are p indices) is convenient in that independent components only appear once in the sum.

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

$$\mathbf{A} = A^{ij}\, \mathbf{e}_i \odot \mathbf{e}_j$$

where $\mathbf{e}_i \odot \mathbf{e}_j = \mathbf{e}_j \odot \mathbf{e}_i$ and $A^{ij} = A^{ji}$. (We do not include 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,

$$\mathbf{e}^{*i} \wedge \mathbf{e}^{*j}\,(\mathbf{e}_\mu, \mathbf{e}_\nu) = \delta^i_\mu \delta^j_\nu - \delta^i_\nu \delta^j_\mu$$

and

$$\mathbf{e}^{*i} \wedge \mathbf{e}^{*j}\,(\mathbf{e}_\mu \wedge \mathbf{e}_\nu) = \delta^i_\mu \delta^j_\nu - \delta^i_\nu \delta^j_\mu.$$

We need two terms here because the skew-symmetry of $\mathbf{e}^{*i} \wedge \mathbf{e}^{*j}(\ ,\ )$ in its slots does not allow us the luxury of demanding that the $\mathbf{e}_i$ be inserted in the exact order of the $\mathbf{e}^{*i}$ to get a non-zero answer. Because a p-th order form has $p!$ terms, 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 $\mathbf{A} = \frac{1}{2} A_{ij} \mathbf{e}^{*i} \wedge \mathbf{e}^{*j}$ and $\mathbf{B} = \frac{1}{2} B^{ij} \mathbf{e}_i \wedge \mathbf{e}_j$, we have

$$\mathbf{A}(\mathbf{B}) = \frac{1}{2} A_{ij} B^{ij},$$

and the sum is only over the independent terms in the sum.

The wedge ($\wedge$) product notation is standard in mathematics where skew-symmetry is implied. 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: 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 $(d+p-1)!/p!(d-1)!$.

Bosons and Fermions

Spaces of symmetric and antisymmetric 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-particles states with basis $\mathbf{e}_i$, then the N-boson space is $\mathrm{Sym}^N\mathcal{H}$ consisting of states

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

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

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

The symmetry of the Bose wavefunction

$$\Phi_{i_1 i_2 \ldots i_N} = \Phi_{i_2 i_1 \ldots i_N}$$

etc., and the antisymmetry of the Fermion wavefunction

$$\Psi_{i_1 i_2 \ldots i_N} = -\Psi_{i_2 i_1 \ldots i_N},$$

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

$$\begin{aligned} \mathbf{\Psi} &= \frac{1}{N!} \Psi_{i_1 i_2 \ldots i_N} \, \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N} \\ &= \boldsymbol{\psi}^{(1)} \wedge \boldsymbol{\psi}^{(2)} \wedge \cdots \wedge \boldsymbol{\psi}^{(N)} \\ &= \psi^{(1)}_{i_1} \psi^{(2)}_{i_2} \cdots \psi^{(N)}_{i} \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N}. \end{aligned}$$

Comparing the coefficients of $\mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_N}$ shows that so the many-body wavefunction can be written as

$$\Psi_{i_1 i_2 \ldots i_N} = \begin{vmatrix} \psi^{(1)}_{i_1} & \psi^{(1)}_{i_2} & \cdots & \psi^{(1)}_{i_N} \\ \psi^{(2)}_{i_1} & \psi^{(2)}_{i_2} & \cdots & \psi^{(2)}_{i_N} \\ \vdots & \vdots & \ddots & \vdots \\ \psi^{(N)}_{i_1} & \psi^{(N)}_{i_2} & \cdots & \psi^{(N)}_{i_N} \end{vmatrix}.$$

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 such 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 $\boldsymbol{\psi}^{(a)}$. 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 $\boldsymbol{\Psi}$ to be decomposable into single-particle states. These are that

$$\Psi_{i_1 i_2 \ldots i_{N-1}[j_1} \Psi_{j_1 j_2 \ldots j_{N+1}]} = 0$$

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

$$\Psi_{i_1 i_2 j_1} \Psi_{j_2 j_3 j_4} - \Psi_{i_1 i_2 j_2} \Psi_{j_1 j_3 j_4} + \Psi_{i_1 i_2 j_3} \Psi_{j_1 j_2 j_4} - \Psi_{i_1 i_2 j_4} \Psi_{j_1 j_2 j_3} = 0.$$

These conditions are called the *Plücker relations* after Julius Plücker who discovered them long before before the advent of quantum mechanics[3]. It is easy to show that they are necessary conditions. It is not so easy to show that they are sufficient, and we will defer proving this until we have more tools at our disposal. 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.

1.2.4 Tensor Character of Linear Maps and Quadratic Forms

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

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

examining the action of the map on the basis elements:

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

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

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

We therefore have

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

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

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

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

$$M^\nu{}_\mu = A^\rho_\mu (A^{-1})^\nu_\sigma M'^\sigma{}_\rho,$$

or, conversely,

$$M'^\nu{}_\mu = (A^{-1})^\rho_\mu A^\nu_\sigma M^\sigma{}_\rho.$$

Thus a matrix representing a linear map has the tensor character suggested by the position of its indicices, *i.e.* it transforms as a type $(1,1)$ tensor. $\mathbf{M}$ is therefore simultaneously an element of $\mathrm{Map}\,(V \to V)$ and an element of $V \otimes V^*$.

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

We can write

$$\mathbf{Q}(\mathbf{x}) = Q_{ij} x^i x^j = x^i Q_{ij} \, x^j = \mathbf{x}^T \mathbf{Q} \mathbf{x}$$

where $Q_{ij} = \mathbf{Q}(\mathbf{e}_i, \mathbf{e}_j)$ is a symmetric matrix, and $\mathbf{x}^T \mathbf{Q} \mathbf{x}$ is standard matrix multiplication notation. Just as with the metric tensor, the coefficients Q_{ij} transform as a doubly covariant, type $(0,2)$ tensor. Thus although both linear maps and quadratic forms can be represented by matrices, these matrices correspond to different types of tensor and transform quite differently under a change of basis. For example, a matrix representing a linear map has a basis-independent determinant. One can certainly compute the determinant of the matrix representing a quadratic form in some particular basis, but

when you change basis and calculate the determinant of the resulting new matrix, you will get a different number. Notice also, that the *trace* of a matrix representing a linear map

$$\mathrm{tr}\,\mathbf{M} = M^{\mu}{}_{\mu}$$

is a tensor of type $(0,0)$, i.e. a scalar, and therefore basis independent.

Basis independent quantities such as the determinant and trace of linear map are called *invariants*.

Exercise: Use the distinction between the transformation law of a quadratic form and that of a linear map to resolve the following "paradox".

a) In quantum mechanics we are taught that the matrices representing two operators can be simultaneously diagonalized only if they commute.
b) 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 coordinates Q_i such that L becomes

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

In b) 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.

1.2.5 Numerically Invariant Tensors

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

$$\delta'^i_j = A^i_{i'}(A^{-1})^{j'}_j \delta^{i'}_{j'} = A^i_k (A^{-1})^k_j = \delta^i_j.$$

In other words the Kroneker 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 δ_{ij}.

Now consider an n-dimensional space with a tensor $\eta_{i_1 i_2 \ldots i_n}$ whose components, in some basis, coincides with the Levi-Civita symbol $\epsilon_{i_1 i_2 \ldots i_n}$. We find that in a new frame the components are

$$\begin{aligned} \eta'_{i_1 i_2 \ldots i_n} &= (A^{-1})^{j_1}_{i_1} (A^{-1})^{j_2}_{i_2} \cdots (A^{-1})^{j_n}_{i_n} \epsilon_{j_1 j_2 \ldots j_n} \\ &= \det (A^{-1})\, \epsilon_{i_1 i_2 \ldots i_n} \\ &= \det (A^{-1})\, \eta_{i_1 i_2 \ldots i_n}. \end{aligned}$$

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

Consider also the quantity

$$\sqrt{g} \stackrel{\text{def}}{=} \sqrt{\det [g_{ij}]}.$$

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

$$\det [g'_{ij}] = \det [(A^{-1})^{i'}_{i} (A^{-1})^{j'}_{j} g_{i'j'}] = (\det A)^{-2} \det [g_{ij}],$$

we see that

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

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

$$\varepsilon_{i_1 i_2 \ldots i_n} = \sqrt{g}\, \epsilon_{i_1 i_2 \ldots i_n},$$

and assume that $\varepsilon_{i_1 i_2 \ldots i_n}$ 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 for which $\det A$ is positive, we see that factors of $\det A$ conspire to give

$$\varepsilon'_{i_1 i_2 \ldots i_n} = \sqrt{g'}\, \epsilon_{i_1 i_2 \ldots i_n}.$$

A similar exercise indictes that if we define $\epsilon^{i_1 i_2 \ldots i_n}$ to be numerically equal to $\epsilon_{i_1 i_2 \ldots i_n}$, then

$$\varepsilon^{i_1 i_2 \ldots i_n} = \frac{1}{\sqrt{g}}\, \epsilon^{i_1 i_2 \ldots i_n}$$

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 we are in an even-dimensional space and are given a skew-symetric tensor F_{ij}, we can therefore construct an invariant

$$\varepsilon^{i_1 i_2 \ldots i_n} F_{i_1 i_2} \cdots F_{i_{n-1} i_n} = \frac{1}{\sqrt{g}} \epsilon^{i_1 i_2 \ldots i_n} F_{i_1 i_2} \cdots F_{i_{n-1} i_n}.$$

Similarly, given an skew-symmetric covariant tensor $F_{i_1 \ldots i_m}$ with $m < n$ indices we can form its *dual*, F^*, a $(n-m)$-contravariant tensor with components

$$(F^*)^{i_{m-1} \ldots i_n} = \frac{1}{m!} \varepsilon^{i_1 i_2 \ldots i_n} F_{i_1 \ldots i_m} = \frac{1}{\sqrt{g}} \frac{1}{m!} \epsilon^{i_1 i_2 \ldots i_n} F_{i_1 \ldots i_m}.$$

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

1.3 Cartesian Tensors

If we restrict ourselves to Cartesian co-ordinate systems with 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. If we further only allow orthogonal transformations A^i_j with $\det A = 1$ (the so-called *proper* orthogonal transformations), then both δ_{ij} and $\epsilon_{i_1 i_2 \ldots i_n}$ are tensors whose components are numerically the same in all bases. Objects which are tensors under the proper orthogonal group are called Cartesian tensors. We shall usually write their indices as suffixes.

For many physics purposes Cartesian tensors are all we need. The rest of this section is devoted to some examples.

1.3.1 Stress and Strain

Tensor calculus arose from the study of elasticity — hence the name.

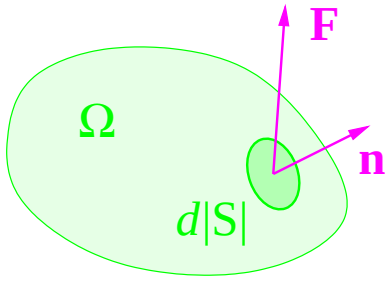
Suppose that an elastic body is deformed so that the point that was at Cartesian co-ordinate x_i is moved to $x_i + \eta_i$. We define the *strain tensor*, e_{ij}, by

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

It is automatically symmetric in its indices. We will leave for later a discussion of why this is the natural definition of strain, and also the modifications necessary if we were to use a non-cartesian co-ordinate system.

To define the *stress tensor*, σ_{ij} we consider a portion of the body Ω, 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

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



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 coordinate 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[4] that $\sigma_{ij} = \sigma_{ji}$.

The stress is related to the strain via the tensor of *elastic constants*, c_{ijkl}, by

$$\sigma_{ij} = c_{ijkl} e_{kl}.$$

The fourth rank tensor of eleastic constants has the symmetry properties,

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

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

[4]If the material is subject to a torque per unit volume, as in the case of a magnetic material in a magnetic field, then the stress tensor is no longer symmetric.

Exercise: Show that these symmetries imply that a general homogeneous material has 21 independent elastic constants. (This result was originally obtained by George Green, of Green's function fame.)

For an *isotropic* material, that is a material whose properties are invariant under the full rotation group, the tensor of elastic constants must be made up of numerically invariant tensors, and the most general such combination with the required symmetries is

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

and so there are only *two* independent elastic constants. In terms of them

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

The quantities λ and μ are called the *Lamé* constants. 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 them.

The bulk modulus κ is defined by

$$\frac{dV}{V} = -\kappa dP$$

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

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

The strains are therefore

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

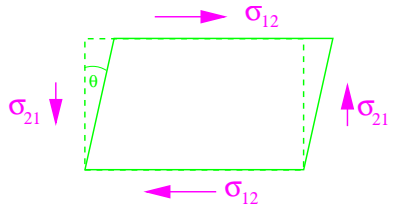
Plugging into the stress-strain relation gives

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

Thus

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

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



Shear strain.

The applied shear stress is $\sigma_{12} = \sigma_{21}$, and the shear modulus, n, is defined so that $n\theta = \sigma_{12}$. Plugging into the stress-strain relation gives

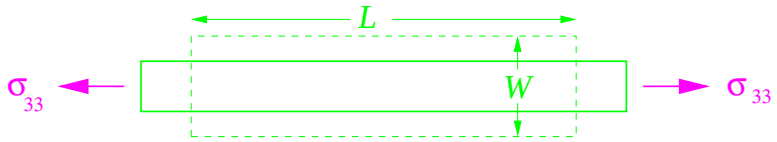
$$n = \mu.$$

We could therefore have written

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

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

Young's modulus, Y, is defined in terms of stretching a wire of initial length L and square cross section of side W under an applied tension $T = \sigma_{33}W^2$ at the ends.



Stretched wire.

We then have

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

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

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

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

$$\eta_3 = z\left(\frac{dL}{L}\right), \quad \eta_1 = x\left(\frac{dW}{W}\right) = -\sigma x\left(\frac{dL}{L}\right), \quad \eta_2 = y\left(\frac{dW}{W}\right) = -\sigma y\left(\frac{dL}{L}\right),$$

so the strain components are

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

We therefore have

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

leading to

$$Y = \lambda(1-2\sigma) + 2\mu.$$

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

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

This tells us that

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

and hence

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

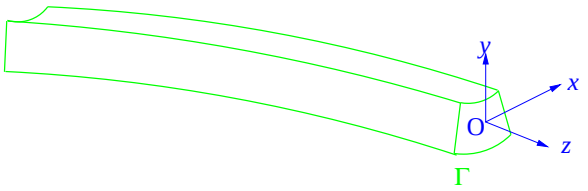
Other relations, following from those above, are

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

Exercise: A steel beam is forged so that its cross section has the shape of a region $\Gamma \in \mathbf{R}^2$. The centroid, O, of each cross section is defined so that

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

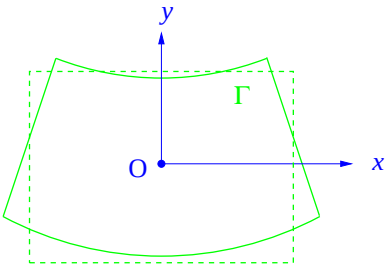
where the co-ordinates x, y are defined with the centroid O as the origin. The beam is slightly bent so that near a particular cross-section it has radius of curvature R.



Bent beam.

Assume that the deformation is such that

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



The original (dashed) and deformed (solid) cross-section.

Notice how, for positive Poisson ratio, the cross section is deformed *anticlastically* — the sides bends *up* as the beam bends *down*. Show that

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

Also show that the other three strain components are zero. Next show that

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

and that all other components of the stress tensor vanish.

Deduce from this that the assumed deformation satisfies the free surface boundary condition, and so is indeed the way the beam deforms. The total elastic energy is given by

$$E = \iiint_{\text{beam}} \frac{1}{2}e_{ij}c_{ijkl}e_{kl}\,d^3x.$$

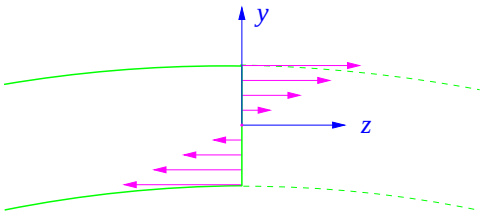
Show that for our bent rod, this reduces to

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

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

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

is the moment of inertia of the region Γ about the y axis — *i.e.* an axis through the centroid, and perpendicular both to the length of the beam and to the plane into which it is bent. On the right hand side y'' denotes the second derivative of the deflection of the beam with respect to the arc-length. This last formula for the strain energy was used several times in MMA.



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

1.3.2 The 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 numerically equal, but pointing in opposite directions. There would therefore be no net external force on the cube. When σ_{ij} is *not* constant then the net force acting on a element of volume dV is

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

Consequently, 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.

Let $\mathbf{E}$ and $\mathbf{B}$ be the electric and magnetic fields. For simplicity, initially assume them to be static. The force per unit volume exerted by these fields

on a charge and current distribution is

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

Writing $\rho = \operatorname{div} \mathbf{D}$, with $\mathbf{D} = \epsilon_0\mathbf{E}$ we find that the force per unit volume due the electric field can be written as

$$\rho E_i = (\partial_j D_j)E^i = \epsilon_0 \partial_j \left(E_i E_j - \frac{1}{2}\delta_{ij}|E|^2 \right).$$

Here we have used the fact that $\operatorname{curl} \mathbf{E}$ is zero for static fields. 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 is

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

The quantity

$$\sigma_{ij} = \epsilon_0 \left(E_i E_j - \frac{1}{2}\delta_{ij}|E|^2 \right) + \mu_0 \left(H_i H_j - \frac{1}{2}\delta_{ij}|H|^2 \right)$$

is called the *Maxwell stress tensor*.

Michael Faraday was the first to intuit this stress picture of electromagnetic forces, which attributes both a longitudinal tension and a sideways pressure to the field lines.

Exercise: Allow the fields in the preceding calculation to be time dependent. Show that Maxwell's equations 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, so the stress tensor can also be thought of as a *momentum flux* tensor.

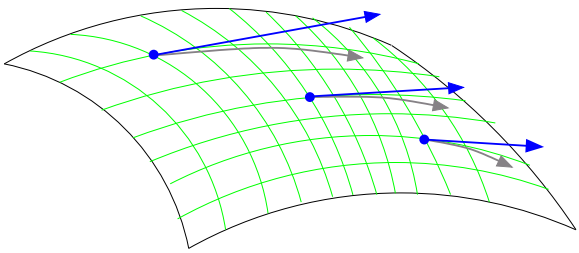
Chapter 2

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, and ultimately to calculus on manifolds.

2.1 Vector Fields and Covector Fields

Physics is full of vector fields — electric, magnetic, velocity fields, and so on. After struggling with it in introductory courses, we rather take the concept for granted. There are some real 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 the field vectors, $\mathbf{E}(x_1)$ and $\mathbf{E}(x_2)$, at two points separated by several meters. 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 vector space. This point of view seems even more reasonable when we consider velocity vectors on a curved surface.



A velocity vector lives in the *tangent space* to the surface at each point, and

each of these spaces is differently oriented subspace of the higher dimensional ambient space. Mathematicians call such a collection of vector spaces — one for each of the points in the surface — a *vector bundle* over the surface. Thus the *tangent bundle* over a surface is the totality of all these different vector spaces tangent to the surface.

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 such that some neighbourhood of each point can be described by means of an n-dimensional co-ordinate system. Where a pair of such *co-ordinate charts* overlap, the transformation formula giving one set of co-ordinates as a function of the other is required to be a smooth (C^∞) function, and to possess a smooth inverse. The collection of all smoothly related coordinate charts is called an *atlas*. There is a more formal definition of a manifold, containing some restrictions, but we won't make use of it. 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

Classical provides a good illustration of these ideas. The configuration space M of a mechanical system is almost always a manifold. When a mechanical 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. The tangent bundle, denoted by TM, is the $2n$ dimensional space whose points consist of a point p in M paired 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

$$\mathbf{X} \cdot \nabla = X^\mu \partial_\mu \tag{2.1}$$

where ∂_μ is shorthand for $\partial/\partial x^\mu$. When the numbers X^μ are functions of the co-ordinates x^σ, this object will be called a tangent-vector field, and we shall write[1]

$$X = X^\mu \partial_\mu. \tag{2.2}$$

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.*

$$\begin{aligned} x^1 &= x^1(z^1, z^2, \ldots, z^n), \\ x^2 &= x^2(z^1, z^2, \ldots, z^n), \\ &\vdots \\ x^n &= x^n(z^1, z^2, \ldots, z^n), \end{aligned} \tag{2.3}$$

then the chain rule for partial differentiation gives

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \frac{\partial z^\nu}{\partial x^\mu}\frac{\partial}{\partial z^\nu} = \left(\frac{\partial z^\nu}{\partial x^\mu}\right)\partial'_\nu. \tag{2.4}$$

By demanding that

$$X = X^\mu \partial_\mu = X'^\nu \partial'_\nu \tag{2.5}$$

we find

$$X'^\nu = \left(\frac{\partial z^\nu}{\partial x^\mu}\right) X^\mu \tag{2.6}$$

or, using

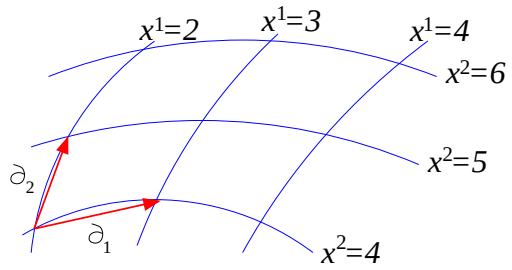
$$\frac{\partial x^\sigma}{\partial z^\nu}\frac{\partial z^\nu}{\partial x^\mu} = \frac{\partial x^\sigma}{\partial x^\nu} = \delta^\sigma_\mu, \tag{2.7}$$

$$X^\nu = \left(\frac{\partial x^\nu}{\partial z^\mu}\right) X'^\mu. \tag{2.8}$$

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

[1] 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 will just be confusing.

It is worth pointing out that the basis vectors ∂_μ are *not* unit vectors. At the moment we have no metric and therefore no notion of length anyway, so we can't 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".



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 coordinates 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 coordinate basis, the tangent vector representing the velocity field due to a one radian per second rigid rotation about the z axis is

$$V_z = \partial_\phi. \tag{2.9}$$

Similarly

$$\begin{aligned} V_x &= -\sin\phi\,\partial_\theta - \cot\theta\cos\phi\,\partial_\phi, \\ V_y &= \cos\phi\,\partial_\theta - \cot\theta\sin\phi,\partial_\phi, \end{aligned} \tag{2.10}$$

represent rigid rotations about the x and y axes.

What about the dual spaces? For these a cute notational game, due to Elié Cartan, is played. We write the basis objects dual to the ∂_μ as $dx^\mu(\ )$. Thus

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

Acting on vector field $X = X^\mu\partial_\mu$, the object dx^μ returns its components

$$dx^\mu(X) = dx^\mu(X^\nu\partial_\nu) = X^\nu dx^\mu(\partial_\nu) = X^\nu\delta^\mu_\nu = X^\mu. \tag{2.12}$$

Actually, any function $f(x)$ on our space (we will write $f \in C^\infty(M)$ for smooth functions on a manifold M) gives rise to a field of covectors in TM^*. This is because our vector field X acts on the scalar function f as

$$Xf = X^\mu \partial_\mu f \tag{2.13}$$

and what we get is another scalar function. This new function gives a number — and thus an element of the field $\mathbf{R}$ — 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". Thus

$$df(X) \stackrel{\text{def}}{=} Xf = X^\mu \frac{\partial f}{\partial x^\mu}. \tag{2.14}$$

If we replace f with the co-ordinate x^ν, we have

$$dx^\nu(X) = X^\mu \frac{\partial x^\nu}{\partial x^\mu} = X^\mu \delta^\nu_\mu = X^\nu, \tag{2.15}$$

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

$$df(X) = \frac{\partial f}{\partial x^\mu} X^\mu = \frac{\partial f}{\partial x^\mu} dx^\mu(X) \tag{2.16}$$

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

$$df = \frac{\partial f}{\partial x^\mu} dx^\mu. \tag{2.17}$$

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 some 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

$$f_\mu \, dx^\mu = f'_\nu \, dz^\nu, \tag{2.18}$$

by using

$$dx^\mu = \left(\frac{\partial x^\mu}{\partial z^\nu}\right) dz^\nu. \tag{2.19}$$

We find

$$f'_\nu = \left(\frac{\partial x^\mu}{\partial z^\nu}\right) f_\mu. \tag{2.20}$$

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

$$Q'^{\lambda\mu}{}_{\rho\sigma\tau}(z) = \frac{\partial z^\lambda}{\partial x^\alpha}\frac{\partial z^\mu}{\partial x^\beta}\frac{\partial x^\gamma}{\partial z^\rho}\frac{\partial x^\delta}{\partial z^\sigma}\frac{\partial x^\epsilon}{\partial z^\tau} Q^{\alpha\beta}{}_{\gamma\delta\epsilon}(x). \tag{2.21}$$

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.

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 combine them. Look at XY, for example:

$$XY = X^\mu \partial_\mu (Y^\nu \partial_\nu) = X^\mu Y^\nu \partial^2_{\mu\nu} + X^\mu \left(\frac{\partial Y^\nu}{\partial x^\mu}\right) \partial_\nu. \tag{2.22}$$

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*:

$$[X, Y] = XY - YX = \left(X^\mu(\partial_\mu Y^\nu) - Y^\mu(\partial_\mu X^\nu)\right)\partial_\nu. \tag{2.23}$$

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

$$[X, Y]^\nu \equiv X^\mu(\partial_\mu Y^\nu) - Y^\mu(\partial_\mu X^\nu) \tag{2.24}$$

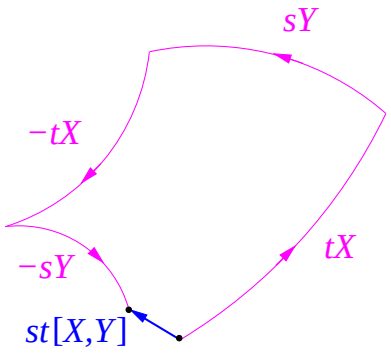
are the components of a new contravariant vector made from the two old vector fields. It is called the *Lie bracket* of the two fields, and has a geometric interpretation.

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

$$\frac{dx^\mu}{dt} = X^\mu(x^1, x^2, \ldots, x^d), \tag{2.25}$$

with initial condition $x^\mu(0) = x_0^\mu$. 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.



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 a d-dimensional manifold M we select $n < d$ linearly independent vector fields $X_i(x)$ at each point x. (Such a set is sometimes called a *distribution* although the concept has nothing to do with objects like "$\delta(x)$" which are also called "distributions".) How can we tell if there is a surface N through each point, such that the X_i form a basis for the tangent space to N at that point? The answer is given by *Frobenius' theorem*.

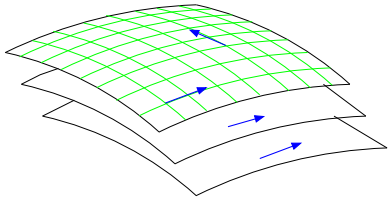
First a definition: If there are functions ${c_{ij}}^k(x)$ such that

$$[X_i, X_j] = {c_{ij}}^k(x) X_k, \tag{2.26}$$

i.e. the Lie brackets close within the set $\{X_i\}$ at each point then the distribution is said to be *involutive.*

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 $\mu = n+1, \ldots, x^d$. 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. Going the other way requires us to form the flows (field lines) of the vector fields and show that they define a surface, or surface. This is not hard, but takes more space than we want to devote to the topic.



A foliation by surfaces.

The stack of surfaces N locally fills out the ambient manifold. It is said to be a *foliation* of the higher dimensional space. For examples the set of spheres of radius r foliate $\mathbf{R}^3$ except at the origin.

Physics Application: Holonomic and anholonomic constraints.

Holonomic constraints are those such as requiring a mass to be at fixed distance from the origin (a spherical pendulum). If we were just told that the velocity vector was constrained to be perpendicular to the radius vector, we would see that such the resulting two-dimensional distribution was involutive, and would deduce that $\mathbf{R}^3$ decomposes into a set of of invariant surfaces which are the spheres of radius r. Thus holonomic constraints restrict the motion to a surface.

If, on the other hand, we have a ball rolling on a table, we have a five-dimensional configuration space parameterized by the centre of mass (x, y) of the ball, and the three Euler angles (θ, ϕ, ψ) defining its orientation. The no-slip rolling condition links the rate of change of the Euler angles to the velocity of the centre of mass. At each point, we are free to roll the ball in two directions, and may expect that the reachable configurations are a two dimensional subspace of the full five dimensional space. The resulting vector fields are not in involution, however, and by calculating enough Lie brackets we eventually obtain five linearly independent velocity vector fields. Thus, starting from one configuration we can reach any other. The no-slip rolling condition is therefore non-integrable, or *anholonomic*. Such systems are tricky to deal with in Lagrangian dynamics.

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

$$\mathcal{L}_X f \stackrel{\text{def}}{=} Xf, \tag{2.27}$$

on a vector field by

$$\mathcal{L}_X Y \stackrel{\text{def}}{=} [X, Y], \tag{2.28}$$

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 function $F(Y)$. Leibniz' rule is then the statement that

$$\mathcal{L}_X F(Y) = (\mathcal{L}_X F)(Y) + F(\mathcal{L}_X Y), \tag{2.29}$$

and 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

$$XF(Y) = (\mathcal{L}_X F)(Y) + F([X, Y]), \tag{2.30}$$

and so

$$(\mathcal{L}_X F)(Y) = XF(Y) - F([X, Y]). \tag{2.31}$$

In components this is

$$\begin{aligned} (\mathcal{L}_X F)(Y) &= X^\nu \partial_\nu (F_\mu Y^\mu) - F_\nu (X^\mu \partial_\mu Y^\nu - Y^\mu \partial_\mu X^\nu) \\ &= (X^\nu \partial_\nu F_\mu + F_\nu \partial_\mu X^\nu) Y^\mu. \end{aligned} \tag{2.32}$$

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 arbitrary spectator Y^ν, we have a formula for $\mathcal{L}_X F$ in components:

$$(\mathcal{L}_X F)_\mu = X^\nu \partial_\nu F_\mu + F_\nu \partial_\mu X^\nu. \tag{2.33}$$

Another example is the Lie derivative of a type $(0, 2)$ tensor, such as the metric tensor, which is

$$(\mathcal{L}_X g)_{\mu\nu} = X^\alpha \partial_\alpha g_{\mu\nu} + g_{\mu\alpha} \partial_\nu X^\alpha + g_{\alpha\nu} \partial_\mu X^\alpha. \tag{2.34}$$

This Lie derivative measures the extent to which a displacement $x^\mu \to x^\mu + \epsilon \eta^\mu$ deforms the geometry.

Exercise: Suppose we have an unstrained block of material in real space. A coordinate system ξ^1, ξ^2, ξ^3, is attached to the atoms of the body. The point with coordinate 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 coordinates.

a) Show that the induced metric in the ξ coordinate 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 a strain vector field $\eta(\xi)$. The point ξ^μ is moved to what was $\xi^\mu + \epsilon\eta^\mu(\xi)$, or equivalently, the atom initially at $x^a(\xi)$ is moved to $x^a + \epsilon\eta^\mu \partial x^a/\partial\xi^\mu$. Show that the new induced metric is

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

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 coordinate 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 *finite strain* tensor is defined as

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

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.

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

Exercise: The metric on the unit sphere equipped with polar coordinates is

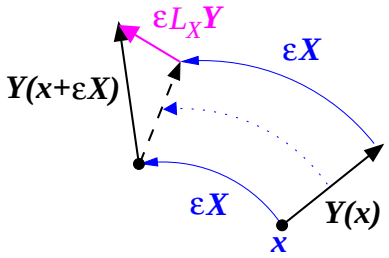
$$g(\ ,\ ) = d\theta \otimes d\theta + \sin^2\theta d\phi \otimes d\phi.$$

Consider

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

the vector field of a rigid rotation about the x axis. Show that $\mathcal{L}_{V_x} g = 0$.

The geometric interpretation of the Lie derivative 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 lives, 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.



In other words, 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 $\epsilon \mathcal{L}_X Y = \epsilon[X,Y]$.

2.3 Exterior Calculus

2.3.1 Differential Forms

The object we introduced in the previous section, the dx^μ, are called one-forms, or differential one-forms. They live in the cotangent bundle, T^*M, of M. (In more precise language, they are *sections* of the cotangent bundle, and vector fields are sections of the tangent bundle.) If we consider the p-th skew-symmetric tensor power $\bigwedge^p(T^*M)$ of the space of one-forms we get objects called p-forms.

For example,

$$A = A_\mu dx^\mu = A_1 dx^1 + A_2 dx^2 + A_3 dx^3, \tag{2.35}$$

is a 1-form,

$$F = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu = F_{12} dx^1 \wedge dx^2 + F_{23} dx^2 \wedge dx^3 + F_{31} dx^3 \wedge dx^1, \tag{2.36}$$

is a 2-form, and

$$\begin{aligned}\Omega &= \frac{1}{3!}\Omega_{\mu\nu\sigma}dx^\mu \wedge dx^\nu \wedge dx^\sigma \\ &= \Omega_{123}dx^1 \wedge dx^2 \wedge dx^3, \end{aligned} \tag{2.37}$$

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

$$\Omega_{\mu\nu\sigma} = \Omega_{\nu\sigma\mu} = \Omega_{\sigma\mu\nu} = -\Omega_{\nu\mu\sigma} = -\Omega_{\mu\sigma\nu} = -\Omega_{\sigma\nu\mu}. \tag{2.38}$$

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

$$dx^\mu \wedge dx^\nu(\partial_\alpha, \partial_\beta) = \delta^\mu_\alpha\delta^\nu_\beta - \delta^\mu_\beta\delta^\nu_\alpha. \tag{2.39}$$

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

$$a \wedge b = (-1)^{pq}\, b \wedge a. \tag{2.40}$$

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 seem rather exotic, 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 these, to two simple formulae! Indeed Cartan's calculus of p-forms is slowly supplanting traditional vector calculus, much as Willard Gibbs' vector calculus supplanted the

tedious component-by-component formulae 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

$$d(a \wedge b) = da \wedge b + (-1)^p a \wedge db. \tag{2.41}$$

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

$$\begin{aligned} d\Omega &= \frac{1}{p!} d(\Omega_{i_1,\ldots,i_p}) dx^{i_1} \cdots dx^{i_p} \\ &= \frac{1}{p!} \left(\partial_k \Omega_{i_1,\ldots,i_p}\right) dx^k dx^{i_1} \cdots dx^{i_p}. \end{aligned} \tag{2.42}$$

Now compute

$$d(d\Omega) = \frac{1}{p!} \left(\partial^2_{lk} \Omega_{i_1,\ldots,i_p}\right) dx^l dx^k dx^{i_1} \cdots dx^{i_p}. \tag{2.43}$$

Fortunately this is zero because $\partial^2_{lk}\Omega = \partial^2_{kl}\Omega$, while $dx^l dx^k = -dx^k dx^l$.
If $A = A_1 dx^1 + A_2 dx^2 + A_3 dx^3$, then

$$\begin{aligned} dA &= \left(\frac{\partial A_2}{\partial x^1} - \frac{\partial A_1}{\partial x^2}\right) dx^1 dx^2 + \left(\frac{\partial A_1}{\partial x^3} - \frac{\partial A_3}{\partial x^1}\right) dx^3 dx^1 + \left(\frac{\partial A_3}{\partial x^2} - \frac{\partial A_2}{\partial x^3}\right) dx^2 dx^3 \\ &= \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu, \end{aligned} \tag{2.44}$$

where

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.45}$$

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

$$dF = \left(\frac{\partial F_{23}}{\partial x^1} + \frac{\partial F_{31}}{\partial x^2} + \frac{\partial F_{12}}{\partial x^3}\right) dx^1dx^2dx^3. \tag{2.46}$$

This looks like a divergence.

In fact $d^2 = 0$, 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: Use axiom ii) to compute $d(d(a \wedge b))$ and confirm that it is zero.

Cartan's formulae

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

$$\begin{aligned} df(X) &= Xf. && (2.47)\\ d\eta(X,Y) &= X\eta(Y) - Y\eta(X) - \eta([X,Y]). && (2.48)\\ d\omega(X,Y,Z) &= X\omega(Y,Z) + Y\omega(Z,X) + Z\omega(X,Y) \\ &\quad -\omega([X,Y],Z) - \omega([Y,Z],X) - \omega([Z,X],Y). && (2.49) \end{aligned}$$

These formulae, and their higher-p analogues, express d in terms of geometric objects, and so make it clear that the exterior derivative is itself an intrinsic 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

$$\partial_\mu\eta_\nu\, dx^\mu dx^\nu(X,Y) = \partial_\mu\eta_\nu(X^\mu Y^\nu - X^\nu Y^\mu). \tag{2.50}$$

The right hand side is equal to

$$X^\mu \partial_\mu(\eta_\nu Y^\nu) - Y^\mu \partial_\mu(\eta_\nu X^\nu) - \eta_\nu(X^\mu \partial_\mu Y^\nu - Y^\mu \partial_\mu X^\nu). \tag{2.51}$$

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.

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

$$i_X\omega(\underbrace{\ldots\ldots}_{p-1\,\text{slots}}) = \omega(\overbrace{X, \underbrace{\ldots\ldots}_{p-1\,\text{slots}}}^{p\,\text{slots}}). \tag{2.52}$$

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

$$\omega_{j_i j_2 \ldots j_p} \to \omega_{k j_2 \ldots j_p} X^k, \tag{2.53}$$

but it is convenient to have a special symbol for this operation. Note that i_X is an anti-derivation, just as is d: if η and ω are p and q forms respectively, then

$$i_X(\eta \wedge \omega) = (i_X\eta) \wedge \omega + (-1)^p \eta \wedge (i_X\omega), \tag{2.54}$$

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

$$\begin{aligned} i_X(dx^\mu \wedge dx^\nu) &= dx^\mu \wedge dx^\nu(X^\alpha \partial_\alpha, \quad), \\ &= X^\mu dx^\nu - dx^\mu X^\nu, \\ &= (i_X dx^\mu) \wedge (dx^\nu) - dx^\mu \wedge (i_X dx^\nu). \end{aligned} \tag{2.55}$$

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

$$\mathcal{L}_X\omega = (d\, i_X + i_X d)\omega. \tag{2.56}$$

This 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

$$(d\, i_X + i_X d)f = i_X df = df(X) = Xf = \mathcal{L}_X f. \tag{2.57}$$

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

$$\begin{aligned}(d\, i_X + i_X d)(f_\nu\, dx^\nu) &= d(f_\nu X^\nu) + i_X(\partial_\mu f_\nu\, dx^\mu dx^\nu) \\ &= \partial_\mu(f_\nu X^\nu)dx^\mu + \partial_\mu f_\nu(X^\mu dx^\nu - X^\nu dx^\mu) \\ &= (X^\nu \partial_\nu f_\mu + f_\nu \partial_\mu X^\nu)dx^\mu. \end{aligned} \tag{2.58}$$

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 antiderivation 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.

2.4 Physical Applications

2.4.1 Maxwell's Equations

In relativistic[2] four-dimensional tensor notation the two source-free Maxwell's equations

$$\begin{aligned}\operatorname{curl} \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\ \operatorname{div} \mathbf{B} &= 0, \end{aligned}$$

reduce to the single equation

$$\frac{\partial F_{\mu\nu}}{\partial x^\lambda} + \frac{\partial F_{\nu\lambda}}{\partial x^\mu} + \frac{\partial F_{\lambda\mu}}{\partial x^\nu} = 0. \tag{2.59}$$

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

where

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{2.60}$$

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

$$\begin{aligned} F &= \frac{1}{2} F_{\mu\nu} dx^\mu dx^\nu \\ &\equiv B_x dydz + B_y dzdx + B_z dxdy + E_x dxdt + E_y dydt + E_z dzdt, \end{aligned} \tag{2.61}$$

is a Minkowski space 2-form.

Exercise: Verify that these Maxwell equations are equivalent to $dF = 0$.

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

The two Maxwell equations with sources

$$\begin{aligned} \operatorname{div} \mathbf{D} &= \rho \\ \operatorname{curl} \mathbf{H} &= \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \end{aligned} \tag{2.62}$$

reduce in 4-tensor notation to the single equation

$$\partial_\mu F^{\mu\nu} = J^\nu. \tag{2.63}$$

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 this 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 Lorentzian signature metrics we should replace $\sqrt{g}$ by $\sqrt{-g}$.) We define "$\star$" to be a linear map

$$\star : \bigwedge^p (T^*M) \to \bigwedge^{(d-p)} (T^*M) \tag{2.64}$$

such that

$$\star\, dx^{i_1} \dots dx^{i_p} \stackrel{\text{def}}{=} \frac{1}{(d-p)!}\sqrt{g} g^{i_1 j_1} \dots g^{i_p j_p} \epsilon_{j_1 \cdots j_p j_{p+1} \cdots j_d} dx^{j_{p+1}} \dots dx^{j_d}. \quad (2.65)$$

Although this definition looks a trifle involved, computations involving it are not so intimidating. The trick is always to work 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

$$\star\,(\mathbf{e}^{*i_1} \wedge \mathbf{e}^{*i_2} \wedge \cdots \wedge \mathbf{e}^{*i_p}) = \mathbf{e}^{*i_{p+1}} \wedge \mathbf{e}^{*i_{p+2}} \wedge \cdots \wedge \mathbf{e}^{*i_d}. \quad (2.66)$$

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

$$\begin{aligned} \star\, dx &= dydz, \\ \star\, dy &= dzdx, \\ \star\, dz &= dxdy. \end{aligned} \quad (2.67)$$

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[3]

$$\begin{aligned} \star\, dxdy &= -dzdt, \\ \star\, dydz &= -dxdt, \\ \star\, dzdx &= -dydt, \\ \star\, dxdt &= dydz, \\ \star\, dydt &= dzdx, \\ \star\, dzdt &= dxdy. \end{aligned} \quad (2.68)$$

For example, the first equation 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 $dxdt$.

[3]Misner, Thorn and Wheeler, *Gravititation*, (MTW) page 108.

The $\star$ map is constructed so that if

$$\alpha = \frac{1}{p!}\alpha_{i_1 i_2 \ldots i_p} dx^{i_1} dx^{i_2} \cdots dx^{i_p}, \tag{2.69}$$

and

$$\beta = \frac{1}{p!}\beta_{i_1 i_2 \ldots i_p} dx^{i_1} dx^{i_2} \cdots dx^{i_p}, \tag{2.70}$$

then

$$\alpha \wedge \star\beta = \beta \wedge \star\alpha = \langle \alpha, \beta \rangle \, \sigma, \tag{2.71}$$

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

$$\langle \alpha, \beta \rangle = \frac{1}{p!} g^{i_1 j_1} g^{i_2 j_2} \cdots g^{i_p j_p} \alpha_{i_1 i_2 \ldots i_p} \beta_{j_1 j_2 \ldots j_p}, \tag{2.72}$$

and σ is the *volume form*

$$\sigma = \sqrt{g}\, dx^1 dx^2 \cdots dx^d. \tag{2.73}$$

We now apply these ideas to Maxwell. From

$$F = B_x dydz + B_y dzdx + B_z dxdy + E_x dxdt + E_y dydt + E_z dzdt, \tag{2.74}$$

we get

$$\star F = -B_x dxdt - B_y dydt - B_z dzdt + E_x dydz + E_y dzdx + E_z dxdy. \tag{2.75}$$

We can check this by taking the wedge product. We find

$$F \star F = \frac{1}{2}(F_{\mu\nu} F^{\mu\nu})\sigma = (B_x^2 + B_y^2 + B_z^2 - E_x^2 - E_y^2 - E_z^2) dtdxdydz. \tag{2.76}$$

Similarly, from

$$J{=}J_\mu dx^\mu = -\rho\, dt + j_x dx + j_y dy + j_z dz, \tag{2.77}$$

we compute

$$\star J = \rho\, dxdydz - j_x dtdydz - j_y dtdzdx - j_z dtdxdy, \tag{2.78}$$

and check that

$$J \star J = (J_\mu J^\mu)\sigma = (-\rho^2 + j_x^2 + j_y^2 + j_z^2) dtdxdydz. \tag{2.79}$$

Observe that

$$d \star J = \left(\frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{j} \right) dt dx dy dz = 0, \tag{2.80}$$

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 $d^2 = 0$. MTW has some nice pictures giving the geometric interpretation of these equations.

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

$$\star \star \omega = (-1)^{p(d-p)} \omega. \tag{2.81}$$

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[4], this is the *cotangent 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 p, q setting. Here we will describe them as they appear in a modern book on Mechanics, such as Abrahams and Marsden's *Foundations of Mechanics*, or V. I. Arnold *Mathematical Methods of Classical Mechanics*.

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

$$\omega = \frac{1}{2} \omega_{ij} dx^i dx^j. \tag{2.82}$$

[4]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.

The word *closed* means that $d\omega = 0$, and non-degenerate means that if $\omega(X, Y) = 0$ for all vectors $Y \in TM_x$ for any point x, then $X = 0$ at that point (or equivalently that the matrix ω_{ij} has an inverse ω^{ij}).

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

$$dH = -i_{v_H}\omega = -\omega(v_H, \ \ ) \tag{2.83}$$

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

$$dH = \frac{\partial H}{\partial q^i}dq^i + \frac{\partial H}{\partial p^i}dp^i \tag{2.84}$$

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

$$v_H = \dot{q}^i\frac{\partial}{\partial q^i} + \dot{p}^i\frac{\partial}{\partial p^i}. \tag{2.85}$$

Now

$$\begin{aligned} i_{v_H}\omega &= dp^idq^i(\dot{q}^j\partial_{q^j} + \dot{p}^j\partial_{p^j}, \quad) \\ &= \dot{p}^idq^i - \dot{q}^idp^i, \end{aligned} \tag{2.86}$$

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

$$\dot{q}^i = \frac{\partial H}{\partial p^i}, \quad \dot{p}^i = -\frac{\partial H}{\partial q^i}. \tag{2.87}$$

Darboux' theorem says that for any point x we can always find coordinates p, q in some neigbourhood x such that $\omega = dp^1dq^1 + dp^2dq^2 + \cdots dp^ndq^n$, so 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 work fine. It may be, however, that the neigbourhood of x where the Darboux coordinates work is not the entire phase space, and we need to cover the space with overlapping p, q coordinate patches. Then, what is a p in one coordinate patch 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 coordinate-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 $\hat{A}$, $\hat{B}$:

$$[\hat{A}, \hat{B}] \quad \leftrightarrow \quad -i\hbar\{A, B\} + O\left(\hbar^2\right). \tag{2.88}$$

We define the Poisson bracket by

$$\{H_1, H_2\} \stackrel{\text{def}}{=} \left.\frac{dH_2}{dt}\right|_{H_1} = v_{H_1} H_2. \tag{2.89}$$

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

$$\{H_1, H_2\} = \omega(v_{H_1}.v_{H_2}). \tag{2.90}$$

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

Since

$$v_{H_1}(H_2 H_3) = (v_{H_1} H_2) H_3 + H_2(v_{H_1} H_3), \tag{2.91}$$

the Poisson bracket is a derivation:

$$\{H_1, H_2 H_3\} = \{H_1, H_2\} H_3 + H_2\{H_1, H_3\}. \tag{2.92}$$

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*:

$$\{\{H_1, H_2\}, H_3\} + \{\{H_2, H_3\}, H_1\} + \{\{H_3, H_1\}, H_2\} = 0. \tag{2.93}$$

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

$$\begin{aligned} d\omega(v_{H_1}, v_{H_2}, v_{H_3}) \;=\; & v_{H_1}\omega(v_{H_2}, v_{H_3}) + v_{H_2}\omega(v_{H_3}, v_{H_1}) + v_{H_3}\omega(v_{H_1}, v_{H_2}) \\ & -\omega([v_{H_1}, v_{H_2}], v_{H_3}) - \omega([v_{H_2}, v_{H_3}], v_{H_1}) - \omega([v_{H_3}, v_{H_1}], v_{H_2}). \end{aligned} \tag{2.94}$$

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

$$v_{H_1}\omega(v_{H_2}, v_{H_3}) = v_{H_1}\{H_2, H_3\} = \{H_1, \{H_2, H_3\}\}. \tag{2.95}$$

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

$$\begin{aligned}
\omega([v_{H_1}, v_{H_2}], v_{H_3}) &= -\omega(v_{H_3}, [v_{H_1}, v_{H_2}]) \\
&= dH_3([v_{H_1}, v_{H_2}]) \\
&= [v_{H_1}, v_{H_2}]H_3 \\
&= v_{H_1}(v_{H_2}H_3) - v_{H_2}(v_{H_1}H_3) \\
&= \{H_1, \{H_2, H_3\}\} - \{H_2, \{H_1, H_3\}\} \\
&= \{H_1, \{H_2, H_3\}\} + \{H_2, \{H_3, H_1\}\}. \qquad (2.96)
\end{aligned}$$

Adding everything togther now shows that

$$\begin{aligned}
0 &= d\omega(v_{H_1}, v_{H_2}, v_{H_3}) \\
&= -\{\{H_1, H_2\}, H_3\} - \{\{H_2, H_3\}, H_1\} - \{\{H_3, H_1\}, H_2\}. \qquad (2.97)
\end{aligned}$$

If we rearrange the Jacobi identity as

$$\{H_1, \{H_2, H_3\}\} - \{H_2, \{H_1, H_3\}\} = \{\{H_1, H_2\}, H_3\}, \qquad (2.98)$$

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 map $H \to v_H$ is not one-to-one, however. Constant functions map to the zero vector field.

We also observe that $\mathcal{L}_{v_H}\omega = 0$, where v_H is the vector field corresponding to H. This last result is *Liouville's theorem* on the conservation of phase-space volume.

The classical mechanics of spin

It is often 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" $\mathbf{S}$ can be regarded

as a fixed length vector that can point in any direction in $\mathbf{R}^3$. We will take it to be of unit length so that its components are

$$\begin{aligned} S_x &= \sin\theta\cos\phi \\ S_y &= \sin\theta\sin\phi \\ S_z &= \cos\theta, \end{aligned} \tag{2.99}$$

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 Nils 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 unnatural in the traditional view of mechanics, but it fits comfortably into modern the symplectic picture.

We want to treat all the points on the sphere alike, and so the natural symplectic 2-form to consider is the element of area $\omega = \sin\theta d\theta d\phi$. 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 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

$$d(BS_x) = B(\cos\theta\cos\phi d\theta - \sin\theta\sin\phi d\phi). \tag{2.100}$$

This is to be set equal to

$$-\omega(v_{BS_x}, \ \ ) = v^\theta(-\sin\theta)d\phi + v^\phi\sin\theta d\theta. \tag{2.101}$$

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

$$v_{(BS_x)} = v^\theta\partial_\theta + v^\phi\partial_\phi = B(\sin\phi\partial_\theta + \cos\phi\cot\theta\partial_\phi). \tag{2.102}$$

This velocity field 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

$$\begin{aligned} v_{S_y} &= -\cos\phi\partial_\theta + \sin\phi\cot\theta\partial_\phi \\ v_{S_z} &= -\partial_\phi. \end{aligned} \tag{2.103}$$

From the 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

$$\begin{aligned} \{S_x, S_y\} &= S_z, \\ \{S_y, S_z\} &= S_x, \\ \{S_z, S_x\} &= S_y. \end{aligned} \tag{2.104}$$

These Poisson brackets for our classical "spin" are to be compared with the commutator 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

Although covariant derivatives are an important topic in physics, this section is outside the main stream of our development and may be omitted at first reading.

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 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

$$\nabla_X f = Xf. \tag{2.105}$$

ii) Its action a basis set of vector fields $\mathbf{e}_a(x)$ (a local frame, or *vielbein*[5]) by introducing a set of functions $\omega^i{}_{jk}(x)$ and setting

$$\nabla_{\mathbf{e}_k}\mathbf{e}_j = \omega^i{}_{jk}\mathbf{e}_i. \tag{2.106}$$

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

The set of functions $\omega^i{}_{jk}(x)$ is called the *connection*. We can choose them at will. Different choices define different covariant derivatives. **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.

If we may take as our basis vectors the co-ordinate vectors $\mathbf{e}_\mu \equiv \partial_\mu$. Then we usually use Γ instead of ω and set

$$\nabla_\mu \mathbf{e}_\nu \equiv \nabla_{\mathbf{e}_\mu}\mathbf{e}_\nu = \Gamma^\lambda_{\mu\nu}\mathbf{e}_\lambda. \tag{2.107}$$

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

Two important quantities which *are* tensors, are associated with ∇_X:

i) The *torsion*

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \tag{2.108}$$

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).

ii) The *Riemann curvature tensor*

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_Z Z - \nabla_{[X,Y]}Z. \tag{2.109}$$

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).

If we require that $T = 0$ and $\nabla_\mu \mathbf{g} = 0$, the connection is uniquely determined, and is called the *Riemann connection*. This is the connection that appears in General relativity.

2.5.2 Cartan's Viewpoint: Local Frames

Let $\mathbf{e}^{*j}(x)$ be the dual basis to the $\mathbf{e}_i(x)$. Introduce the matrix-valued connection one-forms ω with entries $\omega^i{}_j = \omega^i{}_{j\mu}dx^\mu$. In terms of these

$$\nabla_X \mathbf{e}_j = \mathbf{e}_i\omega^i{}_j(X). \tag{2.110}$$

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

We also regard T and R as vector and matrix valued 2-forms

$$T^i = \frac{1}{2} T^i{}_{\mu\nu} dx^\mu dx^\nu, \tag{2.111}$$

$$R^i{}_k = \frac{1}{2} R^i{}_{k\mu\nu} dx^\mu dx^\nu. \tag{2.112}$$

Then we have Cartan's structure equations:

$$d\mathbf{e}^{*i} + \omega^i{}_j \wedge \mathbf{e}^{*j} = T^i \tag{2.113}$$

and

$$d\omega^i{}_k + \omega^i{}_j \wedge \omega^j{}_k = R^i{}_k. \tag{2.114}$$

The last can be written more compactly as

$$d\omega + \omega \wedge \omega = \mathbf{R}, \tag{2.115}$$

where ω and $\mathbf{R}$ are matrices acting on the tangent space.

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

$$I_1 = \int_\Gamma df. \tag{3.1}$$

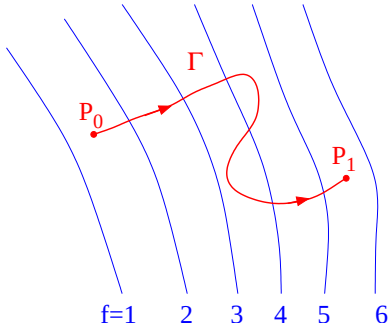
Here Γ is a path in our space starting at some point P_0 and ending at the point P_1. Any reasonable definition of I_1 should end up with the answer we would immediately write down if we saw an expression like I_1 in an elementary calculus class. That is,

$$I_1 = \int_\Gamma df = f(P_1) - f(P_0). \tag{3.2}$$

We will therefore accept this.

Notice that no notion of metric was needed. 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. In the figure $\int_\Gamma df = 5.5 - 1.5 = 4$.



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

$$I_1 = \int_0^1 \left(\frac{df}{ds}\right) ds \tag{3.3}$$

and it is important that we did not have $\left|\frac{df}{ds}\right|$ in this expression. 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

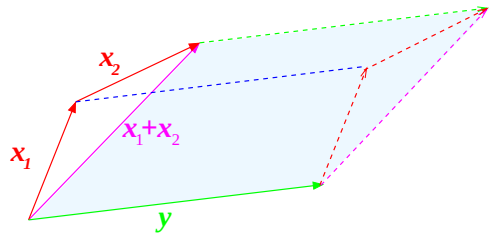
What about integrating 2 and 3-forms? Why the skew-symmetry? To answer these questions, think about assigning some sort of "area" in $\mathbf{R}^2$ to the parallelogram defined by the two vectors $\mathbf{x}, \mathbf{y}$. This is going to be some function of the two vectors. Let us call it $\omega(\mathbf{x}, \mathbf{y})$. 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 following drawing 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.



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 is 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 Manifolds

A manifold or surface is *orientable* if we can chose a single orientation for the entire manifold. 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 defining 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!). In this case we construct a continously varying orientated basis field $\mathbf{e}^{(i)}_\mu(x)$ for each member, labelled by (i), of an atlas of coordinate patches. We should chose the patches so the intersection of any pair forms a connected set. Assuming that this has been done,

the orientation of pair of overlapping patches 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 can be made 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*. The Möbius strip is an example of a non-orientable surface.

3.2 Integrating p-Forms

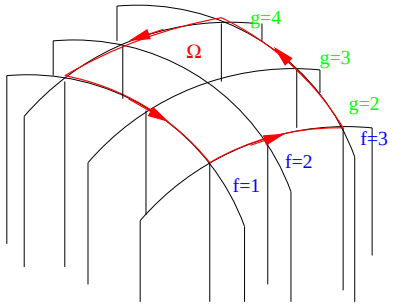
A p-form is naturally integrated over an oriented p-dimensional surface. Rather than start with an abstract definition, I will first give some examples, and hope that the general recipe will then be obvious.

3.2.1 Counting Boxes

To visualize integrating 2-forms begin with

$$\int_\Omega df\, dg, \tag{3.7}$$

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.



We count how many parallelograms (including fractions of a parallelogram) there are, counting them positive if the parallelogram given by the mesh is oriented in the same way as the surface, and negative otherwise.

[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

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 $\mathbf{R}^2$ 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.

Alternatively, we define the integral

$$I_2 = \int_\Omega \omega, \tag{3.10}$$

where ω is a 2-form, and Ω is an oriented surface by thinking about plugging vectors into ω. We tile the surface with collection of (perhaps tiny) parallelograms, each bounded by a ordered pair of vectors. We plug each of these parallelograms into the 2-form at each base point of the pair, and total the resulting numbers. We can generalize this to integrating a p-form over an oriented p-dimensional region, the orientation being determined by the orientation of each p-dimensional parallelepipeds into which the region is decomposed.

3.2.2 General Case

The previous section explained how to think about the integral. Here we explain how to actually do one.

In d=2, if we change variables $x = x(y)$ in

$$I_4 = \int_\Omega f(x) dx^1 dx^2 \tag{3.11}$$

we already know that

$$\begin{aligned} dx^1 &= \frac{\partial x^1}{\partial y^1} dy^1 + \frac{\partial x^1}{\partial y^2} dy^2, \\ dx^2 &= \frac{\partial x^2}{\partial y^1} dy^1 + \frac{\partial x^2}{\partial y^2} dy^2, \end{aligned} \tag{3.12}$$

so

$$dx^1 dx^2 = \left(\frac{\partial x^1}{\partial y^1} \frac{\partial x^2}{\partial y^2} - \frac{\partial x^2}{\partial y^1} \frac{\partial x^1}{\partial y^2} \right) dy^1 dy^2. \tag{3.13}$$

Thus

$$\int_\Omega f(x) dx^1 dx^2 = \int_{\Omega'} f(x(y)) \frac{\partial(x^1, x^2)}{\partial(y^1, y^2)} dy^1 dy^2 \tag{3.14}$$

where $\frac{\partial(x^1,y^1)}{\partial(y^1,y^2)}$ is the Jacobean, and Ω' the integration region in the new variables. This works in the same way if $2 \to p$. 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

$$\omega = \frac{1}{p!} \omega_{\mu_1 \mu_2 \ldots \mu_p} dx^{\mu_1} \cdots dx^{\mu_p} \tag{3.15}$$

over the region Ω of a p dimensional surface in a d dimensional space, substitute a paramaterization

$$\begin{aligned} x^1 &= x^1(\xi^1, \xi^2, \ldots, \xi^p), \\ &\vdots \\ x^d &= x^d(\xi^1, \xi^2, \ldots, \xi^p), \end{aligned} \tag{3.16}$$

into ω. Next, use

$$dx^\mu = \frac{\partial x^\mu}{\partial \xi^i} d\xi^i, \tag{3.17}$$

so that

$$\omega \to \omega(x(\xi))_{i_1 i_2 \ldots i_p} \frac{\partial x^{i_1}}{\partial \xi^1} \cdots \frac{\partial x^{i_p}}{\partial \xi^p} d\xi^1 \cdots d\xi^p, \tag{3.18}$$

which we regard as a p-form on Ω. (The $p!$ is absent here because we have chosen a particular order for the $d\xi$'s.) Then

$$\int_\Omega \omega = \int \omega(x(\xi))_{i_1 i_2 \ldots i_p} \frac{\partial x^{i_1}}{\partial \xi^1} \cdots \frac{\partial x^{i_p}}{\partial \xi^p} d\xi^1 \cdots d\xi^p \tag{3.19}$$

where the right hand side is an ordinary multiple integral. The result does not depend on the chosen parameterization.

Example: To integrate the 2-form $xdydz$ over the surface of a two dimensional sphere of radius R, we parameterize the surface with polar angles as

$$\begin{aligned} x &= R\sin\phi\sin\theta, \\ y &= R\cos\phi\sin\theta, \\ z &= R\cos\theta. \end{aligned} \tag{3.20}$$

Then

$$\begin{aligned} dy &= -R\sin\phi\sin\theta d\phi + R\cos\phi\cos\theta d\theta, \\ dz &= -R\sin\theta d\theta, \end{aligned} \tag{3.21}$$

and so

$$xdydz = R^3\sin^2\phi\sin^3\theta d\phi d\theta. \tag{3.22}$$

We therefore evaluate

$$\begin{aligned} I &= R^3\int_0^{2\pi}\int_0^{\pi}\sin^2\phi\sin^3\theta\, d\phi d\theta \\ &= R^3\int_0^{2\pi}\sin^2\phi\, d\phi\int_0^{\pi}\sin^3\theta\, d\theta \\ &= R^3\pi\int_{-1}^{1}(1-\cos^2\theta)d\cos\theta \\ &= \frac{4}{3}\pi R^3. \end{aligned} \tag{3.23}$$

The volume form

Although we do not need any notion of volume or measure to integrate a differential form, a d-dimensional surface embedded or immersed in $\mathbf{R}^n$ does inherit a metric from the ambient space. If the Cartesian co-ordinates of a point in the surface is given by $x^a(\xi^1,\ldots,\xi^d)$, $a=1,\ldots,n$, then the *induced metric* is

$$\text{“}ds^2\text{”} \equiv \mathbf{g}(\ \ ,\ \ ) \equiv g_{\mu\nu}\, d\xi^\mu\otimes d\xi^\nu = \left(\sum_{a=1}^{n}\frac{\partial x^a}{\partial\xi^\mu}\frac{\partial x^a}{\partial\xi^\nu}\right)d\xi^\mu\otimes d\xi^\nu. \tag{3.24}$$

The *volume form* associated with the metric is

$$d(Volume) = \sqrt{g}\, d\xi^1\cdots d\xi^d, \tag{3.25}$$

where $g = \det(g_{\mu\nu})$. The integral of this over the surface gives the volume, or area, of the surface.

If we change the parameterization of the surface from ξ^μ to ζ^μ, neither the $d\xi^1 \cdots d\xi^d$ nor the $\sqrt{g}$ are separately invariant, but the Jacobean arising from the change of the d-form, $d\xi^1 \cdots dx^d \to d\zeta^1 \cdots d\zeta^d$, cancels against the factor coming from the transformation law of the metric tensor $g_{\mu\nu} \to g'_{\mu\nu}$, leading to

$$\sqrt{g}\, d\xi^1 \cdots d\xi^d = \sqrt{g'} d\zeta^1 \cdots d\zeta^d. \tag{3.26}$$

Example: The induced metric on the surface of a unit-radius two-sphere embedded in $\mathbf{R}^3$, 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

$$d(Area) = \sin\theta\, d\theta d\phi.$$

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 is 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 $\mathbf{R}^2$, then from

$$d[\frac{1}{2}(x\, dy - y\, dx)] = dxdy,$$

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 $\mathbf{R}^2$, then from $d[r^2 d\theta/2] = r\,dr d\theta$ we have

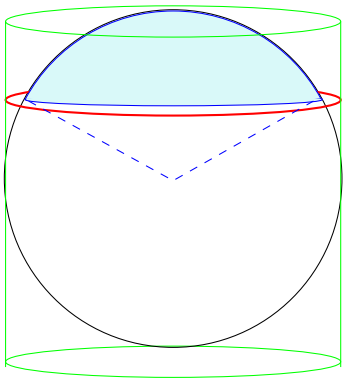
$$\text{Area}\,(\Omega) = \int_\Omega r\,dr d\theta = \frac{1}{2}\int_{\partial\Omega} r^2 d\theta.$$

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 used the example of the previous section to compute the surface integral.

Example: (Archimedes' tombstone.)



Sphere and circumscribed cylinder.

Archimedes gave instructions that his tombstone should have displayed on it a diagram consisting of a sphere and circumscribed cylinder. Cicero, while serving as quaestor 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 above, that the area of the spherical cap cut off by slicing through the cylinder is equal to the area cut off on the cylinder.

[2]Marcus Tullius Cicero, *Tusculan Disputations*, Book V, Sections 64 – 66

We can understand this result via Stokes' theorem: If the two-sphere S^2 is parameterized by polar co-ordinates θ, ϕ, and Ω is a region on 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 gives

$$\text{Area}\,(Cap) = 2\pi(1-\cos\theta)$$

which is indeed the area of the cylinder above the red circle.

Exercise: The sphere S^{n-1} can be thought of as the locus of points in $\mathbf{R}^n$ obeying $\sum_{i=1}^n (x^i)^2 = 1$. Use its invariance under orthogonal transformations to show that the element of surface "area" of the $(n-1)$-sphere can be written as

$$\text{“}d(\text{Area})\text{”} = \frac{1}{(n-1)!}\epsilon_{\alpha_1\alpha_2\ldots\alpha_n} x^{\alpha_1}\, dx^{\alpha_2}\ldots dx^{\alpha_n}.$$

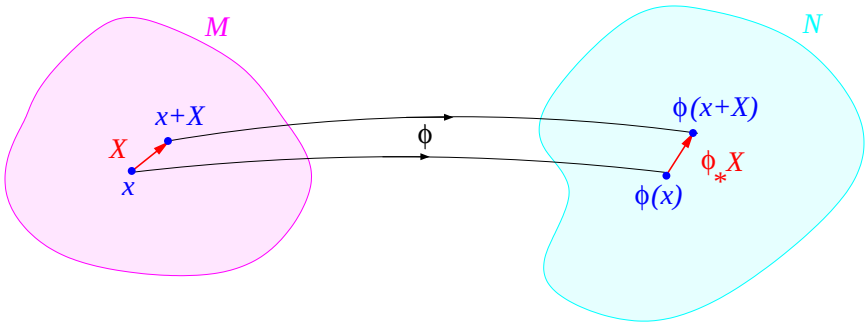
Use Stokes' theorem to relate the integral of this form over the surface of the sphere to the volume of the solid unit sphere. Confirm that we get the correct proportionality between the volume of the solid unit sphere and the "area" of its surface.

3.4 Applications

We now know how to integrate forms. What sort of forms should we seek to integrate? We will now explain that 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 $\phi_* : TM_x \to TN_{\phi(x)}$.



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

$$(\phi_* X)^\mu = \frac{\partial \xi^\mu}{\partial x^\nu} X^\nu. \tag{3.27}$$

This looks very like the transformation formula for contravariant vector components under a change of coordinate system. What we are doing is conceptually different, however. A change of co-ordinates produces a *passive* transformation — *i.e.* a new description for an unchanging vector. What we are doing here is a *active* transformation — we are changing a vector into different one.

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. The map $\phi : M \to N$ induces a natural *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 $X_1, X_2, \ldots, X_p$ at $x \in M$ into it. This we do by pushing the X_i forward to $TN_{\phi(x)}$, plugging them into ω, and declaring the result to be the evaluation of $\phi^*\omega$ on the X_i. Symbolically

$$[\phi^*\omega](X_1, X_2, \ldots, X_p) = \omega(\phi_* X_1, \phi_* X_2, \ldots, \phi_* X_p). \tag{3.28}$$

This all seems rather abstract, but the idea is useful, and in practice quite simple: If the map takes $x \in M \to \xi(x) \in N$, and

$$\omega = \frac{1}{p!} \omega_{i_1 \ldots i_p}(\xi) d\xi^{i_1} \ldots d\xi^{i_p}, \tag{3.29}$$

then

$$\begin{aligned}\phi^*\omega &= \frac{1}{p!}\omega_{i_1 i_2 \ldots i_p}[\xi(x)]d\xi^{i_1}(x)d\xi_2^i(x)\cdots d\xi^{i_p}(x) \\ &= \frac{1}{p!}\omega_{i_1 i_2 \ldots i_p}[\xi(x)]\frac{\partial \xi^{i_1}}{\partial x^{\mu_1}}\frac{\partial \xi^{i_2}}{\partial x^{\mu_2}}\cdots\frac{\partial \xi^{i_p}}{\partial x^{\mu_1}}dx^{\mu_1}\cdots dx^{\mu_p}. \qquad (3.30)\end{aligned}$$

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 $\mathbf{n} : \mathbf{R}^2 \to S^2$ where $\mathbf{n}(x)$ is a unit vector. We can think of $\mathbf{n}$ as the direction of the magnetization field of a two-dimensional ferromagnet. In terms of $\mathbf{n}$, the area 2-form on the sphere can be written

$$\Omega = \frac{1}{2}\mathbf{n}\cdot(d\mathbf{n}\times d\mathbf{n}) \equiv \frac{1}{2}\epsilon_{ijk}n^i dn^j dn^k. \qquad (3.31)$$

The $\mathbf{n}$ map pulls this area-form back to

$$F \equiv \mathbf{n}^*\Omega = \frac{1}{2}(\epsilon_{ijk}n^i\partial_\mu n^j\partial_\nu n^k)dx^\mu dx^\nu = (\epsilon_{ijk}n^i\partial_1 n^j\partial_2 n^k)\,dx^1dx^2 \qquad (3.32)$$

which is a differential form in $\mathbf{R}^2$. We will 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 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 map $\mathbf{n}(x)$ 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 original x sphere wraps round the target $\mathbf{n}$ sphere. A mathematician would call it 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

$$N = \frac{1}{4\pi}\int_{S^2}\left\{\epsilon_{ijk}n^i\partial_1 n^j\partial_2 n^k\right\}dx^1dx^2, \qquad (3.33)$$

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 $F = \mathbf{n}^*\Omega$ 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 $\mathbf{n}$ field in higher dimensions, however, we still have $dF = 0$. This is because

$$dF = \frac{1}{2}\epsilon^{ijk}\partial_\sigma n^i \partial_\mu n^j \partial_\nu n^k dx^\sigma dx^\mu dx^\nu. \tag{3.34}$$

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 and any three such vectors are linearly dependent, so their triple-product is zero.

Although it is closed, $F = \mathbf{n}^*\Omega$ 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

$$\delta F = \mathbf{n}^*[\mathbf{n} \cdot (d\delta\mathbf{n} \times d\mathbf{n})], \tag{3.35}$$

and this variation *can* be written as a total derivative

$$\delta F = d\{\mathbf{n}^*[\mathbf{n} \cdot (\delta\mathbf{n} \times d\mathbf{n})]\} \equiv d\{\epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu\}. \tag{3.36}$$

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

$$\delta N = \int_{S^2} \delta F = \int_{\partial S^2} \epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu. \tag{3.37}$$

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. On $\mathbf{R}^2$, with $\mathbf{n}$ constant at infinity, we have similarly

$$\delta N = \int_2 \delta F = \int_\Gamma \epsilon_{ijk} n^i \delta n^j \partial_\mu n^k dx^\mu, \tag{3.38}$$

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 physical applications, the field $\mathbf{n}$ often winds in localized regions called *Skyrmions*. The winding number counts how many Skyrmions (minus the

number of anti-Skyrmions, which wind with opposite orientation) there are. An example of a smooth map with positive winding number N is

$$e^{\phi} \tan \frac{\theta}{2} = \frac{P(z)}{Q(z)}, \tag{3.39}$$

where P and Q are co-prime polynomials of degree N in $z = x_1 + ix_2$, and θ and ϕ are the polar co-ordinates specifying the direction $\mathbf{n}$. We will later show that this particular field configuration minimizes the energy integral

$$E = \frac{1}{2} \int (\partial_\mu n^i)^2 d^2x \tag{3.40}$$

for the given winding number.

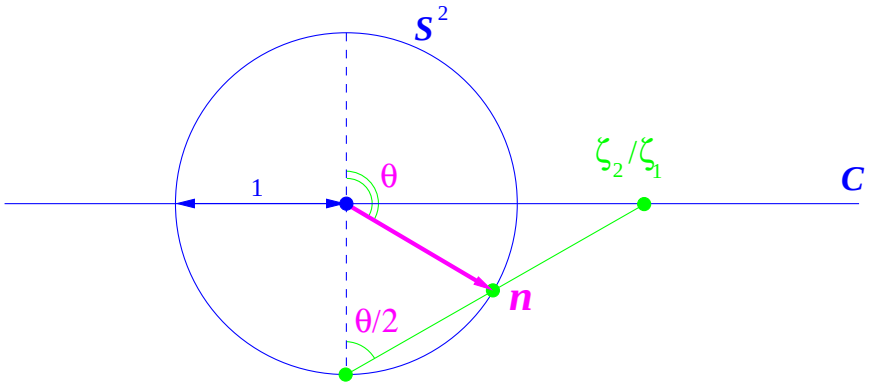
3.4.3 The Hopf Map

The complex projective space $\mathbf{CP}^n$ is defined to be the set of *rays* in a complex $n+1$ dimensional vector space. It consists of equivalence classes of complex vectors $[\zeta_1, \zeta_2, \ldots, \zeta_{n+1}]$, 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. This space is a $2n$-dimensional 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

$$\xi_1 + i\eta_1 = \frac{\zeta_1}{\zeta_{n+1}}, \quad \xi_2 + i\eta_2 = \frac{\zeta_2}{\zeta_{n+1}}, \ldots, \xi_n + i\eta_n = \frac{\zeta_n}{\zeta_{n+1}}. \tag{3.41}$$

Similar co-ordinate systems can be constructed in the regions where other ζ_n are non-zero. Every point in $\mathbf{CP}^n$ lies in at least one of these co-ordinate patches.

The complex projective space $\mathbf{CP}^1$ 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 $[\zeta_1, \zeta_2] \in \mathbf{CP}^1$ 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 $\mathbf{C}$ as being the x, y plane in $\mathbf{R}^3$. We then draw a straight line connecting the plotted point to the south pole of a unit sphere circumscribed in about the origin in $\mathbf{R}^3$. The point where this line (continued if necessary) intersects the sphere is the tip of the unit vector $\mathbf{n}$.



A slice through the unit sphere.

If ζ_2, were zero, we would end up at the north pole where $z = 1$. 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. To find an explicit formula for the map, we observe from the figure that $\zeta_2/\zeta_1 = e^{i\phi} \tan\theta/2$, and this suggests the use of the "t"-substitution formulae

$$\sin\theta = \frac{2t}{1+t^2}, \quad \cos\theta = \frac{1-t^2}{1+t^2}, \tag{3.42}$$

where $t = \tan\theta/2$. Since

$$\begin{aligned} n^1 &= \sin\theta\cos\phi, \\ n^2 &= \sin\theta\sin\phi, \\ n^3 &= \cos\theta, \end{aligned}$$

we then find that

$$n^1 + in^2 = \frac{2(\zeta_2/\zeta_1)}{1+|\zeta_2/\zeta_1|^2}, \quad n^3 = \frac{1-|\zeta_2/\zeta_1|^2}{1+|\zeta_2/\zeta_1|^2}. \tag{3.43}$$

We can multiply through by $|\zeta_1|^2 = \zeta_1^*\zeta_1$, and so write this correspondence in a more symmetrical manner:

$$n^1 = \frac{\zeta_1^*\zeta_2 + \zeta_2^*\zeta_1}{|\zeta_1|^2 + |\zeta_2|^2}$$

$$\begin{aligned} n^2 &= \frac{1}{i}\left(\frac{\zeta_1^*\zeta_2 - \zeta_2^*\zeta_1}{|\zeta_1|^2+|\zeta_2|^2}\right), \\ n^3 &= \frac{|\zeta_1|^2-|\zeta_2|^2}{|\zeta_1|^2+|\zeta_2|^2}. \end{aligned} \tag{3.44}$$

This last form can be conveniently expressed in terms of the Pauli sigma matrices:

$$\begin{aligned} n^1 &= (z_1^*, z_2^*)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \\ n^2 &= (z_1^*, z_2^*)\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \\ n^3 &= (z_1^*, z_2^*)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \end{aligned} \tag{3.45}$$

where

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \frac{1}{\sqrt{|\zeta_1|^2+|\zeta_2|^2}}\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{3.46}$$

is a normalized 2-vector, which we can think of as a *spinor*.

We see that the $\mathbf{CP}^1 \simeq S^2$ correspondence can be given a quantum mechanical interpretation: Any unit 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*

$$n^i = \psi^\dagger \hat{\sigma}^i \psi. \tag{3.47}$$

Now, since

$$1 = |z_1|^2 + |z_2|^2, \tag{3.48}$$

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*:

$$\text{Hopf} : S^3 \to S^2. \tag{3.49}$$

Since the dimension reduces from three to two, 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 $\mathbf{CP}^1$, and hence in S^2. The inverse image of a point in S^2 is a great circle in S^3. Later we will show that any two such great circles are linked and this makes the Hopf map topologically non-trivial in that it cannot be continuously deformed to the identity map.

Exercise:

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 take $\zeta = \xi + i\eta$ as defining a *stereographic co-ordinate system* on the sphere. Show that in these co-ordinates the metric is given by

$$\begin{aligned} ds^2 &\equiv d\theta \otimes d\theta + \sin^2\theta \, d\phi \otimes d\phi \\ &= \frac{2}{(1+|\zeta|^2)^2}(d\overline{\zeta} \otimes d\zeta + d\zeta \otimes d\overline{\zeta}) \\ &= \frac{4}{(1+|\xi|^2+|\eta|^2)^2}(d\xi \otimes d\xi + d\eta \otimes d\eta), \end{aligned}$$

and the area 2-form becomes

$$\begin{aligned} \Omega &\equiv \sin\theta \, d\theta \wedge d\phi \\ &= \frac{2i}{(1+|\zeta|^2)^2} \, d\zeta \wedge d\overline{\zeta} \\ &= \frac{4}{(1+|\xi|^2+|\eta|^2)^2} \, d\xi \wedge d\eta. \end{aligned} \tag{3.50}$$

3.4.4 The Hopf Linking Number

We can use the Hopf map to factor a field of unit vectors $\mathbf{n}(x)$ through the three-sphere by specifying the spinor ψ at each point, instead of the vector $\mathbf{n}$, and so mapping indirectly $x \to \psi \equiv (z_1, z_2)^T \to \mathbf{n}$. It might seem that for a given spin-field $\mathbf{n}(x)$ we can choose the overall phase of $\psi(x)$ 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)

$$(\epsilon_{ijk} n^i \partial_1 n^j \partial_2 n^k) \, dx^1 dx^2 = \frac{2}{i} \sum_{i=1}^{2} (\partial_1 \overline{z}_i \partial_2 z_i - \partial_2 \overline{z}_i \partial_1 z_i) \, dx^1 dx^2, \tag{3.51}$$

and so the topological charge N is given by

$$N = \frac{1}{2\pi i} \int \sum_{i=1}^{2} (\partial_1 \overline{z}_i \partial_2 z_i - \partial_2 \overline{z}_i \partial_1 z_i) \, dx^1 dx^2. \tag{3.52}$$

Since $\mathbf{n}$ 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 $|c_1|^2 + |c_2|^2 = 1$. Now, when written in terms of the z_i variables, the form F becomes a total derivative:

$$\begin{aligned} F &= \frac{2}{i}\sum_{i=1}^{2}\left(\partial_1\overline{z}_i\partial_2 z_i - \partial_2\overline{z}_i\partial_1 z_i\right)dx^1dx^2 \\ &= d\left\{\frac{1}{i}\sum_{i=1}^{2}\left(\overline{z}_i\partial_\mu z_i - (\partial_\mu\overline{z}_i)z_i\right)dx^\mu\right\}. \end{aligned} \tag{3.53}$$

Using Stokes' theorem and observing that, near infinity, we have

$$\frac{1}{2i}\sum_{i=1}^{2}\left(\overline{z}_i\partial_\mu z_i - (\partial_\mu\overline{z}_i)z_i\right) = (|c_1|^2 + |c_2|^2)d\theta = d\theta, \tag{3.54}$$

we find that

$$N = \frac{1}{2\pi i}\int_\Gamma \frac{1}{2}\sum_{i=1}^{2}\left(\overline{z}_i\partial_\mu z_i - (\partial_\mu\overline{z}_i)z_i\right)dx^\mu = \frac{1}{2\pi}\int_\Gamma d\theta, \tag{3.55}$$

where, 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.

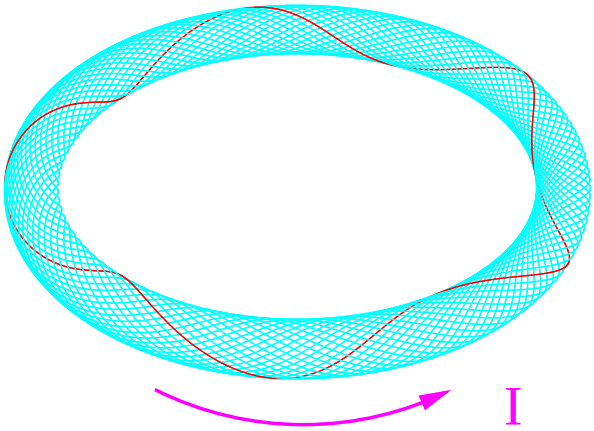
We have uncovered the rather surpring fact that the topological charge of the map $\mathbf{n} : S^2 \to S^2$ is equal to the winding number of the phase angle θ at infinity. This is the topological constraint refered to earlier. As a byproduct, we have confirmed our conjecture that the topological charge N is an integer. The existence of this integer invariant shows that the smooth maps $\mathbf{n} : S^2 \to S^2$ 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

$$\pi_2(S^2) = \mathbf{Z}. \tag{3.56}$$

We will now show that maps $\mathbf{n} : S^3 \to S^2$ also have an associated topological number. Provided that $\mathbf{n}$ tends to a constant direction at infinity so that we can think of $\mathbf{R}^3 \cup \infty$ as being S^3, this number will label the homotopy classes of fields of unit vectors $\mathbf{n}$ in *three* dimensions. If we think of the third dimension as time, a natural set of $\mathbf{n}$ fields to consider are the $\mathbf{n}(x,t)$ corresponding to the world-lines of moving Skyrmions. These 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.

We begin with an amusing problem from magnetostatics. Suppose we are given a cable originally made up of a bundle of many parallel wires. The cable was then twisted N times about its axis and then bent into a closed loop, the end of each individual wire being attached to its begining to make a continuous circuit. A total 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 N knowing only this field?



A twisted cable with $N = 5$

The answer is yes. We use Ampere's law in integral form,

$$\oint_\Gamma \mathbf{B} \cdot d\mathbf{r} = (\text{current encircled by } \Gamma). \tag{3.57}$$

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

$$\sum_{\text{wires } i} \delta I_i \oint \mathbf{B} \cdot d\mathbf{r}_i = \int \mathbf{B} \cdot \mathbf{J}\, d^3x = \int \mathbf{B} \cdot (\nabla \times \mathbf{B})\, d^3x = NI^2. \tag{3.58}$$

We now apply this 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*

$$J^\sigma = \frac{1}{2}\epsilon^{\sigma\mu\nu}\epsilon_{ijk} n^i \partial_\mu n^j \partial_\nu n^k. \tag{3.59}$$

We note that $\nabla \cdot \mathbf{J} = 0$. This is simply another way of saying that the 2-form $F = \mathbf{n}^*\Omega$ is closed.

The flux of $\mathbf{J}$ through a surface S is

$$\int_S \mathbf{J} \cdot d\mathbf{S} = \int_S F \tag{3.60}$$

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 area of the 2-sphere. The Skyrmion world-line will play the role of the cable, and the inverse images 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

$$N_{\text{Hopf}} = \frac{1}{I^2}\int_{S^3} \mathbf{B} \cdot \mathbf{J}\, d^3x = \frac{1}{16\pi^2}\int_{S^3} AF \tag{3.61}$$

will be an integer. This integer is the *Hopf linking number*, 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 R^3 of the forms

$$\begin{aligned} \mathcal{F} &= \frac{1}{i}\sum_{i=1}^{2} \left(d\overline{z}_i dz_i - dz_i d\overline{z}_i\right), \\ \mathcal{A} &= \frac{1}{i}\sum_{i=1}^{2} \left(\overline{z}_i dz_i - z_i d\overline{z}_i\right), \end{aligned} \tag{3.62}$$

respectively. If we substitute $z_{1,2} = \xi_{1,2} + i\eta_{1,2}$, we find that

$$\mathcal{A}F = 8(\xi_1 d\eta_1 d\xi_2 d\eta_2 - \eta_1 d\eta_1 d\xi_2 d\eta_2 + \xi_2 d\eta_2 d\xi_1 d\eta_1 - \eta_2 d\xi_2 d\xi_1 d\eta_1). \quad (3.63)$$

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 $x \to \psi \equiv (z_1, z_2)^T \to \mathbf{n}$ we have that

$$N_{\text{Hopf}} = \frac{1}{16\pi^2}\int_{S^3} AF = \frac{1}{2\pi^2}\int_{S^3} \psi^* d(\text{Volume on } S^3), \quad (3.64)$$

is the number of times the normalized spinor covers S^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

$$\pi_3(S^2) = \mathbf{Z}. \quad (3.65)$$

This result, implying that there are many maps from the three-sphere to the two-sphere that are not smoothly deformable to the constant map, was an great surprise when Hopf discovered it.

One of the principal physics consequences of the existence of the Hopf number is that "quantum lump" quasi-particles like the Skyrmion can be fermions, even though they are described by commuting variables. To understand how this can be, we first explain that the homotopy classes $\pi_n(M)$ are not just *sets*, they have the additional structure of being a *group*. We can compose two homotopy classes to get a third, and each homotopy class has an inverse. To define the group composition law, we think of S^n as 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. While this group structure appears to depend on the fixed point

m_0, but as long as M is arcwise connected, the groups obtained from different m_0's may be shown to be *isomorphic*, or equivalent. In the case of $\pi_2(S^2) = \mathbf{Z}$ and $\pi_3(S^2) = \mathbf{Z}$, the composition law is simply the addition of the integers $N \in \mathbf{Z}$ that label the classes.

When we quantize using Feynman's "sum over histories" path integral, we may multiply the contributions of histories that are not deformable into one another by different phase factors. These phases must must be compatable with the composition of histories by concatenating one after the other – essentially the same operation as composing homotopy classes. This means that the product of the phases for two possible histories must be the phase 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 $\exp(i\pi N_{\text{Hopf}})$ to a history. The rotation of a single Skyrmion through 2π then leads to the wavefunction changing sign. Furthermore, a history where two Skyrmions 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 two Skyrmions therefore changes sign when they are interchanged. This means that the quantized Skyrmion is a fermion.

Chapter 4

Topology of Manifolds

In this chapter we will move from considering *local* properties and consider *global* ones. Our aim is understand and characterize the large-scale connectedness of manifolds. In this chapter we will learn the language of *homology* and *cohomology*, topics which form an important part of the discipline of *algebraic topology*.

4.1 A Topological Miscellany.

Suppose we try to construct a field of unit vectors tangent to the sphere S^2. However you try to do this you will end up in trouble somewhere: you cannot comb a hairy ball. If we try this on the torus, T^2, you will have no problems: you can comb a hairy doughnut!

One way of visualizing a torus without thinking of it as the surface of a doughnut it to remember the old video game *Asteroids*. You could select periodic boundary conditions so that your spaceship would leave of the right-hand side of the screen and instantly re-appear on the left. Suppose we modify the game code so that we now re-appear at the point *diametrically opposite* the point we left. This does not seem like a drastic change until you play a game with a left-hand-drive (US) spaceship. If you take the spaceship off the screen and watch as each point in the ship reappears on the corresponding opposite point, you will observe the ship transmogrify into right-hand-drive (British) craft. If we ourselves made such an excursion, we would end up starving to death because all our left-handed amino acids would have been converted to right-handed ones. The manifold we have constructed

is called the *real projective plane*, and denoted by RP^2. The lack of a global notion of being left or right-handed means it is *non-orientable*, rather like a Möbius strip.

Now consider a *three-dimensional* region with diametrically opposite points identified. What would happen to an aircraft flying through the surface of the region? Would it change handedness, turn inside out, or simply turn upside down?

The effects described in the previous paragraphs all relate to the overall topology of our manifold. These global issues might seem a trifle *recherché* — but they can have practical consequences even for condensed-matter physics. The director field of nematic liquid crystal lives in RP^2, and the global topology of this space influences both the visual appearance of the liquid as well the character of the nematic-isotropic phase transition.

Homeomorphism and Diffeomorphism

The homology and cohomology groups we will study in this chapter are examples of *topological invariants*, quantities that are unaffected by deformations of a manifold that preserve its global topology. They therefore help to distinguish topologically distinct manifolds. If two spaces have different homology then they are certainly distinct. If, however, they have the same homology, we cannot be sure that they are topologically identical. It is somewhat of 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 spaces 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 addition 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 apparantly subtle distinction between homeomorphism and diffeomorphism has consequences for physics. Ed Witten discovered that there are 992 exotic 11-spheres. These are manifolds that are homeomorphic to the 11-sphere, but diffeomorphically inequivalent. This fact is crucial for

the cancellation of global graviational anomalies in the $E_8 \times E_8$ or $SO(32)$ symmetric superstring theories.

4.2 Cohomology

In this section we answer the questions "when can a vector field whose curl vanishes be written as the gradient of something?", and "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 topology of the space the fields inhabit.

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

$$\phi(x) = \int_{x_0}^{x} \mathbf{v} \cdot d\mathbf{r}, \tag{4.1}$$

but probably do not know the corresponding formula for $\mathbf{w}$. Using differential forms, and provided the space in which they 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" refered 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 exists a map φ_t which depends continuously on a parameter $t \in [0,1]$ and for which $\varphi_1(x) = x$ and $\varphi_0(x) = O$. Applied 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 which corresponds to the co-ordinate flow:

$$\frac{dx^\mu}{dt} = \eta^\mu(x,t). \tag{4.2}$$

An easy exercise shows that

$$\frac{d}{dt}\left(\varphi_t^*\omega\right) = \mathcal{L}_\eta(\varphi_t^*\omega). \tag{4.3}$$

We now use the infinitesimal homotopy relation and our assumption that $d\omega = 0$, and hence[1] $d(\varphi_t^*\omega) = 0$, 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 $\mathbf{R}^3$. 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^* gives

$$\varphi_t^*\omega = A(tx,ty,tz)d(ty)d(tz) + (\text{two similar terms}). \tag{4.8}$$

[1]The map φ_t^*, being essentially a change of co-ordinates, commutes with invariant operations such as "d" and "$\mathcal{L}_\eta$".

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 1-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.

We cannot eradicate the condition that Ω be retractable. It is necessary even for $\phi(x) = \int^x \mathbf{v} \cdot d\mathbf{r}$. If we define $\mathbf{v}$ on an annulus $\Omega = \{R_0 < |\mathbf{r}| < R_1\}$, and $\oint_0^{2\pi} \mathbf{v} \cdot d\mathbf{r} \neq 0$, for some closed path wrapping around the annulus, there can be no single-valued ϕ such that $\mathbf{v} = \nabla\phi$. If there were then

$$\oint_\Gamma \mathbf{v} \cdot d\mathbf{r} = \phi(0) - \phi(0) = 0. \tag{4.15}$$

A non-zero value for $\oint_\Gamma \mathbf{v} \cdot d\mathbf{r}$ therefore consititutes an *obstruction* to the existance of an η such that $\mathbf{v} = \nabla\phi$.
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 $\sin\theta d\theta d\phi = d\eta$, then Stokes says that

$$\int_{S^2} \sin\theta d\theta d\phi = \int_{\partial S^2} \eta = 0, \tag{4.19}$$

since $\partial S^2 = 0$. Again a non-zero value for $\int \omega$ over some boundaryless region provides an obstruction to finding an η such that $\omega = d\eta$.

4.2.2 De Rham Cohomology

The question of when $d\omega = 0$ implies that $\omega = d\eta$ is one example of a *cohomolgy* theory. It is known as *de Rham* cohomology after the Swiss mathematician Georges de Rham who did the most to create it.

Given a compact manifold M without boundary consider the space $\Omega^p(M) = \bigwedge^p(T^*M)$ of p-form fields. This is a vector space: we can add p-form fields and multiply them by real constants, but, like the vector space of functions

on M, it is infinite dimensional. The subspace $Z^p(M)$ of *closed* forms, those with $d\omega = 0$, is also infinite dimensional, as is the space $B^p(M)$ of *exact* forms, those that can be written as $\omega = d\eta$ for some globally defined $(p-1)$-form η. Now consider the space $H^p = Z^p/B^p$, which is the space of closed forms *modulo* exact forms. In this space we identify[2] two forms, ω_1 and ω_2, whenever there an η, such that $\omega_1 = \omega_2 + d\eta$. We say that ω_1 and ω_2 are *cohomologous*. Remarkably, for our compact manifold M the space $H^p(M)$ is *finite* dimensional. It is called the p-th (de Rham) cohomology space of the manifold. It depends only on the global topology of M, not on any metric properties. Sometimes we write $H^p_{DR}(M, \mathbf{R})$ to make it clear that we are treating it as a vector space over the real numbers. This is because there is also a space $H^p_{DR}(M, \mathbf{Z})$, where we only allow multiplications by integers.

Cohomology codifies all potential obstructions to solving the problem of finding η such that $d\eta = \omega$: we can find such an η if, and only if, ω is cohomologous to zero.

4.3 Homology

How can we find the cohomolgy spaces of a manifold, and how do we tell if a particular form we are interested in is cohomologous to zero? The most intuitive method is to construct the vector spaces *dual* to the cohomology as these spaces are easy to understand pictorially.

Given a region of space 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 geometric "regions" and define two "regions" as being *homologous* if they differ by the boundary of another "region." We will first make these vague notions precise, and then we will explain how the resulting homology spaces become the duals of de Rham cohomology spaces.

4.3.1 Chains, Cycles and Boundaries

The set of all curves and surfaces in M is infinite dimensional, but the homology spaces we are seeking are finite dimensional. We can make our com-

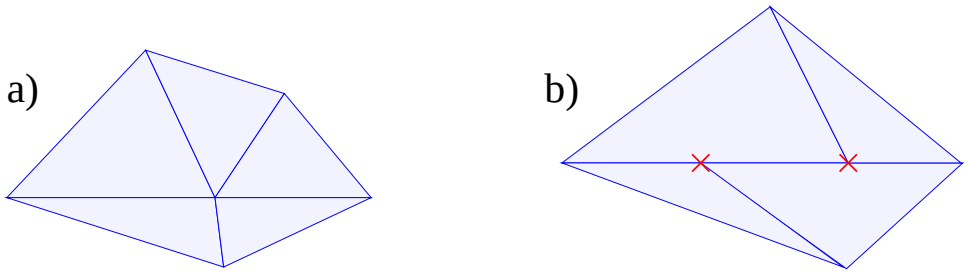
[2]Regard as being the same.

putations easier if we work with finite dimensional spaces throughout. To do this we *triangulate* M.

Simplicial Complexes

We dissect our space 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 this dissection are:

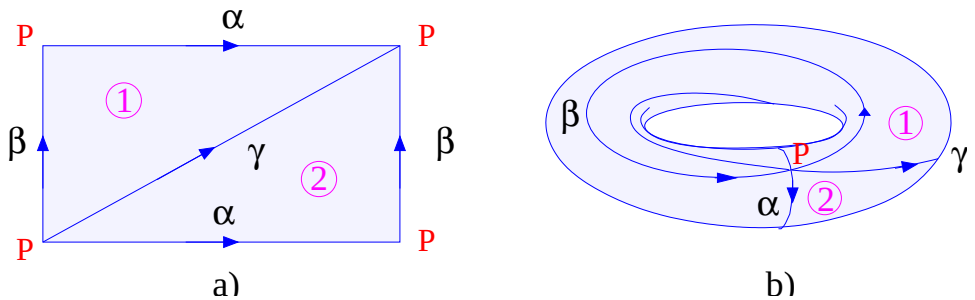
a) Every point must belong to at least one simplex.
b) A point can belong to only a finite number of simplices.
c) Two different simplices either have no points in common, or
   i) one is a face (or edge, or vertex) of the other,
   ii) the set of points in common is the whole of a shared face (or edge, or vertex) edge.



Triangles, or 2-simplices, that are a) allowed, b) not allowed in a dissection. In b) only parts of edges are in common.

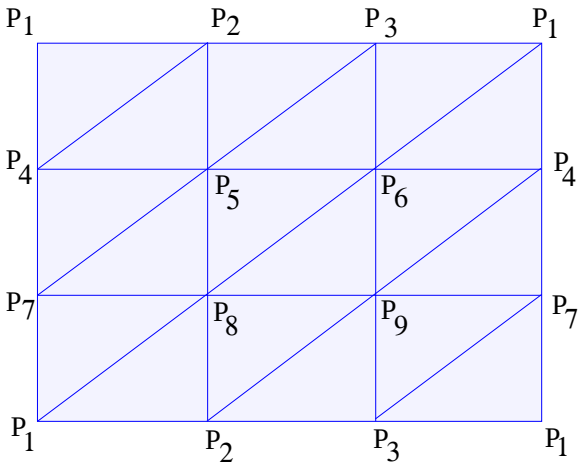
The collection of simplices composing the dissected space is called a *simplicial complex*. We will denote it by S.

Effectively we are replacing our continuous manifold by a discrete triangular lattice, but doing so in such a way as to preserve the global topological properties the space. We often do not require many triangles to do this. For example the 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.



A triangulation of the 2-Torus. Figure a) shows the torus as a rectangle with periodic boundary conditions. The two edges labled α will be glued togther point-by-point when 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. Once we have done this, all four points labeled by P are in the same place, and correspond to the single point P in figure b).

If we want each simplex in the decomposition to be uniquely specified by its vertices, we need a finer dissection. We can, for example, decompose the torus into 18 triangles each of which is uniquely labeled by three points drawn from a set of nine vertices. The resulting simplicial complex then has 27 edges:

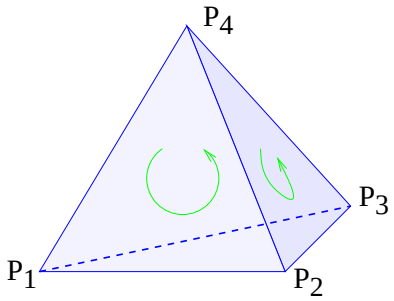


A second triangulation of the 2-Torus.

Again, points with identical labels are to be regarded as the same point, 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.

We may triangulate the sphere, S^2, as a tetrahedron $P_1P_2P_3P_4$.



A tetrahedral triangulation of the 2-sphere. The circulating arrows on the faces indicate the choice of orientation $P_1P_2P_4$ and $P_2P_3P_4$.

Chains

We assign to simplices an orientation defined by the order in which we write their points. The interchange of of any pair of points reverses the orientation, and we assign 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, $C_p(S, \mathbf{R})$, of *p-chains* which have the p-simplices as their basis vectors. The most general elements of $C_2(S, \mathbf{R})$, with S being the tetrahedral triangulation of the sphere S^2, would be

$$c = a_1P_2P_3P_4 + a_2P_1P_3P_4 + a_3P_1P_2P_4 + a_4P_1P_2P_3, \tag{4.20}$$

where $a_1, \ldots a_4$, are real numbers. We regard the distinct faces as being linearly independent basis elements for $C_2(S, \mathbf{R})$. The space is therefore four dimensional. If we had triangulated the sphere with so that its had 16 trianglular faces, the space C_2 would be 16 dimensional.

Similarly, the general element of $C_1(S, \mathbf{R})$ would be

$$c = b_1P_1P_2 + b_2P_1P_3 + b_3P_1P_4 + b_4P_2P_3 + b_5P_2P_4 + b_6P_3P_4, \tag{4.21}$$

and so $C_1(S, \mathbf{R})$ is a six dimensional space spanned by the *edges* of the tetrahedron. For $C_0(S, \mathbf{R})$ we have

$$c = c_1P_1 + c_2P_2 + c_3P_3 + c_4P_4, \tag{4.22}$$

and so $C_0(S, \mathbf{R})$ is four dimensional, and spanned by the *vertices*.

Since our manifold comprises only the *surface* of the two-sphere, there is no such thing as $C_3(S, \mathbf{R})$.

The reason for making the field $\mathbf{R}$ 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 $C_p(S, \mathbf{Z})$. Because a vector space requires that coefficients be drawn from a field, these objects are not vector spaces. They can be thought of as either *modules*—"vector spaces" whose coefficient are drawn from a ring—or as additive groups.

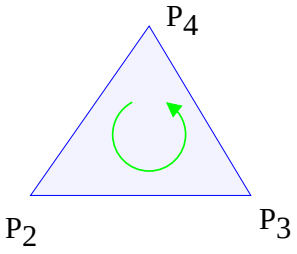
The Boundary Operator

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

$$\partial_p P_{i_1} P_{i_2} \dots P_{i_{p+1}} = \sum_{j=1}^{p+1} (-1)^{j+1} P_{i_1} \dots \hat{P_{i_j}} \dots P_{i_{p+1}}, \tag{4.23}$$

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

$$\partial_2 (P_2 P_3 P_4) = P_3 P_4 - P_2 P_4 + P_2 P_3, \tag{4.24}$$



The oriented triangle $P_2 P_3 P_4$ has boundary $P_3 P_4 + P_4 P_2 + P_2 P_3$.

The boundary of a line segment is the difference of its endpoints

$$\partial_1 (P_1 P_2) = P_2 - P_1. \tag{4.25}$$

Finally, for any point,

$$\partial P_i = 0. \tag{4.26}$$

On a p-chain $c = a_1 s_1 + a_2 s_2 + \cdots + c_n s_n$, where the s_i are p- simplices, we have $\partial c = a_1 \partial s_1 + a_2 \partial s_2 + \cdots + a_n \partial s_n$

For each of the examples we find that $\partial_{p-1}\partial_p s = 0$, and a little effort shows that this is true for any p-simplex. Since chains are sums of simplices and ∂_p is linear, this holds for any $c \in C_p$. Thus $\partial_{p-1}\partial_p = 0$. W will usually abbreviate this as $\partial^2 = 0$.

Any infinite sequence of spaces (vector spaces, modules, groups, *etc.*) $\ldots, C_{-2}, C_{-1}, C_0, C_1, C_2 \ldots$, together with maps $\partial_p C_p \to C_{p-1}$

$$\ldots \stackrel{\partial_{p+1}}{\to} C_p \stackrel{\partial_p}{\to} C_{p-1} \stackrel{\partial_{p-1}}{\to} C_{p-2} \stackrel{\partial_{p-1}}{\to} \ldots, \tag{4.27}$$

such that $\partial_{p-1}\partial_p = 0$, is called a *chain complex*. The finite sequence of C_p's we constructed from our simplicial complex is a chain complex where C_p is zero-dimensional for $p < 0$ or $p > d$.

Cycles, Boundaries and Homology

We next define two important linear subspaces of C_p. The first, the space Z_p of *p-cycles*, consists of those $z \in C_p$ such that $\partial_p z = 0$. The second, the space of *p-boundaries*, B_p, consists of those $b \in C_p$ such that $b = \partial_{p+1} c$ for some $c \in C_{p+1}$. Since $\partial^2 = 0$, the boundaries B_p constitute a subspace of Z_p.

We now form the 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, $z_2 = z_1 + \partial c$. The space $H_p(S)$, or more accurately, $H_p(S, \mathbf{R})$, is called the p-th (simplicial) *homology space* of S. (It becomes the p-th homology *group*, if $\mathbf{R}$ is replaced by the integers).

The remarkable thing is that while the spaces C_p, Z_p, and B_p, depend on the details of how the manifold M has been dissected to form the simplicial complex S, the homology space H_p is independent the dissection. This is neither obvious, nor easy to prove. We will rely on examples to at least make it plausible. Granted this independence, we will write $H_p(M)$, or $H_p(M, \mathbf{R})$, instead of $H_p(S)$ so as to make it clear that H_p is a property of M. The dimension of $H_p(M)$ is called the p-th *Betti number* of the manifold.

Example: The Two-Sphere. For the tetrahedral dissection of the two-sphere, any point is homologous to any other since $P_i - P_j = \partial(P_j P_i)$ and all $P_j P_i$ belong to C_2. Further $\partial P_i = 0$, 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 $H_0(S^2) = \mathbf{R}$, regarding $\mathbf{R}$ 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 they 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

$$\begin{aligned} z_3 &= P_1P_4 + P_4P_3 + P_3P_1, \\ z_4 &= P_1P_3 + P_3P_2 + P_2P_1, \end{aligned}$$

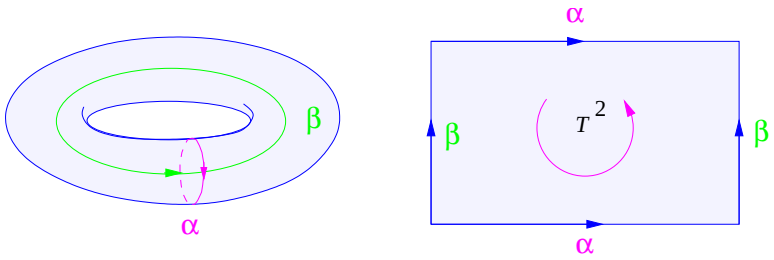
and that these cycles are linearly independent. Since everything is a boundary, we have $H_1(S^2) = \{\mathbf{0}\}$.

We also see that $H_2(S^2) = \mathbf{R}$. In the latter case the basis element is

$$P_2P_3P_4 - P_1P_3P_4 + P_1P_2P_4 - P_1P_2P_3 \tag{4.28}$$

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 have $H_0(T^2) = \mathbf{R}$, $H_1(T^2) = \mathbf{R}^2 \equiv \mathbf{R} \oplus \mathbf{R}$, and $H_2(T^2) = \mathbf{R}$. The basis elements of the two dimensional $H_1(T^2)$ are the 1-cycles α, β running round the torus.



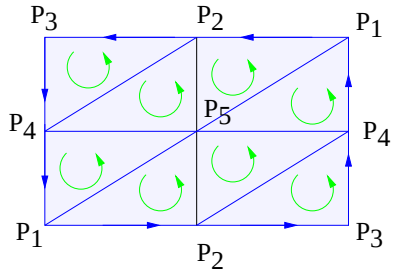
The cycle γ is homologous to $\alpha + \beta$. In terms of the second triangulation of the torus we would have

$$\begin{aligned} \alpha &= P_1P_2 + P_2P_3 + P_3P_1 \\ \beta &= P_1P_7 + P_7P_4 + P_4P_1 \end{aligned} \tag{4.29}$$

and

$$\begin{aligned}\gamma &= P_1P_8 + P_8P_6 + P_6P_1 \\ &= \alpha + \beta + \partial(P_1P_8P_2 + P_8P_9P_2 + P_2P_9P_3 + \cdots). \end{aligned} \tag{4.30}$$

Example: The Projective Plane. The projective plane RP^2 can be regarded as a rectangle with diametrically opposite points identified. Suppose we decompose RP^2 into eight triangles as below:



Triangulating the projective plane.

Consider the entire surface

$$\sigma = P_1P_2P_5 + P_1P_5P_4 + \cdots \in C_2(RP^2). \tag{4.31}$$

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

$$\begin{aligned}\partial(\sigma) &= P_1P_2 + P_2P_3 + P_3P_4 + P_4P_1 + P_1P_2 + P_2P_3 + P_3P_4 + P_1P_2 \\ &= 2(\alpha - \beta) \neq 0. \end{aligned} \tag{4.32}$$

Although RP^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 RP^2. The surface σ of RP^2 is not a two-cycle, therefore. Indeed $Z_2(RP^2)$, and *a fortiori* $H_2(RP^2)$, 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 (4.32) shows that this is the boundary of $\frac{1}{2}\sigma$. Thus $H_2(RP^2, \mathbf{R})$ and $H_1(RP^2, \mathbf{R})$ vanish, while $H_0(RP^2, \mathbf{R}) = \mathbf{R}$.

We can now see the advantge 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 $\mathbf{Z}_2$ to denote the additive group of the integers *modulo* two, we can write $H_1(RP^2, \mathbf{Z}) = \mathbf{Z}_2$. This homology space is a set with only two members $\{0\gamma, 1\gamma\}$. The finite $H_1(RP^2, \mathbf{Z}) = \mathbf{Z}_2$ 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. The torsion becomes invisible when we allow real numbers as a coefficients.

We introduced real-number homology first because the theory of vector spaces is simpler than that of modules, and more familiar to physicists. We were however, buying a simplication at the expense of throwing away information.

The Euler Character

The sum

$$\chi \stackrel{\text{def}}{=} \sum_{p=0}^{d} (-1)^p \dim H_p(M, \mathbf{R})) \tag{4.33}$$

is called the *Euler character* of M. For the 2-sphere, $\chi = 2$, and for the n-torus, $\chi = 0$. This number is manifestly a topological invariant because the individual $\dim H_p(M)$ 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 the number of edges and F 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

$$\chi = \sum_{p=0}^{d} (-1)^p \dim C_p(M). \tag{4.34}$$

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.

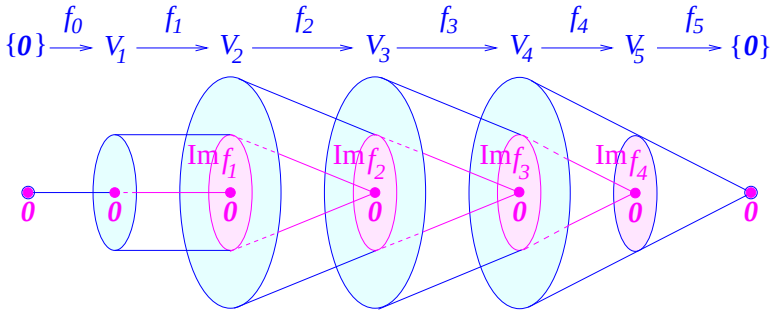
The tool that we will use to relate the alternating sum of the Betti numbers to the alternating sum of the dimensions of the C_p is the *exact sequence*. We say that a set of vector spaces V_p with maps $f_p : V_p \to V_{p+1}$, is an exact sequence if $\text{Ker}\,(f_p) = \text{Im}\,(f_{p-1})$. For example, if all cycles were boundaries then the set of spaces C_p with the map ∂_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 $\mathrm{Im}\,(f_{p-1}) \subset \mathrm{Ker}\,(f_p)$, and the number $\dim H_p = \dim\,(\mathrm{Ker}\,f_p) - \dim\,(\mathrm{Im}\,f_{p-1})$ provides a measure of how far this set inclusion falls short of being an equality.

Suppose that

$$\{\mathbf{0}\} \xrightarrow{f_0} V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} \ldots \xrightarrow{f_{n-1}} V_n \xrightarrow{f_n} \{\mathbf{0}\} \tag{4.35}$$

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 $V_n = \mathrm{Im}\,f_{n-1}$. Similarly, the fact that $\mathrm{Ker}\,f_1 = \mathrm{Im}\,f_0 = \{\mathbf{0}\}$, shows that $\mathrm{Im}\,f_1 \in V_2$ is an isomorphic image of V_1. This situation is represented schematically in the following figure:



A schematic representation of an exact sequence.

Now the *range-nullspace theorem* tells us that

$$\begin{aligned} \dim V_p &= \dim\,(\mathrm{Im}\,f_p) + \dim\,(\mathrm{Ker}\,f_p) \\ &= \dim\,(\mathrm{Im}\,f_p) + \dim\,(\mathrm{Im}\,f_{p-1}). \end{aligned} \tag{4.36}$$

When we take the alternating sum of the dimensions, and use $\dim\,(\mathrm{Im}\,f_0) = 0$ and $\dim\,(\mathrm{Im}\,f_n) = 0$, we find that the sum telescopes to give

$$\sum_{p=0}^{n} (-1)^p \dim V_p = 0. \tag{4.37}$$

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 $\dim(\mathrm{Ker}\,\partial_p) = \dim(\mathrm{Im}\,\partial_{p+1}) + \dim H_p$. Using this and the range-nullspace theorem in the same manner as above, shows that

$$\sum_{p=0}^{d}(-1)^p \dim C_p(M) = \sum_{p=0}^{d}(-1)^p \dim H_p(M). \tag{4.38}$$

This confirms our claim.

4.3.2 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 define

$$(c, \omega) = \sum_i a_i \int_{s_i} \omega. \tag{4.39}$$

The perhaps mysterious notion of "adding" geometric simplices is thus given a concrete interpretation in terms of adding real numbers.

Stokes theorem now reads

$$(\partial c, \omega) = (c, d\omega), \tag{4.40}$$

suggesting that d and ∂ should be regarded as adjoints of each other.

The key observation is that the pairing between chains and forms projects to a pairing of homology classes and cohomology classes. In other words

$$(z + \partial c, \omega + d\chi) = (z, \omega), \tag{4.41}$$

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

$$\begin{aligned}(z, \omega_2) = \int_z \omega_2 &= \int_z \omega_1 + \int_z d\eta, \\ &= \int_z \omega_1 + \int_{\partial z} \eta, \\ &= \int_z \omega_1, \\ &= (z, \omega_1),\end{aligned} \tag{4.42}$$

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 $H^p_{DR}(M)$ 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_{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_{DR}, and there might be distinct elements of H^p_{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, $(H_p(M, \mathbf{R}))^* = H^p_{DR}(M, \mathbf{R})$. We will not try to prove this, but be satisfied with some examples.

The statement $(H_p(M))^* = H^p_{DR}(M)$ 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 iff

$$\int_{z_i} \omega = 0 \tag{4.43}$$

for all cycles $z_i \in Z_p$. It suffices to check this for one representative of each homology class.

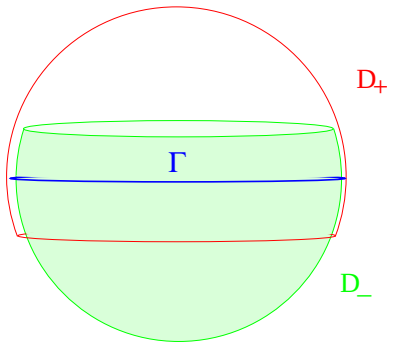
2) If $z_i \in Z_p$, $i = 1, \ldots, \dim H_p$, is a basis for the p-th homology space, and α_i, a set of numbers, one for each z_i, then there exists a closed p-form ω such that

$$\int_{z_i} \omega = \alpha_i. \tag{4.44}$$

If ω^i constitute a basis of the vector space $H^p(M)$, then the matrix of numbers

$$\Omega_i{}^j = (z_i, \omega^j) = \int_{z_i} \omega^j, \tag{4.45}$$

is called the *period matrix*, and the $\Omega_i{}^j$ themselves are the *periods*.
Example: $H_1(T^2) = \mathbf{R} \oplus \mathbf{R}$ is two-dimensional. Since a finite dimensional vector space and its dual have the same dimension, de Rham tells us that $H^1_{DR}(T^2)$ 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 $\omega = \alpha_\theta d\theta + \alpha_\phi d\phi$ returns $\int_{z_\theta} \omega = \alpha_\theta$ and $\int_{z_\phi} \omega = \alpha_\phi$ for any $\alpha_\theta, \alpha_\pi$, so $\{d\theta, d\phi\}$ and $\{z_\theta, z_\phi\}$ are dual bases.
Example: As an illustration of de Rham part 1), observe that it is easy to show that a closed 1-form 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 the proviso ensures that f is not multivalued.
Example: A more subtle problem is to show that, given a 2-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.



Since both D_+ and D_- are contractable, there are 1-forms χ_+ and χ_- such that $\omega = d\chi_+$ in D_+ and $\omega = d\chi_-$ in D_-. Thus

$$d(\chi_+ - \chi_-) = 0, \quad \text{in} \quad D_+ \cap D_-. \tag{4.46}$$

Dividing the sphere into two disjoint sets with a common (but oppositely oriented) boundary $\Gamma \in D_+ \cap D_-$ we have

$$0 = \int_{S^2} \omega = \oint_\Gamma (\chi_+ - \chi_-), \tag{4.47}$$

and this is true for any such curve Γ. Thus, by the previous example,

$$\phi = (\chi_+ - \chi_-) = df \tag{4.48}$$

for some smooth function defined in $\Gamma \in D_+ \cap D_-$. We now introduce a *partition of unity* subordinate to the cover of S^2 by D_+ and D_-. This 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

$$f = \rho_+ f - (-\rho_-) f, \tag{4.49}$$

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 value to assign to it.

We now observe that

$$\chi_+ + df_+ = \chi_- + df_-, \quad \text{in} \quad D_+ \cap D_-. \tag{4.50}$$

Thus $\omega = d\chi$ where χ is defined everywhere by the rule

$$\begin{aligned} \chi &= \chi_+ + df_+, \quad \text{in} \quad D_+ \\ &= \chi_- + df_-, \quad \text{in} \quad D_-. \end{aligned} \tag{4.51}$$

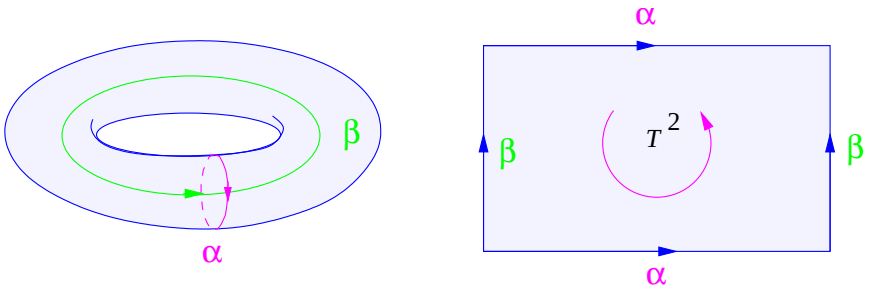
It does not matter which definition we take in the cingular region $D_+ \cap D_-$, because the two definitions coincide there.

This methods of this example, a special case of the *Mayer-Vietoris principle*, can be extended to give a proof of de Rham's claims.

Example: Suppose that the cycles generating the homology group $H_1(T^2)$ of the 2-torus are α and β, and that a and b are closed ($da = db = 0$), but not necessarily exact, 1-forms. We will show that

$$\int_{T^2} a \wedge b = \int_\alpha a \int_\beta b - \int_\alpha b \int_\beta a.$$

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.



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' theorm to evaluate

$$\int_{T^2} a \wedge b = \int_{T^2} d(fb) = \int_{\partial T^2} fb$$

The two integrals on the two vertical sides of the rectangle can be combined to a single integral over the points of the 1-cycle β:

$$\int_{\text{vertical}} fb = \int_\beta [f(L_x, y) - f(0, y)]b.$$

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 1-cycle α, so the difference $f(L_x, y) - f(0, y)$ is equal to $\int_\alpha a$. Thus

$$\int_\beta [f(L_x, y) - f(0, y)]b = \int_\alpha a \int_\beta b.$$

Similarly, the contributions of the two horizontal sides is

$$\int_\alpha [f(x, 0) - f((x, L_y)]b = -\int_\beta a \int_\alpha b.$$

On putting the contributions of both pairs of sides together, the claimed result follows.

4.4 Hodge Theory and the Morse Index

The Laplacian, when acting on a scalar function ϕ is simply div (grad ϕ), but when acting on vectors it becomes

$$\nabla^2 \mathbf{v} = \text{grad}\,(\text{div}\,\mathbf{v}) - \text{curl}\,(\text{curl}\,\mathbf{v}). \tag{4.52}$$

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 $\mathbf{T}$ 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 (4.52) 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.

4.4.1 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

$$\langle a, b\rangle_p = \langle b, a\rangle_p = \int_M a \star b = \frac{1}{p!}\int d^D x \sqrt{g}\, a_{i_1 i_2 \ldots i_p} b^{i_1 i_2 \ldots i_p}. \tag{4.53}$$

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

$$\langle da, b\rangle_{p+1} = \langle a, \delta b\rangle_p, \tag{4.54}$$

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

$$\begin{aligned} d(a\star b) &= da\star b + (-1)^p a(d\star b) \\ &= da\star b + (-1)^p(-1)^{p(D-p)} a\star(\star d\star b) \\ &= da\star b - (-1)^{Dp+1} a\star(\star\, d\star b). \end{aligned} \tag{4.55}$$

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

$$\int_M da\star b = (-1)^{Dp+1}\int_M a\star(\star\, d\star b), \tag{4.56}$$

or

$$\langle da, b\rangle_{p+1} = (-1)^{Dp+1}\langle a, (\star\, d\star) b\rangle_p \tag{4.57}$$

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

$$\delta = (-1)^{Dp+D+1}\star\, d\star\,. \tag{4.58}$$

Observe how the sequence of maps in $\star d \star$ works:

$$\Omega^p(M) \xrightarrow{\star} \Omega^{D-p}(M) \xrightarrow{d} \Omega^{D-p+1}(M) \xrightarrow{\star} \Omega^{p-1}(M). \tag{4.59}$$

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

$$\Delta_p = \delta d + d\delta, \tag{4.60}$$

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,

$$\Delta_p = -\nabla^2 \tag{4.61}$$

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

$$\langle (\delta d + d\delta)a, b\rangle_p = \langle \delta a, \delta b\rangle_{p-1} + \langle da, db\rangle_{p+1} = \langle a, (\delta d + d\delta)b\rangle_p, \tag{4.62}$$

and so we deduce that Δ_p is self-adjoint on $\Omega^p(M)$. The middle terms in (4.62) 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 (4.62) for $a = b$ becomes that

$$0 = \langle \delta a, \delta a\rangle_{p-1} + \langle da, da\rangle_{p+1}. \tag{4.63}$$

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

$$V = \mathrm{Ker}\,(A) \oplus \mathrm{Im}\,(A). \tag{4.64}$$

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

$$\begin{aligned} \omega &= (d\delta + \delta d)\eta + \gamma \\ &= d\alpha + \delta\beta + \gamma, \end{aligned} \tag{4.65}$$

where $\alpha = \delta\eta$, $\beta = d\eta$, and γ is harmonic. This result is known as the *Hodge decomposition* of ω. 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

$$0 = d\alpha + \delta\beta + \gamma \tag{4.66}$$

with non-zero $d\alpha$, $\delta\beta$ and γ. To see that this is not possible, take the d of (4.66) and then the inner product of the result with β. Because $d(d\alpha) = d\gamma = 0$, we end up with

$$\begin{aligned} 0 &= \langle \beta, d\delta\beta \rangle \\ &= \langle \delta\beta, \delta\beta \rangle. \end{aligned} \tag{4.67}$$

Thus $\delta\beta = 0$. Now apply δ to the two remaining terms of (4.66) 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 (4.66) asserts that $\gamma = 0$.

Suppose that ω is closed. Then our strategy of taking the d of the decomposition

$$\omega = d\alpha + \delta\beta + \gamma, \tag{4.68}$$

followed by an inner product with β leads to $\delta\beta = 0$. A closed form can thus be decomposed as

$$\omega = d\alpha + \gamma \tag{4.69}$$

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

$$\dim(\mathrm{Ker}\,\Delta_{\mathrm{p}}) = \dim\left(H^p(M)\right) = b_p. \tag{4.70}$$

Here b_p is the p-th Betti number. From this we immediately deduce that

$$\chi = \sum_{p=0}^{D} (-1)^p \dim(\mathrm{Ker}\,\Delta_{\mathrm{p}}), \tag{4.71}$$

where χ 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:

$$\Delta_p u_\lambda = \lambda u_\lambda. \tag{4.72}$$

Then

$$\begin{aligned} \lambda\, du_\lambda &= d\,\Delta_p u_\lambda \\ &= d(d\delta + \delta d)u_\lambda \\ &= (d\delta)du_\lambda \\ &= (\delta d + d\delta)du_\lambda \\ &= \Delta_{p+1} du_\lambda. \end{aligned} \tag{4.73}$$

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

$$\lambda u_\lambda = (\delta d + d\delta)u_\lambda = d(\delta u_\lambda). \tag{4.74}$$

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

$$\sum_p (-1)^p \dim V_p^\lambda = \begin{cases} = \chi, & \lambda = 0, \\ = 0, & \lambda \neq 0. \end{cases} \tag{4.75}$$

All the topology resides in the null-spaces, therefore.

Exercise: 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 fact is known as Poincaré duality.

4.4.2 Morse Theory

Suppose, as in the previous section, M is a D-dimensional compact, oriented, manifold without boundary and $V : M \to \mathbf{R}$ 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

$$V(x) = V(0) + \frac{1}{2}H_{\mu\nu}x^\mu x^\nu + \dots . \tag{4.76}$$

Here, the matrix $H_{\mu\nu}$ is the *Hessian*

$$H_{\mu\nu} = \left.\frac{\partial^2 V}{\partial x^\mu \partial x^\nu}\right|_0 . \tag{4.77}$$

We can change co-ordinates so as reduce the Hessian to a canonical form with only $\pm 1, 0$ on the diagonal:

$$H_{\mu\nu} = \begin{pmatrix} -I_m & & \\ & I_n & \\ & & 0_{D-m-n} \end{pmatrix} . \tag{4.78}$$

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

$$\sum_{m=0}^{D} (-1)^m N_m = \chi. \tag{4.79}$$

Here χ 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 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. The Poincaré-Hopf theorem now states that, for a compact orientable manifold and a vector field with only finitely many zeros,

$$\sum_{\text{zeros } n} i(z_n) = \chi. \tag{4.80}$$

A tangent vector field must therefore always have at least one zero unless $\chi = 0$. Since the two-sphere has $\chi = 2$, it cannot be combed.

Supersymmetric Quantum Mechanics

Ed Witten gave a beautiful proof of the Morse index theorem 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 *deja 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 we take to obey

$$\{\psi^{\dagger\mu}, \psi^\nu\} \equiv \psi^{\dagger\mu}\psi^\nu + \psi^\nu\psi^{\dagger\mu} = g^{\mu\nu}. \tag{4.81}$$

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

$$(\psi^{\dagger 1})^{p_1}(\psi^{\dagger 2})^{p_2}\dots(\psi^{\dagger n})^{p_n}|0\rangle, \tag{4.82}$$

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

$$\langle 0|\psi^{\mu_p}\dots\psi^{\mu_2}\psi^{\mu_1}\dots\psi^{\dagger\nu_1}\psi^{\dagger\nu_2}\dots\psi^{\dagger\nu_q}|0\rangle$$

is zero unless $p = q$, in which case it is equal to

$$g^{\mu_1\nu_1}g^{\mu_2\nu_2}\dots g^{\mu_p\nu_p} \pm (\text{permutations}).$$

We now make the correspondence

$$\frac{1}{p!}f_{\mu_1\mu_2\dots\mu_p}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\dots\psi^{\dagger\mu_p}|0\rangle \leftrightarrow \frac{1}{p!}f_{\mu_1\mu_2\dots\mu_p}(x)dx^{\mu_1}dx^{\mu_2}\dots dx^{\mu_p}, \tag{4.83}$$

to identify p-fermion states with p-forms. We think of $f_{\mu_1\mu_2\dots\mu_p}(x)$ 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

$$|a\rangle = \frac{1}{p!}a_{\mu_1\mu_2\dots\mu_p}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\dots\psi^{\dagger\mu_p}|0\rangle \tag{4.84}$$

and

$$|b\rangle = \frac{1}{q!}b_{\mu_1\mu_2\dots\mu_q}(x)\psi^{\dagger\mu_1}\psi^{\dagger\mu_2}\dots\psi^{\dagger\mu_q}|0\rangle \tag{4.85}$$

to be

$$\begin{aligned}\langle a,b\rangle &= \int_M d^Dx\sqrt{g}\,\frac{1}{p!q!}\,a^*_{\mu_1\mu_2\dots\mu_p}b_{\nu_1\nu_2\dots\nu_q}\langle 0|\psi^{\mu_p}\dots\psi^{\mu_1}\psi^{\dagger\nu_1}\dots\psi^{\dagger\nu_q}|0\rangle \\ &= \delta_{pq}\int_M d^Dx\sqrt{g}\,\frac{1}{p!}\,a^*_{\mu_1\mu_2\dots\mu_p}b^{\mu_1\mu_2\dots\mu_p}.\end{aligned} \tag{4.86}$$

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

$$\delta = d^\dagger = (\psi^{\dagger\mu}\partial_\mu)^\dagger = \partial_\mu^\dagger \psi^\mu = -\partial_\mu \psi^\mu = -\psi^\mu \partial_\mu. \tag{4.87}$$

The hermitian adjoint of ∂_μ is here being taken with respect to the standard $L^2(\mathbf{R}^D)$ inner product. This computation becomes more complicated when when $g_{\mu\nu}$ becomes position dependent. The adjoint $\partial_\mu^\dagger$ 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

$$d\delta + \delta d = -\{\psi^{\dagger\mu}, \psi^\nu\}\, \partial_\mu \partial_\nu = -\delta^{\mu\nu}\partial_\mu\partial_\nu = -\nabla^2. \tag{4.88}$$

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

$$d_t = e^{-tV(x)} d\, e^{tV(x)}, \quad \delta_t = e^{tV(x)} \delta\, e^{-tV(x)}, \tag{4.89}$$

and

$$\Delta_t = d_t\delta_t + \delta_t d_t. \tag{4.90}$$

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

$$\omega = d_t\alpha + \delta_t\beta + \gamma, \tag{4.91}$$

where $\Delta_t \gamma = 0$, and concude that

$$\dim\left(\mathrm{Ker}\left(\Delta_t\right)_p\right) = b_p \tag{4.92}$$

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

$$\begin{aligned} d_t &= \psi^{\dagger\mu}(\partial_\mu + t\partial_\mu V) \\ \delta_t &= \psi^{\mu}(\partial_\mu - t\partial_\mu V) \end{aligned} \tag{4.93}$$

and find

$$d_t\delta_t + \delta_t d_t = -\nabla^2 + t^2|\nabla V|^2 + t[\psi^{\dagger\mu}, \psi^\nu]\,\partial^2_{\mu\nu}V. \tag{4.94}$$

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

$$\Delta_{p,t} \approx \sum_{i=1}^{D}\left\{-\frac{\partial^2}{\partial x_i^2} + t^2\lambda_i^2 x_i^2 + t\lambda_i[\psi^{\dagger i}, \psi^i]\right\}. \tag{4.95}$$

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

$$t\sum_{i=1}^{D}\left\{|\lambda_i|(1+2n_i) \pm \lambda_i\right\}. \tag{4.96}$$

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 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. The harmonic oscillator approximation thus suggests that $b_p = N_p$.

If we could trust our computation of the energy spectrum, we would have established the Morse theorem

$$\sum_{m=0}^{D} (-1)^m N_m = \sum_{p=0}^{D} (-1)^m b_p = \chi, \tag{4.97}$$

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

$$N_p \geq b_p. \tag{4.98}$$

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 paramater 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

We now discuss the complications that arise in Witten's fermionic calculus when we have to take curvature into account. In doing so we will make manifest the Riemannian geometry that is almost completely concealed by Hodge's d, δ calculus. We will find ourselves introducing the covariant derivative in an unconventional, but powerful, manner.

We assume that our manifold M is equipped with a torsion-free connection $\Gamma^{\mu}_{\nu\lambda} = \Gamma^{\mu}_{\lambda\nu}$, and we use it 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:

$$\begin{aligned} [\hat{\nabla}_\mu, f] &= \partial_\mu f, \\ [\hat{\nabla}_\mu, \psi^{\dagger\nu}] &= -\Gamma^{\nu}_{\mu\lambda}\psi^{\dagger\lambda}, \\ [\hat{\nabla}_\mu, \psi^{\nu}] &= -\Gamma^{\nu}_{\mu\lambda}\psi^{\lambda}. \end{aligned} \tag{4.99}$$

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

$$\begin{aligned} \hat{\nabla}_\mu\left(f_\nu\psi^{\dagger\nu}|0\rangle\right) &= \left([\hat{\nabla}_\mu, f_\nu\psi^{\dagger\nu}] + f_\nu\psi^{\dagger\nu}\hat{\nabla}_\mu\right)|0\rangle \\ &= \left([\hat{\nabla}_\mu, f_\nu]\psi^{\dagger\nu} + f_\alpha[\hat{\nabla}_\mu, \psi^{\dagger\alpha}]\right)|0\rangle \\ &= (\partial_\mu f_\nu - f_\alpha\Gamma^{\alpha}_{\mu\nu})\psi^{\dagger\nu}|0\rangle \\ &= (\nabla_\mu f_\nu)\,\psi^{\dagger\nu}|0\rangle, \end{aligned} \tag{4.100}$$

where

$$\nabla_\mu f_v = \partial_\mu f_\nu - \Gamma^{\alpha}_{\mu\nu} f_\alpha, \tag{4.101}$$

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 may see somewhat shocking to someone familiar with covariant derivatives—being able pass the metric through a covariant derivative is a basic compatibilty condition in Riemann geometry—but all is not lost because $\hat{\nabla}_\mu$ (with a caret) is not quite the same beast as ∇_μ. We proceed as follows:

$$\begin{aligned} \partial_\alpha g^{\mu\nu} &= [\hat{\nabla}_\alpha, g^{\mu\mu}] \\ &= [\hat{\nabla}_\alpha, \{\psi^{\dagger\mu}, \psi^{\nu}\}] \end{aligned}$$

$$\begin{aligned} &= [\hat{\nabla}_\alpha, \psi^{\dagger\mu}\psi^\nu] + [\hat{\nabla}_\alpha, \psi^\nu\psi^{\dagger\mu},] \\ &= -\{\psi^{\dagger\mu}, \psi^\nu\}\Gamma^\nu_{\alpha\lambda} - \{\psi^{\dagger\nu}, \psi^\lambda\}\Gamma^\mu_{\alpha\lambda} \\ &= -g^{\mu\lambda}\,\Gamma^\nu_{\alpha\lambda} - g^{\nu\lambda}\,\Gamma^\mu_{\alpha\lambda}. \end{aligned} \tag{4.102}$$

We conclude that

$$\partial_\alpha g^{\mu\nu} + g^{\mu\lambda}\Gamma^\nu_{\alpha\lambda} + g^{\nu\lambda}\Gamma^\mu_{\alpha\lambda} \equiv \nabla_\alpha g^{\mu\nu} = 0. \tag{4.103}$$

Metric compatibility is therefore implicit in the formalism. The connection will therefore be the standard Riemannian one

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}g^{\alpha\lambda}\left(\partial_\mu g_{\lambda\nu} + \partial_\nu g_{\mu\lambda} - \partial_\lambda g_{\mu\nu}\right). \tag{4.104}$$

Knowing this, we can compute the adjoint of $\hat{\nabla}_\mu$.

$$\begin{aligned} \left(\hat{\nabla}_\mu\right)^\dagger &= -\frac{1}{\sqrt{g}}\hat{\nabla}_\mu\sqrt{g} \\ &= -\left(\hat{\nabla}_\mu + \partial_\mu \ln\sqrt{g}\right) \\ &= -(\hat{\nabla}_\mu + \Gamma^\nu_{\mu\nu}). \end{aligned} \tag{4.105}$$

That $\Gamma^\nu_{\mu\nu}$ is the logarithmic derivative of $\sqrt{g}$ is a standard identity for the Riemann connection. The resultant formula for $(\hat{\nabla}_\mu)^\dagger$ can be used to verify that the second and third equations in (4.99) are compatible with each other.

We can also compute $[[\hat{\nabla}_\mu, \hat{\nabla}_\nu], \psi^\alpha]$ and from it deduce that

$$[\hat{\nabla}_\mu, \hat{\nabla}_\nu] = R_{\sigma\lambda\mu\nu}\psi^{\dagger\sigma}\psi^\lambda, \tag{4.106}$$

where

$$R^\alpha{}_{\beta\mu\nu} = \partial_\mu\Gamma^\alpha_{\beta\nu} - \partial_\nu\Gamma^\alpha_{\beta\mu} + \Gamma^\alpha_{\lambda\mu}\Gamma^\lambda_{\beta\nu} - \Gamma^\alpha_{\lambda\nu}\Gamma^\lambda_{\beta\mu} \tag{4.107}$$

is the Riemann curvature tensor.

We now define d to be

$$d = \psi^{\dagger\mu}\hat{\nabla}_\mu. \tag{4.108}$$

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

$$\begin{aligned} \delta &\equiv \left(\psi^{\dagger\mu}\hat{\nabla}_\mu\right)^\dagger \\ &= \hat{\nabla}^\dagger_\mu\,\psi^\mu \\ &= -(\hat{\nabla}_\mu + \Gamma^\nu_{\mu\nu})\psi^\mu \\ &= -\psi^\mu(\hat{\nabla}_\mu + \Gamma^\nu_{\mu\nu}) + \Gamma^\mu_{\mu\nu}\psi^\nu \\ &= -\psi^\mu\hat{\nabla}_\mu. \end{aligned} \tag{4.109}$$

The Laplace-Beltrami operator can now be worked out as

$$\begin{aligned} d\delta + \delta d &= -\left(\psi^{\dagger\mu}\hat{\nabla}_\mu\psi^\nu\hat{\nabla}_\nu + \psi^\nu\hat{\nabla}_\nu\psi^{\dagger\mu}\hat{\nabla}_\mu\right) \\ &= -\left(\{\psi^{\dagger\mu},\psi^\nu\}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^\sigma_{\mu\nu}\hat{\nabla}_\sigma) + \psi^\nu\psi^{\dagger\mu}[\hat{\nabla}_\nu,\hat{\nabla}_\mu]\right) \\ &= -\left(g^{\mu\nu}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^\alpha_{\mu\nu}\hat{\nabla}_\sigma) + \psi^\nu\psi^{\dagger\mu}\psi^{\dagger\sigma}\psi^\lambda R_{\sigma\lambda\nu\mu}\right) \end{aligned} \tag{4.110}$$

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

$$d\delta + \delta d = -g^{\mu\nu}(\hat{\nabla}_\mu\hat{\nabla}_\nu - \Gamma^\sigma_{\mu\nu}\hat{\nabla}_\sigma) - \psi^{\dagger\alpha}\psi^\beta\psi^{\dagger\mu}\psi^\nu R_{\alpha\beta\mu\nu}. \tag{4.111}$$

This result is called the *Weitzenböck formula*. An equivalent formula can be derived directly from (4.58), 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

$$B = \hat{\nabla}^\dagger_\mu\, g^{\mu\nu}\,\hat{\nabla}_\nu \tag{4.112}$$

which shows that it is a positive operator.

Chapter 5

Groups and Representation Theory

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 with the same multiplication table as the group. These sets are called *representations* of the group.

Given a group, we will endeavour to find and classify all possible representations.

5.1 Basic Ideas

We will 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 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*.

iii) Existence of inverse: For each $g \in G$ there is an element g^{-1} such that $g^{-1}g = e$.

From these axioms there follows some conclusions that are so basic that they are often included in the axioms themselves, but since they are not independent, we will 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 $\mathbf{Z}$ under addition. The binary operation is $(n, m) \to n+m$. This is not a finite group.
2) The integers modulo n under addition. $(m, n) \to m + n$, mod n. This group is denoted by $\mathbf{Z}_n$.
3) The non-zero integers modulo p (a prime) under *multiplication* $(m, n) \to mn$, mod p. If the modulus is not a prime number, we do not get a group (why not?).
4) 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 $(f_i, f_j) \to f_i \circ f_j$. Here the "$\circ$" is a standard notation for composition of functions: $(f_i \circ f_j)(z) = f_i(f_j(z))$.

5) The set of rotations in three dimensions, equivalently the set of 3×3 real matrices O, obeying $O^T O = I$, and $\det O = 1$. This is the group $SO(3)$. Other groups $SO(n)$ are defined analogously. If we relax the condition on the determinant we get the groups $O(n)$. These are examples of *Lie groups*, *i.e.* groups which are also a manifold M and whose multiplication law is a smooth function $M \times M \to M$.
6) Groups are often specified by giving a list of *generators* and *relations*. For example the *cyclic group* of order n, 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 with 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 will need:

i) *Subgroups*: If a subset of elements of a group forms a group, it is called a subgroup. For example, $\mathbf{Z}_{12}$ 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 $H \subseteq G$, with 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 intersect, they coincide (if $g_1 h_1 = g_2 h_2$, then $g_2 = g_1(h_1 h_2^{-1})$ and $g_1 H = g_2 H$.). If H is a finite group, each coset has the same number of distinct elements as H (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 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 $G/H \equiv \{g_1H,\, g_2H, \ldots\}$ by taking a representative element from each of g_iH, and g_jH, taking the product of these elements, and defining $(g_iH)(g_jH)$ 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*[2].

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 $g_2 = g^{-1}g_1g$. If g_1 is conjugate to g_2 we will write $g_1 \sim g_2$. Conjugacy is an *equivalence relation*[3], 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 $h^{-1}gh = g$ form a subgroup. The set elements of 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 $n \times n$ unitary matrices, the conjugacy classes are the set of matrices with the same eigenvalues.

---

[2]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 monster is the automorphism group of a certain algebra, called the Griess algebra. The mysterious "Monstrous moonshine" links its representation theory to the elliptic modular function $J(\tau)$ and to string theory.

[3]An equivalence relation, $\sim$, is a binary relation which 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.*

Example: Permutations. The permutation group on n objects, S_n, has order $n!$. Suppose we consider a permutation π in S_8 that takes

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 1 & 5 & 4 & 7 & 6 & 8 \end{pmatrix}$$

We can write this out in cycle notation

$$\pi = (123)(45)(67)(8).$$

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 $\pi(1) = 2$, $\pi(2) = 3$, $\pi(3) = 1$. The "8", being both first and last in its cycle, maps to itself: $\pi(8) = 8$. Any permutation with this cycle pattern, $(***)(**)(**)(*)$, will be in the same conjugacy class as π. We say that there is 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

$$\operatorname{sgn}(\pi_1)\operatorname{sgn}(\pi_2) = \operatorname{sgn}(\pi_1\pi_2),$$

so the *even* ($\operatorname{sgn}\pi = +1$) permutations form an invariant subgroup called the *Alternating group*, A_n. The alternating group, A_n, is simple for $n \geq 5$, and, as Galois showed, this simplicity prevents the solution of the general quintic (or any higher degree) equation by radicals.

If we write out the group elements is 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 *Cayley's Theorem*.

Exercise: Let H_1, H_2 be two subgroups of a group G. Show that $H_1 \cap H_2$ is also a subgroup.

Exercise: 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.

Exercise: 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.

Exercise: 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.

Exercise: 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: Show that the set of powers a^n of an element $a \in G$ form a subgroup. Let p be prime. Show that the set $\{1, 2, \ldots p-1\}$ forms a group of order $(p-1)$ under multiplication modulo p, and, by the use of 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: 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[4] $(p-1)(q-1)$, 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.

[4]Has no factors in common with.

Exercise: Consider the group with multiplication table[5]

$\mathcal{G}$	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

It 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.

Exercise: 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.

5.1.3 Group Actions on Sets

Groups usually appear in physics as symmetries: they act on some physical object to change it in some way, perhaps while leaving some other property invariant.

Suppose X is a set. We will 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 will call $gx \in X$, and such that $g_2(g_1 x) = (g_1 g_2)x$, and $ex = x$. There is some controlled vocabulary for group actions:

[5] To find AB look in row A column B.

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 implies that $g = e$. Equivalently, if the map $G \to \mathrm{Map}\,(X \to X)$ is 1-1. If the action is not faithful, the set of g corresponding to to the identity map is an invariant subgroup H of G, and we can take G/H as having 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 *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 $S^2 \simeq SO(3)/SO(2)$.

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 homomorphism from G to a subgroup of $GL(n, \mathbf{C})$, the group of invertible $n \times n$ matrices with complex entries. In other words it is a set of $n \times n$ matrices that obeys the group

multiplication law

$$D(g_1)D(g_2) = D(g_1 g_2). \tag{5.2}$$

Given such a representation, we can form another one $D'(g)$ by conjugation with any 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 will call them *equivalent* representations and write $D \sim D'$, since we can think of them as being matrices representing the same linear map, but in different bases. Our task in this chapter will be to find and classify representations up to equivalence.

Real and Pseudoreal 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 such real matrices. In this case we say that D is *pseudo-real*.
Example: Consider the defining representation of $SU(2)$ (the group of 2×2 unitary matrices with unit determinant.) Such matrices are necessarily of the form

$$U = \begin{pmatrix} a & -b^* \\ b & a^* \end{pmatrix}, \tag{5.4}$$

with $|a|^2 + |b|^2 = 1$. They are therefore specified by *three* real parameters and so the group manifold is threee 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 impossible to find 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, while we have seen that we need three dimensions to describe all possible $SU(2)$ matrices.
Exercise: Show that if $D(g)$ is a representation, then so is

$$D'(g) = (D^{-1})^T, \tag{5.6}$$

where the superscript T denotes the transposed matrix.

Direct Sum and Direct Product

Another way to get new representations from old is 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 matrix

$$\begin{pmatrix} D^1(g) & 0 \\ 0 & D^2(g) \end{pmatrix}. \tag{5.7}$$

We will be 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*, $D^1 \otimes D^2$, to be the natural action on the tensor product of the representation spaces. In other words, if $D^1(g)\mathbf{e}_j^{\{1\}} = \mathbf{e}_i^{\{1\}} D^1_{ij}(g)$ and $D^2(g)\mathbf{e}_j^{\{2\}} = \mathbf{e}_i^{\{2\}} D^2_{ij}(g)$ we define

$$[D^1 \otimes D^2](g)(\mathbf{e}_i^{\{1\}} \otimes \mathbf{e}_j^{\{2\}}) = (\mathbf{e}_k^{\{1\}} \otimes \mathbf{e}_l^{\{2\}}) D^1_{ik}(g) D^2_{lj}(g). \tag{5.8}$$

We think of $D^1_{ik}(g)D^2_{lj}(g)$ being a matrix $D^{1\otimes 2}_{il,jk}(g)$ 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.

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. We call such representations *irreducible*. You cannot usually tell by just looking at a representation whether is is reducible or not. We need to develop some tools. We will begin with a more powerful definition of irreducibilty.

To define irreducibility we need 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 $\{0\}$. If there is a non-trivial invariant subspace, then the set[6] 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 form

$$A_\alpha = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}. \tag{5.9}$$

If we can find a[7] complementary subspace U' which is also invariant, then

$$A_\alpha = \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}, \tag{5.10}$$

and we say that the operators 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*, which, by unitarity, will automatically be invariant. In this case 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

$$\Lambda A_\alpha = B_\alpha \Lambda, \tag{5.11}$$

for all α, then *either*

a) $\Lambda = 0$,
   or
b) Λ is 1-1 and onto (and hence invertible), in which case U and V have the same dimension and $A_\alpha = \Lambda^{-1} B_\alpha \Lambda$.

[6] 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.

[7] Complementary subspaces are not unique.

The proof is straightforward: The relation (5.11) 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

$$\Lambda A_\alpha = A_\alpha \Lambda, \tag{5.12}$$

then either $\Lambda = 0$ or $\Lambda = \lambda I$. To see this observe that (5.12) 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 $D(g) : V \to V$ be a representation. 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

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{g \in G} (D(g)\mathbf{x}, D(g)\mathbf{y}). \tag{5.13}$$

It is easy to see that this new inner product has the same properties as the old one, and in addition

$$\langle D(g)\mathbf{x}, D(g)\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{5.14}$$

This means that the representation is automatically unitary with respect to the new product. If we work with bases that are orthonormal with respect to the new product, and we usually will, then the $D(g)$ are unitary matrices, $D(g^{-1}) = D^{-1}(g) = [D(g)]^\dagger$.

Thus representations of finite groups can always be taken to be unitary. As a consequence, 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 $D^J(g) : V_J \to V_J$ denote 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 an arbitrary matrix with the appropriate number of rows and columns so that the matrix product $D^J M D^K$ is defined. The sum

$$\Lambda = \sum_{g\in G} D^J(g^{-1}) M D^K(g) \tag{5.15}$$

obeys $D^J(g)\Lambda = \Lambda D^K(g)$ for any g. Consequently, Schur's lemma tells us that

$$\Lambda_{il} = \sum_{g\in G} D^J_{ij}(g^{-1}) M_{jk} D^K_{kl}(g) = \lambda \delta_{il}\delta^{JK}. \tag{5.16}$$

Now take M_{ij} to be zero except for one entry, then we have

$$\sum_{g\in G} D^J_{ij}(g^{-1}) D^K_{kl}(g) = \lambda_{jk}\delta_{il}, \delta^{JK} \tag{5.17}$$

where we have taken note that the constant λ depends on the location of the one non-zero entry in M. We can find the constant λ_{jk} by assuming that $K = J$, setting $i = l$, and summing over i. We find

$$|G|\delta_{jk} = \lambda_{jk} \dim J. \tag{5.18}$$

Putting these results together we find that

$$\frac{1}{|G|}\sum_{g\in G} D^J_{ij}(g^{-1}) D^K_{kl}(g) = (\dim J)^{-1}\delta_{jk}\delta_{il}\delta^{JK}. \tag{5.19}$$

If our matrices $D(g)$ are unitary, we can write this as

$$\frac{1}{|G|}\sum_{g\in G} \left(D^J_{ij}(g)\right)^* D^K_{kl}(g) = (\dim J)^{-1}\delta_{ik}\delta_{jl}\delta^{JK}. \tag{5.20}$$

If we regard the complex-valued functions on the set G as forming a vector space, then the entries in the representation matrices are orthogonal with respect to the natural inner product on that space.

There can be no more orthogonal functions on G than the dimension of the function space itself, which is $|G|$. Thus

$$\sum_J (\dim J)^2 \leq |G|. \tag{5.21}$$

In fact, as you will show later, the equality holds. The matrix elements form a complete orthonormal set of functions on G, and the sum of the squares of the dimensions of the inequivalent irreducible representations is equal to the order of G.

Class Functions and Characters

Since

$$\mathrm{tr}\,(C^{-1}DC) = \mathrm{tr}\,D, \tag{5.22}$$

the trace of a representation matrix is the same for equivalent representations. Further since

$$\mathrm{tr}\,\left(D^{-1}(g)D(g_1)D(g)\right) = \mathrm{tr}\,D(g), \tag{5.23}$$

the trace is the same for group elements in the same conjugacy class. The *character*,

$$\chi(g) = \mathrm{tr}\,D(g), \tag{5.24}$$

is therefore said to be a *class function*.

By taking the trace of the matrix element orthogonality relation we see that the characters $\chi^J = \mathrm{tr}\,D^J$ of the irreducible representations obey

$$\frac{1}{|G|}\sum_{g\in G}\left(\chi^J(g)\right)^*\chi^K(g) = \frac{1}{|G|}\sum_i d_i\left(\chi_i^J\right)^*\chi_i^K = \delta^{JK}, \tag{5.25}$$

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 conjugacy classes. 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 can be decomposed into, and how many times, n_J, they occur. We do this forming the *compound character*

$$\chi(g) = \mathrm{tr}\,D(g) \tag{5.26}$$

and observing that if we can find a basis in which

$$D(g) = \underbrace{(D^1(g) \oplus D^1(g) \oplus \cdots)}_{n_1 \text{ terms}} \oplus \underbrace{(D^2(g) \oplus D^2(g) \oplus \cdots)}_{n_2 \text{ terms}} \oplus \cdots, \tag{5.27}$$

then

$$\chi(g) = n_1\chi^1(g) + n_2\chi^2(g) + \cdots \tag{5.28}$$

From this we find

$$n_J = \frac{1}{|G|}\sum_{g\in G} (\chi(g))^* \chi^J(g) = \frac{1}{|G|}\sum_i d_i \ (\chi_i)^* \chi_i^J. \tag{5.29}$$

There are extensive tables of group characters. Here, in particular, is the character table for 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

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

$$1 + 1 + 2^2 + 3^2 + 3^2 = 24 = 4!$$

which is the order of the group.

5.2.3 The Group Algebra

Given a group G, we may take the elements of the group to be the basis of a vector space. We will denote these basis elements by $\mathbf{g}$ to distinguish them from the elements of the group. We retain the multiplication rule, however, so $g_1 \to \mathbf{g}_1$, $g_2 \to \mathbf{g}_2 \Longrightarrow g_3 = g_1g_2 \to \mathbf{g}_1\mathbf{g}_2 = \mathbf{g}_3$. The resulting mathematical object is called the *group algebra*, or *Frobenius algebra*.

The group algebra, considered as a vector space, is automatically a representation. We define the action of g in the most natural way as

$$D(g)\mathbf{g}_i = \mathbf{g}\,\mathbf{g}_i = \mathbf{g}_j D_{ji}(g). \tag{5.30}$$

The matrices $D_{ji}(g)$ make up the *regular representation.* Their entries consist of 1's and 0's, with exactly one non-zero entry in each row and each column.

Exercise: Show that the character of the regular representation has $\chi(e) = |G|$, and $\chi(g) = 0$, for $g \neq e$.

Exercise: 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 of the group automatically gives us a representation of the group algebra. Certain linear combinations of the group elements turn out to be very useful becuase the corresponding matrices can be used to project out vectors with desirable symmetry properties.

Consider the elements

$$\mathbf{e}^J_{\alpha\beta} = \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* \mathbf{g} \tag{5.31}$$

of the group algebra. These have the property that

$$\begin{aligned}
\mathbf{g}_1 \mathbf{e}^J_{\alpha\beta} &= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g)\right]^* (\mathbf{g}_1\mathbf{g}) \\
&= \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\alpha\beta}(g_1^{-1}g)\right]^* \mathbf{g} \\
&= \left[D^J_{\alpha\gamma}(g_1^{-1})\right]^* \frac{\dim J}{|G|} \sum_{g\in G} \left[D^J_{\gamma\beta}(g)\right]^* \mathbf{g} \\
&= \mathbf{e}^J_{\gamma\beta} D^J_{\gamma\alpha}(g_1).
\end{aligned} \tag{5.32}$$

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 this it follows that

$$\begin{aligned}
\mathbf{e}^J_{\alpha\beta}\,\mathbf{e}^K_{\gamma\delta} &= \frac{\dim J}{|G|}\sum_{g\in G}\left[D^J_{\alpha\beta}(g)\right]^*\mathbf{g}\,\mathbf{e}^K_{\gamma\delta} \\
&= \frac{\dim J}{|G|}\sum_{g\in G}\left[D^J_{\alpha\beta}(g)\right]^* D^K_{\epsilon\gamma}(g)\mathbf{e}^K_{\epsilon\delta} \\
&= \delta^{JK}\delta_{\alpha\epsilon}\delta_{\beta\gamma}\,\mathbf{e}^K_{\epsilon\delta} \\
&= \delta^{JK}\delta_{\beta\gamma}\,\mathbf{e}^J_{\alpha\delta},
\end{aligned} \tag{5.33}$$

which, for each J, is the multiplication rule for matrices having zero entries everywhere except for the (i,j)-th, which has a "1". There will be n^2 of these $n\times n$ matrices for each n-dimensional representation, so the Frobenius algebra is isomorphic to a direct sum of simple matrix algebras.

Every element of G can be reconstructed as

$$\mathbf{g} = \sum_J D^J_{ij}(g)\mathbf{e}^J_{ij} \tag{5.34}$$

and once again we deduce that $|G| = \sum_J (\dim J)^2$.

We now define

$$\mathbf{P}^J = \sum_i \mathbf{e}^J_{ii} = \frac{\dim J}{|G|}\sum_{g\in G}\left[\chi^J(g)\right]^*\mathbf{g}. \tag{5.35}$$

We have

$$\mathbf{P}^J\mathbf{P}^K = \delta^{JK}\mathbf{P}^K, \tag{5.36}$$

so these are projection operators. The completeness of the characters shows that

$$\sum_J \mathbf{P}^J = I. \tag{5.37}$$

It should be clear that, if $D(g)$ is any representation, then replacing $\mathbf{g}$ by $D(g)$ in $\mathbf{P}^J$ gives a projection onto the representation space J. In other words, if $\mathbf{v}$ is a vector in the represention space and we set

$$\mathbf{v}_i = \mathbf{e}^J_{ip}\mathbf{v} \tag{5.38}$$

for any fixed p, then

$$D(g)\mathbf{v}_i = D(g)\mathbf{e}^J_{ip}\mathbf{v} = \mathbf{e}^J_{jp}\mathbf{v}D^J_{ji}(g) = \mathbf{v}_j D^J_{ji}(g). \tag{5.39}$$

Of course, if the representation space J does not occur in the decomposition of $D(g)$, then all these terms are identically zero.

5.3 Physics Applications

5.3.1 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

$$L = \frac{1}{2}\dot{\mathbf{x}}^T M \dot{\mathbf{x}} - \frac{1}{2}\mathbf{x}^T V \mathbf{x} \tag{5.40}$$

where M and V are symmetric $n \times n$ matrices with M being positive definite. This gives rise to the equations of motion

$$M\ddot{\mathbf{x}} = V\mathbf{x} \tag{5.41}$$

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

$$-\omega_i^2 M\mathbf{x}_i = V\mathbf{x}_i. \tag{5.42}$$

The normal-mode frequencies are solutions of the secular equation

$$\det\left(V - \omega^2 M\right) = 0, \tag{5.43}$$

and modes with distinct frequencies are orthogonal with respect to the inner product defined by M,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T M \mathbf{y}. \tag{5.44}$$

We will be interested in solving this problem for vibrations about the equilibrium configuration of a molecule. Suppose this equilibrium configuration has a symmetry group G. This will give rise to an n dimensional representation on the space of $\mathbf{x}$'s

$$\mathbf{x} \to D(g)\mathbf{x}, \tag{5.45}$$

which leaves both the intertia matrix M and the potential matrix V unchanged.

$$[D(g)]^T M D(g) = M, \qquad [D(g)]^T V D(g) = V. \tag{5.46}$$

Consequently, if we have an eigenvector $\mathbf{x}_i$ with frequency ω_i,

$$-\omega_i^2 M\mathbf{x}_i = V\mathbf{x}_i \tag{5.47}$$

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 comprised by $D(g)$.

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 character table is

	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

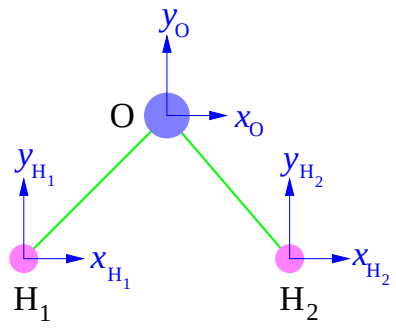
The group is Abelian, so all the representations are one dimensional.

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

$$\mathbf{x} = (x_O, y_O, z_O, x_{H_1}, y_{H_1}, z_{H_1}, x_{H_2}, y_{H_2}, z_{H_2}). \tag{5.48}$$

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.



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

$$\begin{pmatrix} 9 \\ -1 \\ 1 \\ 3 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + 2\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} + 3\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \tag{5.49}$$

or

$$\chi^D = 3\chi^{A_1} + \chi^{A_2} + 2\chi^{B_1} + 3\chi^{B_2}. \tag{5.50}$$

Thus the nine-dimensional representation decomposes as

$$D = 3A_1 \oplus A_2 \oplus 2B_1 \oplus 3B_2. \tag{5.51}$$

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

$$P^{A_1} = \frac{1}{4}[(\chi^{A_1}(e))^* D(e) + (\chi^{A_1}(a))^* D(b) + (\chi^{A_1}(b))^* D(b) + (\chi^{A_1}(ab))^* D(ab)] \tag{5.52}$$

to $\mathbf{v}_{H_1,x}$, a small displacement of H_1 in the x direction. We find

$$\begin{aligned} P^{A_1}\mathbf{v}_{H_1,x} &= \frac{1}{4}(\mathbf{v}_{H_1,x} - \mathbf{v}_{H_2,x} - \mathbf{v}_{H_2,x} + \mathbf{v}_{H_1,x}) \\ &= \frac{1}{2}(\mathbf{v}_{H_1,x} - \mathbf{v}_{H_2,x}). \end{aligned} \tag{5.53}$$

This mode will be 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

$$\begin{aligned} P^{A_1}\mathbf{v}_{H_1,y} &= \frac{1}{2}(\mathbf{v}_{H_1,y} + \mathbf{v}_{H_2,y}), \\ P^{A_1}\mathbf{v}_{O,y} &= \mathbf{v}_{O,y}, \end{aligned} \tag{5.54}$$

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

$$P^{B_2}\mathbf{v}_{H_1,x} = \frac{1}{2}(\mathbf{v}_{H_1,x} + \mathbf{v}_{H_2,x})$$

$$\begin{aligned} P^{B_2}\mathbf{v}_{H_1,y} &= \frac{1}{2}(\mathbf{v}_{H_1,y} - \mathbf{v}_{H_2,y}) \\ P^{B_2}\mathbf{v}_{O,x} &= \mathbf{v}_{O,x} \end{aligned}$$

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

$$y_{H_1} = -y_{H_2} = \alpha x_{H_1} = \alpha x_{H_2} = \beta x_0 = a_2 \tag{5.55}$$

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.

5.3.2 Crystal Field Splittings

A quantum mechanical system has a symmetry G if the hamiltonian $\hat{H}$ obeys

$$D^{-1}(g)\hat{H}D(g) = \hat{H}, \tag{5.56}$$

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 $D(g)$.

A common problem is to understand how degeneracy is lifted by perturbations that break G down to a smaller subgroup H. Now an n-dimensional irreducible representation of G is automatically a representation of any subgroup of G, but in general it will no longer be irreducible. Thus the n-fold degenerate level will split into multiplets, one for each of the irreducible representations of H contained in the original 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[8].

Suppose the crystal has octohedral symmetry. The character table of the octohedral group is

[8]The following discussion and tables are taken from chapter 9 of M. Hamermesh *Group Theory*.

	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

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

$$\chi^l(\theta) = \frac{\sin(2l+1)\theta/2}{\sin\theta/2}, \tag{5.57}$$

and the first few χ^l's evaluated on the rotation angles of the classes of O are

	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

The 9-fold degenerate $l = 4$ multiplet thus decomposes as

$$\begin{pmatrix} 9 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \\ 2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \\ -1 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \tag{5.58}$$

or

$$\chi^4_{SO(3)} = \chi^{A_1} + \chi^E + \chi^{F_1} + \chi^{F_2}. \tag{5.59}$$

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.

I have considered only the simplest case here, ignoring the complications introduced by reflection symmetries, and by 2-valued spinor represenations of the rotation group. If you need to understand these, read Hamermesh *op cit.* some of

Chapter 6

Lie Groups

A Lie group[1] is a manifold G equipped with a group multiplication rule $g_1 \times g_2 \to g_3$ which is an smooth function of the g's, as is the operation of taking the inverse of a group element. The most commonly met Lie groups in physics are the infinite families of *matrix groups* $GL(n)$, $SL(n)$, $O(n)$, $SO(n)$, $U(n)$, $SU(n)$, and $Sp(n)$. There is also a 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 dimension of the group and most of the group structure can be understood by examining group elements in 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*, $GL(n, \mathcal{F})$, which consists of invertible $n \times n$ matrices over the field $\mathcal{F}$. We will only consider the cases $\mathcal{F} = \mathbf{C}$ or $\mathcal{F} = \mathbf{R}$.

[1] Named for the Norwegian mathematician Sophus Lie.

A near-identity matrix in $GL(n, \mathbf{R})$ can be written $g = I + \epsilon A$ where A is an arbitrary $n \times n$ real matrix. This matrix contains n^2 real entries, so we can thus 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 $GL(n, \mathbf{C})$ has n^2 *complex* dimensions, and this corresponds to $2n^2$ real dimensions.

If we restrict the determinant of a $GL(n, \mathcal{F})$ matrix to be unity, we get the *special linear group*, $SL(n, \mathcal{F})$. 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 $SL(n, \mathbf{R})$ has dimension $n^2 - 1$.

6.1.1 Unitary Groups and Orthogonal Groups

Perhaps the most important of the matrix groups are the unitary and orthogonal groups.

The Unitary group

The unitary group $U(n)$ is the set of $n \times n$ 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

$$\begin{aligned} I + O(\epsilon^2) &= (I + \epsilon A)(I + \epsilon A^\dagger) \\ &= I + \epsilon(A + A^\dagger) + O(\epsilon^2) \end{aligned} \tag{6.3}$$

and so $A_{ij} = -A^*_{ji}$. The matrix A is therefore skew-hermitian and 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 $n(n-1)/2$ term 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, and this property leads to the group manifold of $U(n)$ being a compact set.

When the 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. This leads to them having many properties in common with the finite groups we studied in the last chapter.

The group $U(n)$ is not simple. Its centre is an invariant $U(1)$ subgroup consisting of matrices of the form $U = e^{i\theta} I$. The *special unitary group* $SU(n)$, consists of $n \times n$ unimodular (having determinant $+1$) unitary matrices. Although not strictly simple (its center, Z, is the discrete subgroup of matrices $U_m = \omega^m I$ with ω an n-th root of unity, and this is obviously an invariant subgroup) it 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 $SU(n)$ group manifold is $n^2 - 1$ dimensional.

The Orthogonal Group

The orthogonal group $O(n)$, is the set of real matrices such that $O^T = O^{-1}$. For an orthogonal matrix in the neighbourhood of the identity, $O = I + \epsilon A$, this condition requires that $A_{ij} = -A_{ij}$. The the group is therefore $n(n-1)/2$ dimensional. The rows or columns are again orthonormal, and thus bounded. This means that $O(n)$ is compact.

Since $1 = \det(O^T O) = \det O^T \det O = (\det O)^2$ we have $\det O = \pm 1$. The set of orthogonal matrices with $\det O = +1$ compose the *special orthogonal group*, $SO(n)$. The unimodularity condition discards a disconnected part of the group manifold, and so does not reduce the dimension of the space which is still $n(n-1)/2$.

6.1.2 Symplectic Groups

The symplectic (from the Greek word meaning to "fold together") groups are slightly more exotic, and merit a more extended discussion. This section should probably be read after the rest of the chapter, because we will use some notations that are defined later.

Let ω be a non-degenerate skew-symmetric matrix. The *symplectic group*, $Sp(2n, \mathcal{F})$ is defined by

$$Sp(2n, \mathcal{F}) = \{S \in GL(2n, \mathcal{F}) : S^T \omega S = \omega\}. \tag{6.4}$$

Here $\mathcal{F}$ is a commutative field, such as $\mathbf{R}$ or $\mathbf{C}$. Note that, even when $\mathcal{F} = \mathbf{C}$, we still use the transpose "T", not †, in this definition. Setting $S = I_{2n} + \epsilon A$, and plugging into the definition shows that $A^T\omega + \omega A = 0$.

We can always reduce ω to its canonical form

$$\omega = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}. \tag{6.5}$$

Having done so, then A short computation shows that the most general form for A is

$$A = \begin{pmatrix} a & b \\ c & -a^T \end{pmatrix}, \tag{6.6}$$

where a is any $n \times n$ matrix, and $b^T = b$, $c^T = c$. If we assume that the matrices are real, then counting the degrees of freedom gives the dimension of the group as

$$\dim Sp(2n, \mathbf{R}) = n^2 + 2 \times \frac{n}{2}(n+1) = n(2n+1). \tag{6.7}$$

The entries in a, b, c can be arbitrarily large, so $Sp(2n, \mathbf{R})$ is not compact.

The determinant of any symplectic matrix is $+1$. To see this take the elements of ω be ω_{ij}, and let

$$\omega(x, y) = \omega_{ij} x^i y^j \tag{6.8}$$

be the associated skew bilinear form (*not* sesquilinear!). Then Weyl's identity

$$\begin{aligned} &\mathrm{Pf}\,(\omega) \det |x_1, x_2, \ldots x_{2n}| \\ &\quad = \frac{1}{2^n n!} \sum_{\pi \in S_{2n}} \mathrm{sgn}\,(\pi) \omega(x_{\pi(1)}, x_{\pi(2)}) \ldots \omega(x_{\pi(2n-1)}, x_{\pi(2n)}), \end{aligned} \tag{6.9}$$

shows that

$$\begin{aligned} &\mathrm{Pf}\,(\omega)\,(\det M) \det |x_1, x_2, \ldots x_{2n}| \\ &\quad = \frac{1}{2^n n!} \sum_{\pi \in S_{2n}} \mathrm{sgn}\,(\pi) \omega(Mx_{\pi(1)}, Mx_{\pi(2)}) \ldots \omega(Mx_{\pi(2n-1)}, Mx_{\pi(2n)}), \end{aligned}$$

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 $Sp(2n, \mathcal{F})$ are automatically unimodular there is no "special symplectic" group.

Unitary Symplectic Group

The intersection of two groups is also a group. We can therefore define the *unitary symplectic group* as

$$Sp(n) = Sp(2n, \mathbf{C}) \cap U(2n). \tag{6.10}$$

This group is compact. We will soon see that its dimension is $n(2n+1)$, the same as the non-compact $Sp(2n, \mathbf{R})$. The group $Sp(n)$ may also be defined as $U(n, \mathbf{H})$ where $\mathbf{H}$ are the quaternions.
Warning: Physics papers often make no distinction between $Sp(n)$, which is a compact group, and $Sp(2n, \mathbf{R})$ which is non-compact. To add to the confusion the compact $Sp(n)$ is also sometimes called $Sp(2n)$. You have to judge from the context which group the author means.
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, single-electron Hamiltonian containing $\mathbf{L} \cdot \mathbf{S}$ spin-orbit interactions obeys

$$C^{-1}HC = H^*. \tag{6.12}$$

If we regard H as being and $n \times n$ matrix of 2×2 matrices

$$H_{ij} = h_{ij}^0 + i \sum_{n=1}^{3} h_{ij}^n \hat{\sigma}_n,$$

then this 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 orthogonal, therefore all states are doubly degenerate. This is *Kramers'* degeneracy.

H may be diagonalized by an element of $U(n, \mathbf{H})$, that is an element of $U(2n)$ obeying $C^{-1}UC = U^*$. We may rewrite this condition as

$$C^{-1}UC = U^* \Rightarrow UCU^T = C,$$

therefore $U(n, \mathbf{H})$ is a unitary matrix which preserves the skew bilinear matrix C and is an element of $Sp(n)$. Further investigation shows that $U(n, \mathbf{H}) = Sp(n)$.

We can exploit the quaternionic viewpoint to count the dimensions. Let $U = I + \epsilon B$ be in $U(n, \mathbf{H})$, 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

$$\dim U(n, \mathbf{H}) = \dim Sp(n) = 3n + 4 \times \frac{n}{2}(n-1) = n(2n+1). \tag{6.15}$$

Thus, as promised, we see that the compact group $Sp(n)$ and the non-compact group $Sp(2n, \mathbf{R})$ have the same dimension.

We can also count the dimension of $Sp(n)$ 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

$$\dim Sp(n) = (n^2) + n(n+1) = n(2n+1)$$

as before.

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 2×2 unitary matrix can be written

$$U = \begin{pmatrix} x^0 + ix^3 & ix^1 + x^2 \\ ix^1 - x^2 & x^0 - ix^3 \end{pmatrix}. \tag{6.16}$$

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 in addition the x^i are real, we have $U^\dagger = U^{-1}$. The group manifold of $SU(2)$ is therefore 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 system is only good for one-half of the three-sphere, but this is typical when we have a non-trivial manifold. Other co-ordinate patches can be constructed as needed.

We can simplify our notation by introducing 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

$$[\hat{\sigma}_i, \hat{\sigma}_j] = 2i\epsilon_{ijk}\hat{\sigma}_k. \tag{6.19}$$

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 $e = I$ by real linear combinations of the $i\hat{\sigma}_i$. The three-dimensional vector space spanned by these matrices is therefore the tangent space TM_e at the identity element. For any Lie group this tangent space is called the *Lie algebra*, $\mathcal{G} = \mathrm{Lie}\, G$ 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

$$[\hat{\lambda}_i, \hat{\lambda}_j] = if_{ij}{}^k\hat{\lambda}_k \tag{6.22}$$

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 we usually desire the $\hat{\lambda}_i$ to be hermitian. They are usually absent in books written for mathematicians.

6.2.1 Invariant vector fields

Consider a group element, $I+\epsilon\hat{L}$, in the neighbourhood of the identity, with $\hat{L} = a^i(i\hat{\sigma}_i)$. We can map this infinitesimal element to the neighbourhood an arbitrary group element g by multiplying on the left to get $g(I+\epsilon\hat{L})$. For example, with $\hat{L}_3 = i\hat{\sigma}_3$, we find

$$\begin{aligned} g(I+\epsilon\hat{L}_3) &= (x^0 + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3)(I + i\epsilon\hat{\sigma}_3) \\ &= (x^0 - \epsilon x^3) + i\hat{\sigma}_1(x^1 - \epsilon x^2) + i\hat{\sigma}_2(x^2 + \epsilon x^1) + i\hat{\sigma}_3(x^3 + \epsilon x^0) \end{aligned} \tag{6.23}$$

Another way of looking at this process is that multiplication of any element g on the right by $(I+\epsilon\hat{L}_3)$ moves g, and so changes its co-ordinates by an amount

$$\delta \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} = \epsilon \begin{pmatrix} -x^3 \\ -x^2 \\ x^1 \\ x^0 \end{pmatrix}. \tag{6.24}$$

This suggests the introduction of the *left-invariant vector field*

$$L_3 = -x^2\partial_1 + x^1\partial_2 + x^0\partial_3. \tag{6.25}$$

Similarly we define

$$\begin{aligned} L_1 &= x^0\partial_1 - x^3\partial_2 + x^2\partial_3 \\ L_2 &= x^3\partial_1 + x^0\partial_2 - x^1\partial_3. \end{aligned} \tag{6.26}$$

These are "left invariant" because the push-forward of the vector $L_i(g_0)$ at g_0 by multiplication on the left by any g produces a vector $g_*[L_i(g_0)]$ at gg_0 that coincides with the $L_i(gg_0)$ already at that point. We can express this statement tersely as $g_*L_i = L_i$.

Using $\partial_i x^0 = -x^i/x_0$, we can compute the Lie brackets and find

$$[L_1, L_2] = -2L_3. \tag{6.27}$$

In general

$$[L_i, L_j] = -2\epsilon_{ijk}L_k. \tag{6.28}$$

This construction works for all matrix groups. For each basis element $\hat{L}_i = i\hat{\lambda}_i$ of the Lie algebra we multiply group elements on the right by$I + i\epsilon\hat{L}_i$ and

so construct the corresponding left-invariant vector field L_i. The Lie bracket of these vector fields will be

$$[L_i, L_j] = -f_{ij}{}^k L_k, \tag{6.29}$$

which coincides with the commutator of the matrices $\hat{L}_i$. The coefficients $f_{ij}{}^k$ are guaranteed to be position independent because 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.

The Exponential Map

Given any vector field, X, we can define the flow along it by solving the equation

$$\frac{dx^\mu}{dt} = X^\mu(x(t)). \tag{6.30}$$

If we do this for the left-invariant vector field L, with $x(0) = e$, we get the element denoted by $g(x(t)) = \mathrm{Exp}\,(tL)$. The symbol "Exp" stands for the *exponential map* which takes us from elements of the Lie algebra to elements of the group. The reason for this 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 elements of the Lie algebra: if $\hat{L} = ia^i\hat{\lambda}_i$, then

$$g(t) = \exp(t\hat{L}), \tag{6.31}$$

is an element of the group and

$$\frac{d}{dt}g(t) = \hat{L}g(t). \tag{6.32}$$

Right-invariant vector fields

We can repeat the exercise of the previous section, multiplying the infinitesimal group element $(I + \epsilon\hat{R})$ in from the left instead. For $\hat{R} = i\hat{\sigma}_3$, for example,

$$\begin{aligned}(I + \epsilon\hat{R}_3)g &= (I + i\epsilon\hat{\sigma}_3)(x^0 + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3) \\ &= (x^0 - \epsilon x^3) + i\hat{\sigma}_1(x^1 + \epsilon x^2) + i\hat{\sigma}_2(x^2 - \epsilon x^1) + i\hat{\sigma}_3(x^3 + \epsilon x^0)\end{aligned} \tag{6.33}$$

This motion corresponds to the *right-invariant vector field*

$$R_3 = x^2\partial_1 - x^1\partial_2 + x^0\partial_3. \tag{6.34}$$

Again, we can also define

$$\begin{aligned} R_1 &= x^3\partial_1 - x^0\partial_2 + x^1\partial_3 \\ R_2 &= x^0\partial_1 + x^3\partial_2 - x^2\partial_3. \end{aligned} \tag{6.35}$$

We find that

$$[R_1, R_2] = +2R_3, \tag{6.36}$$

or, in general,

$$[R_i, R_j] = +2\epsilon_{ijk}R_k. \tag{6.37}$$

For a general Lie group, the Lie brackets of the right-invariant fields will be

$$[R_i, R_j] = +f_{ij}{}^k R_k. \tag{6.38}$$

whenever

$$[L_i, L_j] = -f_{ij}{}^k L_k, \tag{6.39}$$

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 mechanics.

6.2.2 Maurer-Cartan Forms

If $g \in G$, then $dgg^{-1} \in \mathrm{Lie}\, G$. For example, starting from

$$\begin{aligned} g &= x^0 + ix^1\hat{\sigma}_1 + ix^2\hat{\sigma}_2 + ix^3\hat{\sigma}_3 \\ g^{-1} &= x^0 - ix^1\hat{\sigma}_1 - ix^2\hat{\sigma}_2 - ix^3\hat{\sigma}_3 \end{aligned} \tag{6.40}$$

we have

$$\begin{aligned} dg &= dx^0 + idx^1\hat{\sigma}_1 + idx^2\hat{\sigma}_2 + idx^3\hat{\sigma}_3 \\ &= (x^0)^{-1}(-x^1dx^1 - x^2dx^2 - x^3dx^3) + idx^1\hat{\sigma}_1 + idx^2\hat{\sigma}_2 + idx^3\hat{\sigma}_3. \end{aligned} \tag{6.41}$$

From this we find

$$\begin{aligned} dgg^{-1} &= i\hat{\sigma}_1\left((x^0+(x^1)^2/x^0)dx^1+(x^3+(x^1x^2)/x^0)dx^2+(-x^2+(x^1x^3)/x^0)dx^3\right) \\ &\quad +i\hat{\sigma}_2\left((-x^3+(x^2x^1)/x^0)dx^1+(x^0+(x^2)^2/x^0)dx^2+(x^1+(x^2x^3)/x^0)dx^3\right) \\ &\quad +i\hat{\sigma}_3\left((x^2+(x^3x^1)/x^0)dx^1+(-x^1+(x^3x^2)/x^0)dx^2+(x^0+(x^3)^2/x^0)dx^3\right) \end{aligned} \tag{6.42}$$

and we see that 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

$$dgg^{-1} = \omega_R = (i\hat{\sigma}_i)\omega_R^i. \tag{6.43}$$

We evaluate

$$\begin{aligned} \omega_R^1(R_1) &= (x^0+(x^1)^2/x^0)x^0+(x^3+(x^1x^2)/x^0)x^3+(-x^2+(x^1x^3)/x^0)(-x^2) \\ &= (x^0)^2+(x^1)^2+(x^2)^2+(x^3)^2 \\ &= 1. \end{aligned} \tag{6.44}$$

Working similarly we find

$$\begin{aligned} \omega_R^1(R_2) &= (x^0+(x^1)^2/x^0)(-x^3)+(x^3+(x^1x^2)/x^0)x^0+(-x^2+(x^1x^3)/x^0)x^1 \\ &= 0. \end{aligned} \tag{6.45}$$

In general we will discover that $\omega_R^i(R_j) = \delta_j^i$, and so these Maurer Cartan forms constitute the dual basis to the right-invariant vector fields.

We may also define

$$g^{-1}dg = \omega_L = (i\hat{\sigma}_i)\omega_L^i, \tag{6.46}$$

and discover that $\omega_L^i(L_j) = \delta_j^i$. The ω_L are therfore the dual basis to the left-invariant vector fields.

Now acting with the exterior derivative d on $gg^{-1} = I$ tells us that $d(g^{-1}) = -g^{-1}dgg^{-1}$. Using this 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

$$d\omega_R = d(dgg^{-1}) = (dgg^{-1})\wedge(dgg^{-1}) = \omega_R\wedge\omega_R. \tag{6.47}$$

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

$$\begin{aligned} \omega_R \wedge \omega_R &= \omega_R^i \wedge \omega_R^j (i\hat{\sigma}_i)(i\hat{\sigma}_j) \\ &= \frac{1}{2}\omega_R^i \wedge \omega_R^j [i\hat{\sigma}_i, i\hat{\sigma}_j] \\ &= -\frac{1}{2} f_{ij}{}^k (i\hat{\sigma}_k)\, \omega_R^i \wedge \omega_R^j, \end{aligned} \tag{6.48}$$

so

$$d\omega_R^k = -\frac{1}{2} f_{ij}{}^k \omega_R^i \wedge \omega_R^j. \tag{6.49}$$

These equations are known as the *Maurer-Cartan relations* for the right-invariant forms.

For the left-invariant forms we have

$$d\omega_L = d(g^{-1}dg) = -(g^{-1}dg) \wedge (g^{-1}dg) = -\omega_L \wedge \omega_L \tag{6.50}$$

or

$$d\omega_L^k = +\frac{1}{2} f_{ij}{}^k \omega_L^i \wedge \omega_L^j. \tag{6.51}$$

These 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

Physicists often Use Euler angles to parameterize $SU(2)$. We write an arbitrary $SU(2)$ unitary matrix U as

$$\begin{aligned} U &= \exp\{-i\phi\hat{\sigma}_3/2\} \exp\{-i\theta\hat{\sigma}_2/2\} \exp\{-i\psi\hat{\sigma}_3/2\}, \\ &= \begin{pmatrix} e^{-i\phi/2} & 0 \\ 0 & e^{i\phi/2} \end{pmatrix} \begin{pmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{pmatrix} \begin{pmatrix} e^{-i\psi/2} & 0 \\ 0 & e^{i\psi/2} \end{pmatrix}, \\ &= \begin{pmatrix} e^{-i(\phi+\psi)/2}\cos\theta/2 & -e^{i(\psi-\phi)/2}\sin\theta/2 \\ e^{i(\phi-\psi)/2}\sin\theta/2 & e^{+i(\psi+\phi)/2}\cos\theta/2 \end{pmatrix}. \end{aligned} \tag{6.52}$$

Comparing with the earlier expression for U in terms of the x^μ, we obtain the Euler-angle parameterization of the three-sphere

$$x^0 = \cos\theta/2 \cos(\psi+\phi)/2,$$

$$\begin{aligned} x^1 &= \sin\theta/2\sin(\phi-\psi)/2, \\ x^2 &= -\sin\theta/2\cos(\phi-\psi)/2, \\ x^3 &= -\cos\theta/2\sin(\psi+\phi)/2. \end{aligned} \tag{6.53}$$

The ranges of the angles can be taken to be $0 \leq \phi < 2\pi$, $0 \leq \theta < \pi$, $0 \leq \psi < 4\pi$.

Exercise: 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.

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 later) 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

$$\begin{aligned} ds^2 &= (dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2, \\ &= \frac{1}{4}\left(d\theta^2 + \cos^2\theta/2(d\psi+d\phi)^2 + \sin^2\theta/2(d\psi-d\phi)^2\right), \\ &= \frac{1}{4}\left(d\theta^2 + d\psi^2 + d\phi^2 + 2\cos\theta d\phi d\psi\right). \end{aligned} \tag{6.54}$$

Here I've used the traditional physics way of writing a metric. In the more formal notation from chapter one, where we think of the metric as being a bilinear function, we would write the last line as

$$\mathbf{g}(\ ,\ ) = \frac{1}{4}[d\theta\otimes d\theta + d\psi\otimes d\psi + d\phi\otimes d\phi + \cos\theta(d\phi\otimes d\psi + d\psi\otimes d\phi)] \tag{6.55}$$

From this we find

$$\begin{aligned} g = \det(g_{\mu\nu}) &= \frac{1}{4^3}\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & \cos\theta \\ 0 & \cos\theta & 1 \end{vmatrix} \\ &= \frac{1}{64}(1-\cos^2\theta) = \frac{1}{64}\sin^2\theta. \end{aligned} \tag{6.56}$$

The volume element, $\sqrt{g}\, d\theta d\phi d\psi$, is therefore

$$d(Volume) = \frac{1}{8}\sin\theta d\theta d\phi d\psi, \tag{6.57}$$

and the total volume of the sphere is

$$Vol(S^3) = \frac{1}{8}\int_0^{\pi}\sin\theta d\theta \int_0^{2\pi} d\phi \int_0^{4\pi} d\psi = 2\pi^2. \tag{6.58}$$

This coincides with the standard expression for the volume of S^{d-1}, the surface of the d-dimensional unit ball,

$$Vol(S^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}, \tag{6.59}$$

when $d = 4$.

Exercise: Evaluate the Maurer-Cartan form $\omega_L^3 = \mathrm{tr}\,(\sigma_3 g^{-1}dg)$ in terms of the Euler angle parameterization and show that

$$\omega_L^3 = i(-d\psi - \cos\theta d\phi). \tag{6.60}$$

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 (θ, ϕ). Thus, if we set $\omega_L^3 = i\eta$, then

$$d\eta = \sin\theta\, d\theta\, d\phi = i\,\mathrm{Hopf}^*(d[Area\, S^2]). \tag{6.61}$$

Also observe that

$$\eta \wedge d\eta = -\sin\theta\, d\theta\, d\phi\, d\psi. \tag{6.62}$$

From this show that

$$\frac{1}{16\pi^2}\int_{S^3}\eta \wedge d\eta = -1. \tag{6.63}$$

6.2.5 $SO(3) \simeq SU(2)/\mathbf{Z}_2$

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 $SU(2)$. 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 Cartan in 1913, long before they were needed in physics.

The simplest way to motivate the spinor/rotation connection is via the Pauli matrices. The sigma matrices are hermitian, traceless, and obey

$$\hat{\sigma}_i\hat{\sigma}_j + \hat{\sigma}_j\hat{\sigma}_i = 2\delta_{ij}, \tag{6.64}$$

If, for any $U \in SU(2)$, we define

$$\hat{\sigma}'_i = U\hat{\sigma}_i U^{-1} \tag{6.65}$$

we see that the $\hat{\sigma}'_i$ have exactly the same properties. Since the original $\hat{\sigma}_i$ form a basis for the space of hermitian traceless matrices, we must have

$$\hat{\sigma}'_i = \hat{\sigma}_j A_{ji} \tag{6.66}$$

for some real 3×3 matrix A_{ij}. From (6.64) we find that

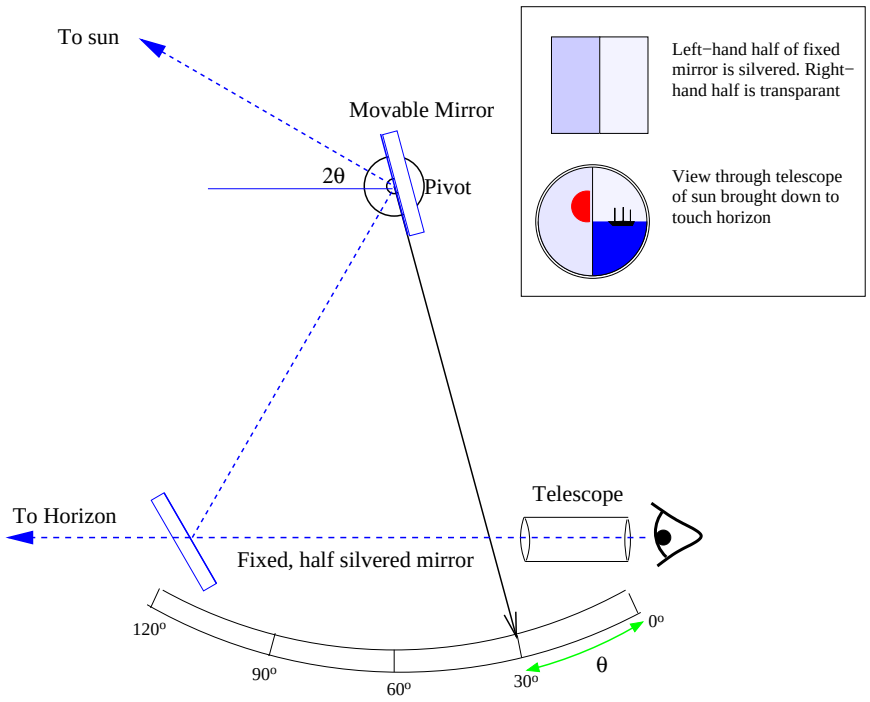
$$\begin{aligned} 2\delta_{ij} &= \hat{\sigma}'_i\hat{\sigma}'_j + \hat{\sigma}'_j\hat{\sigma}'_i \\ &= (\hat{\sigma}_l A_{li})(\hat{\sigma}_m A_{mj}) + (\hat{\sigma}_m A_{mj})(\hat{\sigma}_l A_{li}) \\ &= (\hat{\sigma}_l\hat{\sigma}_m + \hat{\sigma}_m\hat{\sigma}_l)A_{li}A_{mj} \\ &= 2\delta_{lm}A_{li}A_{mj}, \end{aligned}$$

so

$$A_{mi}A_{mk} = \delta_{ik}. \tag{6.67}$$

In other words $A^T A = I$, and A is an element of $O(3)$. The determinant of any orthogonal matrix is ± 1, but $SU(2)$ is simply connected, and $A = I$, when $U = I$. Continuity therefore tells us that $\det A = 1$. The A matrices are therefore in $SO(3)$.

By exploiting the principle of the sextant we may construct a $U(R)$ for any element $R \in SO(3)$.



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 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 object $\hat{\mathbf{x}} = x^i \hat{\sigma}_i$. Now, if $\mathbf{n}$ is a unit vector, then

$$(-\hat{\sigma}_i n^i)(x^j \hat{\sigma}_j)(\hat{\sigma}_k n^k) = \left(x^j - 2(\mathbf{n}\cdot\mathbf{x})(n^j)\right)\hat{\sigma}_j \tag{6.68}$$

is the $\mathbf{x}$ vector reflected in the plane perpendicular to $\mathbf{n}$. So, for example

$$-(\hat{\sigma}_1 \cos\theta/2 + \hat{\sigma}_2 \sin\theta/2)(-\hat{\sigma}_1)\hat{\mathbf{x}}(\hat{\sigma}_1)(\hat{\sigma}_1 \cos\theta/2 + \hat{\sigma}_2 \sin\theta/2) \tag{6.69}$$

performs two succesive reflections, first in the "1" plane, and then in a plane at an angle $\theta/2$ to it. Multiplying the factors, and using the $\hat{\sigma}_i$ algebra, we find

$$\begin{aligned}&(\cos\theta/2 - \hat{\sigma}_1\hat{\sigma}_2 \sin\theta/2)\hat{\mathbf{x}}(\cos\theta/2 + \hat{\sigma}_1\hat{\sigma}_2 \sin\theta/2)\\ &\quad = \hat{\sigma}_1(\cos\theta\, x^1 - \sin\theta\, x^2) + \hat{\sigma}_2(\sin\theta\, x^1 + \cos\theta\, x^2) + \hat{\sigma}_3 x^3,\end{aligned}$$

and this is a rotation through θ as claimed. We can write this as

$$e^{-i\frac{1}{4i}[\hat{\sigma}_1,\hat{\sigma}_2]\theta}(x^i\hat{\sigma}_i)e^{i\frac{1}{4i}[\hat{\sigma}_1,\hat{\sigma}_2]\theta} = e^{-i\hat{\sigma}_3\theta/2}(x^i\hat{\sigma}_i)e^{i\hat{\sigma}_3\theta/2} = \hat{\sigma}_j R_{ji}x^i, \tag{6.70}$$

where R is the 3×3 rotation matrix for a rotation through angle θ in the 1-2 plane. It should be clear that this construction allows *any* rotation to be performed. 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 fruit of our labours is a two-dimensional unitary matrix , $U(R)$, such that

$$U(R)\hat{\sigma}_i U^{-1}(R) = \hat{\sigma}_j R_{ji}, \tag{6.71}$$

for any $R \in SO(3)$. This $U(R)$ is the *spinor* represenation of the rotation group.

Exercise: Verify that $U(R_2)U(R_1) = U(R_2R_1)$ and observe that we *must* write the R on the right, for this composition to work.

If $U(R) \in SU(2)$, so is $-U(R)$, and $U(R)$ and $-U(R)$ give exactly the same rotation R. The mapping between $SU(2)$ and $SO(3)$ is $2 \to 1$, and 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 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 is not contractable to a point. If we continue on from the antipode back to the original point, the combined path *is* contractable. Expressing these facts mathematically, we say that the first *Homotopy group*, the group of based paths with composition given by concatenation, is $\pi_1(SO(3)) = \mathbf{Z}_2$. This is the topology behind the Phillipine (or Balinese) Candle Dance, and how the electron knows whether a sequence of rotations that eventually bring it back to its original orientation should be counted as a 2π rotation ($U = -I$) or a $4\pi \equiv 0$ rotation ($U = +I$).

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 *Clifford algebra*

$$\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\delta_{\mu\nu}. \tag{6.72}$$

This is a generalization of the key algebraic property of the Pauli sigma matrices.

If $N\,(=2n)$ is even, then we can find $2^n \times 2^n$ matrices, $\hat{\gamma}_\mu$, satisfying this algebra. If $N\,(=2n+1)$ is odd, we append to the matrices for $N=2n$ the matrix $\hat{\gamma}_{2n+1} = -(i)^n \hat{\gamma}_1 \hat{\gamma}_2 \cdots \hat{\gamma}_n$. 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

$$e^{-i\frac{1}{4i}[\hat{\gamma}_m,\hat{\gamma}_n]\theta}(x^i\hat{\gamma}_i)e^{i\frac{1}{4i}[\hat{\gamma}_m,\hat{\gamma}_n]\theta} = \hat{\gamma}_j R_{ji} x^i, \tag{6.73}$$

and we find that the matrices $\hat{\Gamma}_{mn} = \frac{1}{4i}[\hat{\gamma}_m, \hat{\gamma}_n]$ obey the lie algebra of $SO(N)$. The $2^{[N/2]}$ dimensional space on which they act is the spinor representation of $SO(N)$.

If N is even then we can still construct the matrix $\hat{\gamma}_{2n+1}$ and find that it anticommutes with all the other $\hat{\gamma}$'s. It cannot be the identity matrix, therefore, but it still commutes with all the Γ_{mn}. By Schur's lemma, this means that the $SO(2n)$ spinor representation space V is *reducible*. Now $\gamma_{2n+1}^2 = I$, and so γ_{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

$$V = V_{odd} \oplus V_{even}. \tag{6.74}$$

Here V_{even} and V_{odd}, the plus and minus eigenspaces of γ_{2n+1}, are called the spaces of right and left *chirality*. When N is odd the spinor representation is irreducible.

The Adjoint Representation

The idea of obtaining a representation by conjugation works for an arbitrary Lie group. Given an infinitesimal element $I + \epsilon\hat{L}$, the conjugate element

$g(I + \epsilon\hat{L})g^{-1}$ will also be an infinitesimal element. This means that $g\hat{L}_i g^{-1}$ must be expressible as a linear combination of the $\hat{L}_i$ matrices. Consequently we can define a linear map acting on the element $X = X^i\hat{L}_i$ of the Lie algebra by setting

$$\mathrm{Ad}(g)\hat{L}_i \equiv g\hat{L}_i g^{-1} = \hat{L}_j(\mathrm{Ad}\,(g))^j{}_i.$$

The matrices $(\mathrm{Ad}\,(g))^j{}_i$ form the *adjoint* representation of the group. The dimension of the adjoint representation coincides with that of the group.

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

$$\int d[g]\left(D^J_{ij}(g)\right)^* D^K_{lm}(g) = \frac{1}{\dim J}\delta^{JK}\delta_{il}\delta_{jm}.$$

The Peter-Weyl theorem asserts that the representation matrices, $D^J_{mn}(g)$, form a complete set of orthogonal function on the group manifold. In the case of $SU(2)$ this tells us that the spin J representation matrices

$$\begin{aligned} D^J_{mn}(\theta,\phi,\psi) &= \langle J,m|e^{-iJ_3\phi}e^{-iJ_2\theta}e^{-iJ_3\psi}|J,n\rangle, \\ &= e^{-im\phi}d^J_{mn}(\theta)e^{-in\psi}, \end{aligned}$$

which you will know from quantum mechanics courses[2], are a complete set of functions on the three-sphere with

$$\frac{1}{16\pi^2}\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi\int_0^{4\pi}d\psi\left(D^J_{mn}(\theta,\phi,\psi)\right)^* D^{J'}_{m'n'}(\theta,\phi,\psi)$$
$$=\frac{1}{2J+1}\delta^{JJ'}\delta_{mm'}\delta_{nn'}.$$

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

$$\frac{1}{4\pi}\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi}d\phi\left(D^L_{m0}(\theta,\phi)\right)^* D^{L'}_{m'0}(\theta,\phi)=\frac{1}{2L+1}\delta^{LL'}\delta_{mm'}.$$

Comparing with the definition of the spherical harmonics, we see that we can identify

$$Y^L_m(\theta,\phi)=\sqrt{\frac{2L+1}{4\pi}}\left(D^L_{m0}(\theta,\phi,\psi)\right)^*.$$

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, $\chi^J(g)=D^J_{nn}(g)$ 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 have

$$\begin{aligned}\chi(\theta) &= e^{iJ\theta}+e^{i(J-1)\theta}+\cdots+e^{-i(J-1)\theta}+e^{-iJ\theta},\\ &= \frac{\sin(2J+1)\theta/2}{\sin\theta/2}.\end{aligned}$$

Warning: The angle θ in this formula 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

$$\frac{1}{\pi}\int_0^{2\pi}\chi^J(\theta)\chi^{J'}(\theta)\sin^2\left(\frac{\theta}{2}\right)d\theta=\delta^{JJ'},$$

[2]See, for example, G. Baym *Lectures on Quantum Mechanics*, Ch 17.

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: 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

$$e^{-t_2\hat{Y}}e^{-t_1\hat{X}}e^{t_2\hat{Y}}e^{t_1\hat{X}} = I - t_1t_2[\hat{X}, \hat{Y}] + \cdots,$$

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:

$$L_x: \quad \left\{\begin{array}{rcl}\delta y &=& -z\,\delta\theta_x\\ \delta z &=& +y\,\delta\theta_x\end{array}\right\} \Longrightarrow L_x = y\partial_z - z\partial_y, \quad \hat{L}_x = \begin{pmatrix}0&0&0\\0&0&-1\\0&1&0\end{pmatrix},$$

$$L_y: \quad \left\{\begin{array}{rcl}\delta z &=& -x\,\delta\theta_y\\ \delta x &=& +z\,\delta\theta_y\end{array}\right\} \Longrightarrow L_y = z\partial_x - x\partial_z, \quad \hat{L}_y = \begin{pmatrix}0&0&1\\0&0&0\\-1&0&0\end{pmatrix},$$

$$L_z: \quad \left\{\begin{array}{rcl}\delta x &=& -y\,\delta\theta_z\\ \delta y &=& +x\,\delta\theta_z\end{array}\right\} \Longrightarrow L_z = x\partial_y - y\partial_x, \quad \hat{L}_y = \begin{pmatrix}0&-1&0\\1&0&0\\0&0&0\end{pmatrix}.$$

From this we find

$$[L_x, L_y] = -L_z,$$

as a Lie bracket of vector fields, but

$$[\hat{L}_x, \hat{L}_y] = +\hat{L}_z,$$

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 invariant fields on the representation matrices, $D^J_{mn}(g)$. For example

$$\begin{aligned} L_i D^J_{mn}(g) &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn}(g(1+i\epsilon\hat{\lambda}_i)) - D^J_{mn}(g)\right)\right] \\ &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn'}(g)D^J_{n'n}(1+i\epsilon\hat{\lambda}_i) - D^J_{mn}(g)\right)\right] \\ &= \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\left(D^J_{mn'}(g)(\delta_{n'n} + i\epsilon(\hat{\Lambda}^J_i)_{n'n}) - D^J_{mn}(g)\right)\right] \\ &= D^J_{mn'}(g)(i\hat{\Lambda}^J_i)_{n'n} \end{aligned} \tag{6.75}$$

where $\hat{\Lambda}^J_i$ is the matrix representing $\hat{\lambda}_i$ in the representation J. Repeating this exercise we find that

$$L_i\left(L_j D^J_{mn}(g)\right) = D^J_{mn''}(g)(i\hat{\Lambda}^J_i)_{n''n'}(i\hat{\Lambda}^J_j)_{n'n},$$

Thus

$$[L_i, L_j]D^J_{mn}(g) = D^J_{mn'}(g)[i\hat{\Lambda}^J_i, i\hat{\Lambda}^J_j]_{n'n},$$

and we get the commutator of the representation matrices in the right order only if we multiply 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 Abstract Lie Algebras

A Lie algebra $\mathcal{G}$ is a (real or complex) vector space with a non-associative binary operation $\mathcal{G}\times\mathcal{G}\to\mathcal{G}$ 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 $n\times n$ matrices. As a vector space this is n^2 dimensional. Setting $[A, B] = AB - BA$, makes $M(n)$ into a Lie Algebra.

Example: Let b^+ denote the subset of $M(n)$ consisting of upper triangular matrices with anything allowed on the diagonal. Then b^+ with the above bracket is a Lie algebra. (The "b" stands for *Borel*).
Example: Let n^+ denote the subset of b^+ consisting of strictly upper triangular matrices — those with zero on the diagonal. Then 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

$$[L_i, L_j] = f_{ij}{}^k L_k$$

where $[\ \ ,\ \ ]$ is the Lie bracket of vector fields. The resulting Lie algebra, $\mathcal{G} = \mathrm{Lie}\, G$ is the Lie algebra of the group.
Observation: The set N^+ of upper triangular matrices with 1's on the diagonal forms a Lie group, with n^+ as its Lie algebra. Similarly, the set B^+ consisting of upper triangular matrices with anything allowed on the diagonal, is also a Lie group, and has 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 invariant subgroups. This is provided by the notion of an *ideal*. A ideal is a subalgebra $\mathcal{I} \subseteq \mathcal{G}$ with the property that

$$[\mathcal{I}, \mathcal{G}] \in \mathcal{I}.$$

That is, taking the bracket of any element of $\mathcal{G}$ with any element of $\mathcal{I}$ gives an element in $\mathcal{I}$. With this definition we can form $\mathcal{G} - \mathcal{I}$ by identifying $X \sim X + I$ for any $I \in \mathcal{I}$. Then

$$[X + \mathcal{I}, Y + \mathcal{I}] = [X, Y] + \mathcal{I},$$

and the bracket of two equivalence classes is insensitive to the choice of representatives. (This is the same definition that is used to define quotient rings.)

If a Lie group G has an invariant subgroup H which is also a Lie group, then the Lie algebra $\mathcal{H}$ of the subgroup is an ideal in $\mathcal{G} = \mathrm{Lie}\, G$ and the Lie algebra of the quotient group G/H is the quotient algebra $\mathcal{G} - \mathcal{H}$.

6.3.1 Adjoint Representation

Given an element $X \in \mathcal{G}$ let it act on the Lie algebra considered as a vector space by a linear map $\mathrm{ad}\,(x)$ defined by

$$\mathrm{ad}\,(X)Y = [X, Y].$$

The Jacobi identity is then equivalent to the statement

$$(\mathrm{ad}\,(X)\mathrm{ad}\,(Y) - \mathrm{ad}\,(Y)\mathrm{ad}\,(X))\,Z = \mathrm{ad}\,([X, Y])Z.$$

Thus

$$(\mathrm{ad}\,(X)\mathrm{ad}\,(Y) - \mathrm{ad}\,(Y)\mathrm{ad}\,(X)) = \mathrm{ad}\,([X, Y]),$$

or

$$[\mathrm{ad}\,(X), \mathrm{ad}\,(Y)] = \mathrm{ad}\,([X, Y]),$$

and the map $X \to \mathrm{ad}\,(X)$ is a representation of the algebra called the *adjoint representation*.

The linear map "$\mathrm{ad}\,(X)$" exponentiates to give a map $\exp[\mathrm{ad}\,(tX)]$ defined by

$$\exp[\mathrm{ad}\,(tX)]Y = Y + t[X, Y] + \frac{1}{2}t^2[X, [X, Y]] + \cdots.$$

You probably know the matrix identity[3]

$$e^{tA}Be^{-tA} = B + t[A, B] + \frac{1}{2}t^2[A, [A, B]] + \cdots.$$

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

$$\mathrm{Ad}\,(\mathrm{Exp}\,Y) = \exp(\mathrm{ad}\,(Y)).$$

6.3.2 The Killing form

Using ad we can define an inner product $\langle \quad , \quad \rangle$ on the Lie algebra by

$$\langle X, Y\rangle = \mathrm{tr}\,(\mathrm{ad}\,(X)\mathrm{ad}\,(Y)).$$

[3]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 also satisfies this equation.

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

$$\langle \mathrm{ad}\,(X)Y, Z\rangle + \langle Y, \mathrm{ad}\,(X)Z\rangle = 0$$

so "$\mathrm{ad}\,(X)$" is skew-symmetric with respect to it. This means, in particular, that

$$\langle e^{\mathrm{ad}\,(X)}Y, e^{\mathrm{ad}\,(X)}Z\rangle = \langle Y, Z\rangle,$$

and the Killing form remains invariant under the action of the adjoint representation on the algebra. When our group is simple, any other invariant inner product will be proportional to this Killing form product.
Definition: If the Killing form is non degenerate, the Lie Algebra is said to be *semi-simple*.

This definition of semi-simplicity is equivalent (although not obviously so) to the definition of a Lie algebra being semi-simple if it contains no Abelian ideal. A semisimple algebra is (again not obviously) the direct sum of simple algebras — those with no ideals except $\{0\}$ and $\mathcal{G}$ itself. Simple and semi-simple algebras are the easiest to study. The Lie algebras b^+ and n^+ are not semi-simple.

Exercise: Show that if $\mathcal{G}$ is a semisimple Lie algebra and $\mathcal{I}$ an ideal, then $\mathcal{I}^\perp$, the orthogonal complement with respect to the Killing form, is also an ideal and

$$\mathcal{G} = \mathcal{I} \oplus \mathcal{I}^\perp.$$

The symbol $\mathcal{G}_1 \oplus \mathcal{G}_2$ denotes a direct sum of the algebras. This implies both a direct sum as vector spaces and the statement $[\mathcal{G}_1, \mathcal{G}_2] = 0$.

Definition: If the Killing form is negative definite, the Lie Algebra is said to be *compact*, and is the Lie algebra of a compact group. (Physicists like to put "i"'s in some of these definitions, so as to make "ad" hermitian, and the Killing form of compact groups positive definite.) The map $\mathrm{Ad}\,(\mathrm{Exp}\,X) : \mathcal{G} \to \mathcal{G}$ is then orthogonal.

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 we find matrices $\hat{L}_i$ obeying the Lie algebra

$$[\hat{L}_i, \hat{L}_j] = i f_{ij}{}^k \hat{L}_k$$

and then the matrices

$$\hat{g} = \exp\left\{ i \sum_i a_i \hat{L}_i \right\}$$

will form a representation of the group, or, to be more precise, a representation of that part of the group which is connected to the identity element. In these equations we have inserted factors of "i" in the locations where they are usually found in physics texts. With these factors, for example, the Lie algebra of $SU(n)$ consists of traceless hermitian matrices instead of skew-hermitian matrices.

SU(2)

The quantum-mechanical angular momentum algebra consists of the commutation relation

$$[J_1, J_2] = i\hbar J_3,$$

together with two similar equations related by cyclic permutations. This is, with $\hbar = 1$, the Lie algebra of $SU(2)$. The goal of representation theory is to find all possible sets of matrices which have the same commutation relations as these operators.

Remember how the 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

$$J_+ = J_1 + iJ_2, \quad J_- = J_1 - iJ_2,$$

which are eigenvectors of $\mathrm{ad}\,(J_3)$

$$\mathrm{ad}\,(J_3) J_\pm = [J_3, J_\pm] = \pm J_\pm.$$

From this 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 $m \pm 1$.

We next assume the existence of 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 this 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

$$\begin{aligned} J_+J_- &= (J_1^2+J_2^2+J_3^2)-(J_3^2-J_3), \\ J_-J_+ &= (J_1^2+J_2^2+J_3^2)-(J_3^2+J_3). \end{aligned}$$

The resulting set of normalized states $|j,m\rangle$ obey

$$\begin{aligned} J_3|j,m\rangle &= m|j,m\rangle, \\ J_-|j,m\rangle &= \sqrt{j(j+1)-m(m-1)}|j,m-1\rangle, \\ J_+|j,m\rangle &= \sqrt{j(j+1)-m(m+1)}|j,m+1\rangle. \end{aligned}$$

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 chose 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 negative-$(norm)^2$ vectors, and the resultant representation cannot be unitary.

This strategy *works for any (semi-simple) Lie algebra*!

SU(3)

Consider, for example, $SU(3)$. The matrix Lie algebra $su(3)$ is spanned by the Gell-Mann λ-matrices

$$\begin{aligned} \hat{\lambda}_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat{\lambda}_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat{\lambda}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ \hat{\lambda}_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \hat{\lambda}_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \hat{\lambda}_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\ \hat{\lambda}_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \hat{\lambda}_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \end{aligned} \tag{6.76}$$

which form a basis for the 3×3 traceless, hermitian matrices. They have been chosen and normalized so that

$$\mathrm{tr}\,(\hat{\lambda}_i\hat{\lambda}_j) = 2\delta_{ij},$$

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

$$\begin{aligned} t_\pm &= \frac{1}{2}(\hat{\lambda}_1 \pm i\hat{\lambda}_2), \\ v_\pm &= \frac{1}{2}(\hat{\lambda}_4 \pm i\hat{\lambda}_5), \\ u_\pm &= \frac{1}{2}(\hat{\lambda}_6 \pm i\hat{\lambda}_7). \end{aligned}$$

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

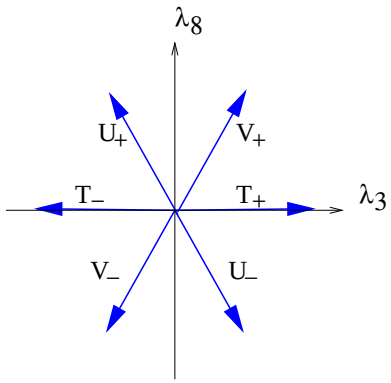
$$\begin{aligned} T_\pm &= \frac{1}{2}(\Lambda_1 \pm i\Lambda_2), \\ V_\pm &= \frac{1}{2}(\Lambda_4 \pm i\Lambda_5), \\ U_\pm &= \frac{1}{2}(\Lambda_6 \pm i\Lambda_7). \end{aligned}$$

These are simultaneous eigenvectors of the commuting pair of operators $\mathrm{ad}\,(\Lambda_3)$ and $\mathrm{ad}\,(\Lambda_8)$:

$$\begin{aligned} \mathrm{ad}\,(\Lambda_3)T_\pm &= [\Lambda_3, T_\pm] = \pm 2T_\pm, \\ \mathrm{ad}\,(\Lambda_3)V_\pm &= [\Lambda_3, V_\pm] = \pm V_\pm, \\ \mathrm{ad}\,(\Lambda_3)U_\pm &= [\Lambda_3, U_\pm] = \mp U_\pm, \\ \mathrm{ad}\,(\Lambda_8)T_\pm &= [\Lambda_8, T_\pm] = 0 \\ \mathrm{ad}\,(\Lambda_8)V_\pm &= [\Lambda_8, V_\pm] = \pm\sqrt{3}V_\pm, \\ \mathrm{ad}\,(\Lambda_8)U_\pm &= [\Lambda_8, U_\pm] = \pm\sqrt{3}U_\pm, \end{aligned}$$

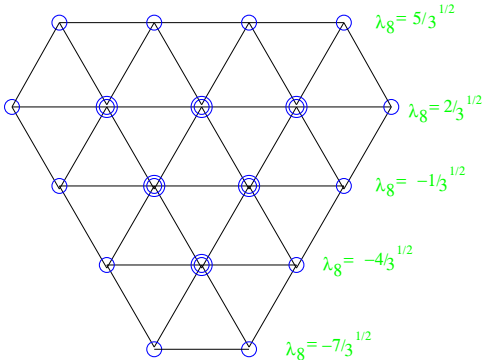
Thus in any representation the $T_\pm$, $U_\pm$, $V_\pm$, act as ladder operators, changing the simultaneous eigenvalues of the commuting pair Λ_3, Λ_8. Their eigenvalues, λ_3, λ_8, are called the *weights*, and there will be a set of such weights for each possible representation. By using the ladder 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.



The root vectors of $su(3)$.

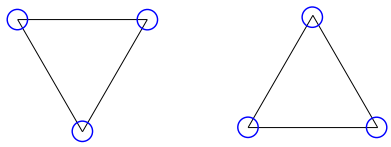
The weights in a representation of $su(3)$ lie on a hexagonal lattice, and the representations are labelled by pairs of integers (zero allowed) p, q which give the length of the sides of the "crystal". These representations have dimension $d = \frac{1}{2}(p+1)(q+1)(p+q+2)$.



The 24 dimensional irrep with $p = 3$, $q = 1$.

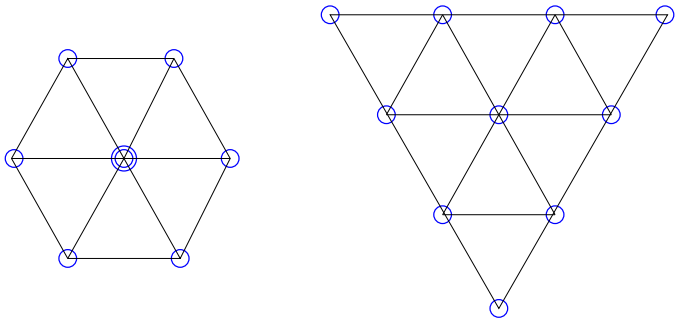
In the figure each circle represents a state with a given weight. A double circle indicates that there are two independent states with this weight, so the total number of weights, and hence the dimension of the representation is 24. In general the degeneracy of the weights increases by one at each "layer", until we reach a triangular inner core all of whose weights have the same degeneracy.

The representations are often labeled by the dimension. The defining representation of $su(3)$ and its complex conjugate are denotes by 3 and $\bar{3}$,



The irreps with $p = 1$, $q = 0$, and $p = 0$, $q = 1$, also known as the 3 and the $\overline{3}$.

while the eight dimensional adjoint represention and the 10 have weights



The irreps 8 (the adjoint) and 10.

For a general simple Lie algebra we play the same game. We find a maximal set of commuting operators, h_i, which make up the *Cartan subalgebra*, $\mathcal{H}$. The number of h_i in this maximally commuting set is called the *rank* of the Lie algbera. We now diagonalize the "ad" action of the h_i on the rest of the algebra. The simultaneous eigenvectors are denoted by e_α where the α, with components α_i, are the *roots*, or root vectors.

$$\text{ad}\,(h_i)e_\alpha = [h_i, e_\alpha] = \alpha_i e_\alpha.$$

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 $su(3)$ 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 $\alpha_i \langle e_\alpha, h_j\rangle = \langle \text{ad}\,(h_i)e_\alpha, h_j\rangle = -\langle e_\alpha, [h_i, h_j]\rangle = 0$ 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 is $[e_\alpha, e_\beta]$ is non-zero, it 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 $\mathcal{H}$ 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 $su(2)$ subalgebra of $\mathcal{G}$. In particular h_α (being the analogue of $2J_3$) has only *integer* eigenvalues. For example in $su(3)$

$$\begin{aligned}
[T_+, T_-] &= h_T = \Lambda_3, \\
[V_+, V_-] &= h_V = \frac{1}{2}\Lambda_3 + \frac{\sqrt{3}}{2}\Lambda_8, \\
[U_+, U_-] &= h_U = -\frac{1}{2}\Lambda_3 + \frac{\sqrt{3}}{2}\Lambda_8,
\end{aligned}$$

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

$$\text{ad}\,(h_\alpha)e_\beta = [h_\alpha, e_\beta] = \frac{2\alpha\cdot\beta}{\alpha^2}e_\beta,$$

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 $\text{ad}\,(e_{-\alpha})e'_\alpha = 0$, and e'_α is killed by the step-down operator. It would therefore be the lowest weight in some $su(2)$ representation. At the same time, however, $\text{ad}\,(h_\alpha)e'_\alpha = 2e'_\alpha$, and we know that the lowest weight in any spin J representation cannot have positive eigenvalue.

The conditions that

$$\frac{2\alpha\cdot\beta}{\alpha^2} \in \mathcal{Z}$$

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 degrees.

These constraints lead to a complete classification of possible Lie algebras into the infinite families

$$\begin{aligned} \mathcal{A}_n, &\quad n = 1, 2, \cdots. && sl(n+1, \mathbf{C}), \\ \mathcal{B}_n, &\quad n = 2, 3, \cdots. && so(2n+1, \mathbf{C}), \\ \mathcal{C}_n, &\quad n = 3, 3, \cdots. && sp(2n, \mathbf{C}), \\ \mathcal{D}_n, &\quad n = 4, 5, \cdots. && so(2n, \mathbf{C}), \end{aligned}$$

together with the exceptional algebras $\mathcal{G}_2$, $\mathcal{F}_4$, $\mathcal{E}_6$, $\mathcal{E}_7$, $\mathcal{E}_8$. These do not correspond to any of the classical matrix algebras. For example $\mathcal{G}_2$ is the 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 $\mathcal{C}_1 = \mathcal{D}_1 = \mathcal{A}_1$. This corresponds to $sp(2, \mathbf{C}) \simeq so(3, \mathbf{C}) \simeq sl(2, \mathbf{C})$. Similarly $\mathcal{D}_2 = \mathcal{A}_1 + \mathcal{A}_1$, corresponding to isomorphism $SO(4) \simeq SU(2) \times SU(2)/Z_2$, while $\mathcal{C}_2 = \mathcal{B}_2$ implies that, locally, the compact $Sp(2) \simeq SO(5)$. Finally $\mathcal{D}_3 = \mathcal{A}_3$ implies that $SU(4)/Z_2 \simeq SO(6)$.

6.3.4 Product Representations

Given two representations $\Lambda_i^{(1)}$ and $\Lambda_i^{(2)}$ of $\mathcal{G}$, we can form a new representation that exponentiates to the tensor product of the corresponding representations of the group G. We set

$$\Lambda_i^{(1\otimes 2)} = \Lambda_i^{(1)} \otimes I + I \otimes \Lambda_i^{(2)}.$$

This process 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

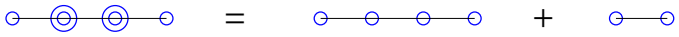
$$\begin{aligned} \Lambda_i^{(1\otimes 2)}|m^{(1)}, m^{(2)}\rangle &= (\Lambda_i^{(1)} \otimes 1 + 1 \otimes \Lambda_i^{(2)})|m^{(1)}, m^{(2)}\rangle \\ &= (m_i^{(1)} + m_i^{(2)})|m^{(1)}, m^{(2)}\rangle \end{aligned}$$

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 $j > j'$,

$$j \otimes j' = (j + j') \oplus (j + j' - 1) \oplus \cdots (j - j').$$

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

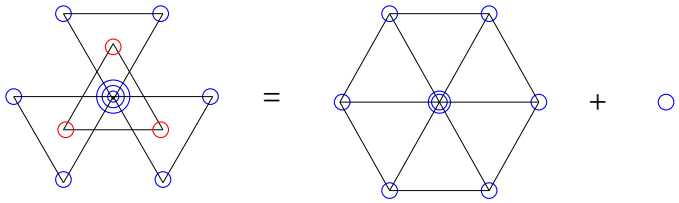


The weights for $1/2 \otimes 1 = 3/2 \oplus 1/2$.

The rules for decomposing products in other groups are more complicated than for $SU(2)$, but can be obtained from weight diamgrams in the same manner. In $SU(3)$, we have, for example

$$\begin{aligned} 3 \otimes \bar{3} &= 1 \oplus 8, \\ 3 \otimes 8 &= 3 \oplus \bar{6} \oplus 15, \\ 8 \otimes 8 &= 1 \oplus 8 \oplus 8 \oplus 10 \oplus \overline{10} \oplus 27. \end{aligned}$$

To illustrate the first of these we consider adding the weights for the $\bar{3}$ (blue) to each of the weights in the 3 (red)



The resultant weights decompose (uniquely) into the weight diagrams for the 8 together with a singlet.

Chapter 7

Complex Analysis I

Although this chapter is called complex *analysis*, we will try to develop the subject as complex *calculus* — meaning that we will 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.

7.1 Cauchy-Riemann equations

We will 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

$$\begin{aligned} \sin z \equiv \sin(x+iy) &= \sin x \cos iy + \cos x \sin iy \\ &= \sin x \cosh y + i \cos x \sinh y. \end{aligned} \tag{7.1}$$

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 will 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

$$\frac{\partial u}{\partial x}, \quad \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x}, \quad \frac{\partial u}{\partial y}, \tag{7.2}$$

exist and are continuous then $f = u + iv$ is differentiable as a complex-valued function of two real variables. This means that we can linearly approximate the variation in f as

$$\delta f = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y + \cdots \tag{7.3}$$

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

$$\delta f = \frac{\partial f}{\partial z}\delta z + \frac{\partial f}{\partial \overline{z}}\delta \overline{z} + \cdots, \tag{7.4}$$

where

$$\begin{aligned} \frac{\partial f}{\partial z} &\equiv \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \\ \frac{\partial f}{\partial \overline{z}} &\equiv \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right). \end{aligned} \tag{7.5}$$

Now our function $f(z)$ is not supposed to depend on $\overline{z}$, so it should satisfy

$$\partial_{\overline{z}} f \equiv \frac{\partial f}{\partial \overline{z}} = 0. \tag{7.6}$$

Thus, with $f = u + iv$,

$$0 = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)(u + iv), \tag{7.7}$$

or

$$\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right) + i\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) = 0. \tag{7.8}$$

Since the vanishing of a complex number requires the real and imaginary parts to be separately zero, this implies that

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial v}{\partial y}, \\ \frac{\partial v}{\partial x} &= -\frac{\partial u}{\partial y}. \end{aligned} \tag{7.9}$$

These 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 the function is *complex differentiable* (or just differentiable) at that point, and, taking $\delta z = z - z_0$, we have

$$\delta f \equiv f(z) - f(z_0) = \frac{\partial f}{\partial z}(z - z_0) + \cdots, \tag{7.10}$$

where the remainder, represented by the dots, tends to zero faster than $|z - z_0|$ as $z \to z_0$. This linear approximation to the variation in $f(z)$ is equivalent to the statement that the ratio

$$\frac{f(z) - f(z_0)}{z - z_0} \tag{7.11}$$

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 "is not supposed to depend on $\overline{z}$". Since we no longer need $\bar{z}$, it is natural to drop the partial derivative signs and write the limit as an ordinary derivative

$$\frac{df}{dz}, \quad \text{or} \quad f'(z). \tag{7.12}$$

This complex derivative obeys exactly the same calculus rules as the ordinary real derivatives:

$$\begin{aligned} \frac{d}{dz} z^n &= n z^{n-1}, \\ \frac{d}{dz} \sin z &= \cos z, \\ \frac{d}{dz}(fg) &= \frac{df}{dz} g + f \frac{dg}{dz}, \quad \textit{etc.} \end{aligned} \tag{7.13}$$

If the function is differentiable at all points in an arcwise-connected open set, or *domain*, D, the function is said to be *analytic* there. The words *regular* or *holomorphic* are also used.

7.1.1 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

$$\begin{aligned}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)u &= 0,\\ \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)v &= 0.\end{aligned} \tag{7.14}$$

and so both the real and imaginary parts of $f(z)$ are automatically *harmonic* functions of x, y.

Further, from Cauchy-Riemann again, we deduce that

$$\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}=0. \tag{7.15}$$

This means that $\nabla u\cdot\nabla v=0$, and so any pair of curves $u=const.$ and $v=const.$ intersect at right angles. If we regard u as the potential ϕ solving some electrostatics problem, then the curves $v=const.$ are the associated field lines.

In fluid mechanics, if $\mathbf{v}$ is the velocity field of an irrotational ($\nabla\times\mathbf{v}=\mathbf{0}$) flow, then we can wrote the flow field as a gradient

$$\begin{aligned}v_x &= \partial_x\phi,\\ v_y &= \partial_y\phi,\end{aligned} \tag{7.16}$$

where ϕ is a *velocity potential*. If the flow is incompressible ($\nabla\cdot\mathbf{v}=0$), then we can write it as a curl

$$\begin{aligned}v_x &= \partial_y\chi,\\ v_y &= -\partial_x\chi,\end{aligned} \tag{7.17}$$

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

$$\begin{aligned}\partial_x\phi &= \partial_y\chi,\\ \partial_y\phi &= -\partial_x\chi.\end{aligned} \tag{7.18}$$

Thus ϕ and χ are harmonic conjugates, and so the combination $\Phi=\phi+i\chi$ is an analytic function called the *complex stream function*.

A conjugate v exists for any harmonic function u. Here is an existence proof: First, the motivation for the construction. Observe that if we assume we have a u, v pair obeying Cauchy-Riemann in some domain D then we can write

$$\begin{aligned} dv &= \frac{\partial v}{\partial x}dx + \frac{\partial v}{\partial y}dy \\ &= -\frac{\partial u}{\partial y}dx + \frac{\partial u}{\partial x}dy. \end{aligned} \tag{7.19}$$

This observation suggests that if we are given only a harmonic function u we can *define* a v by

$$v(z) - v(z_0) = \int_{z_0}^{z} \left(-\frac{\partial u}{\partial y}dx + \frac{\partial u}{\partial x}dy \right). \tag{7.20}$$

The integral is path independent, and hence well defined, because

$$\frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) - \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = -\nabla^2 u = 0. \tag{7.21}$$

We now observe that we can 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

$$v(z) = \int^{x} \left(-\frac{\partial u}{\partial y}\right) dx' + \text{ rest of integral}, \tag{7.22}$$

and may use

$$\frac{d}{dx}\int^{x} f(x', y)\, dx' = f(x, y) \tag{7.23}$$

to see that

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}, \tag{7.24}$$

at (x, y). If we approach along the y axis we may similarly compute

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}. \tag{7.25}$$

Thus our newly defined v does indeed obey the Cauchy-Riemann equations.

Because of the utility the harmonic conjugate it is worth giving a practical recipe for finding it. The method we give below is one we learned from John d'Angelo. It is more efficient that those given in the regular textbooks. We first observe that if f is a function of z only, then $\overline{f}$ depends only on $\overline{z}$, so we will write $\overline{f(z)} = \overline{f}(\overline{z})$. Now

$$u(x,y) = \frac{1}{2}\left(f(z) + \overline{f(z)}\right). \tag{7.26}$$

Set

$$x = \frac{1}{2}(z+\overline{z}), \quad y = \frac{1}{2i}(z-\overline{z}), \tag{7.27}$$

so

$$u\left(\frac{1}{2}(z+\overline{z}), \frac{1}{2i}(z-\overline{z})\right) = \frac{1}{2}\left(f(z) + \overline{f}(\overline{z})\right). \tag{7.28}$$

Now set $\overline{z} = 0$, while keeping z fixed! Thus

$$f(z) + \overline{f(0)} = 2u\left(\frac{z}{2}, \frac{z}{2i}\right). \tag{7.29}$$

The function f is not completely determined of course, because we can always add an imaginary constant to v, and the above is equivalent to

$$f(z) = 2u\left(\frac{z}{2}, \frac{z}{2i}\right) + iC, \quad C \in \mathbf{R}. \tag{7.30}$$

For example, let $u = x^2 - y^2$. We find

$$f(z) + \overline{f(0)} = 2\left(\frac{z}{2}\right)^2 - 2\left(\frac{z}{2i}\right)^2 = z^2, \tag{7.31}$$

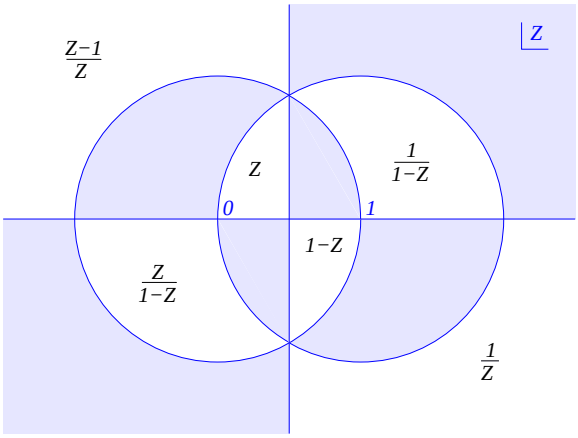
or

$$f(z) = z^2 + iC, \quad C \in \mathbf{R}. \tag{7.32}$$

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.

7.1.2 Conformal Mapping

An analytic function $w = f(z)$ will map 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 electrostatics or two dimensional fluid flow. Their simplest property is geometrical: such maps are *conformal*.



The unshaded triangle marked z is mapped conformally into the other five unshaded regions by the functions labeling them. Observe that the angles of the triangle is preserved the maps.

Suppose that the derivative of $f(z)$ at a point z_0 is non-zero. Then

$$f(z) - f(z_0) \approx A(z - z_0), \tag{7.33}$$

where

$$A = \left.\frac{df}{dz}\right|_{z_0}. \tag{7.34}$$

If you think about the geometric interpretation of complex multiplication (multiply the magnitudes, add the arguments) you will see that "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 $z \to f(z) = w$ 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.[1] The conformal

[1] If f were a function of $\overline{z}$ only, then the map would still be isogonal, but would reverse the orientation. We might call these maps *antiholomorphic* and *anti-conformal*.

property will fail at points where the derivative vanishes.

If we can find a conformal map $z \ (\equiv x + iy) \rightarrow w \ (\equiv u + iv)$ of some domain D to another D' then a function $f(z)$ that solves a potential problem (a Dirichlet boundary-value problem, for example) in D will lead to $f(z(w))$ solving an analogous problem in D'.

Example: The map $z \rightarrow w = z + e^z$ maps the strip $-\pi \leq y \leq \pi$, $-\infty < x < \infty$ into 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 lines $y = \pm\pi$ get folded back on themselves at $w = -1 \pm i\pi$, where the derivative of $w(z)$ vanishes.

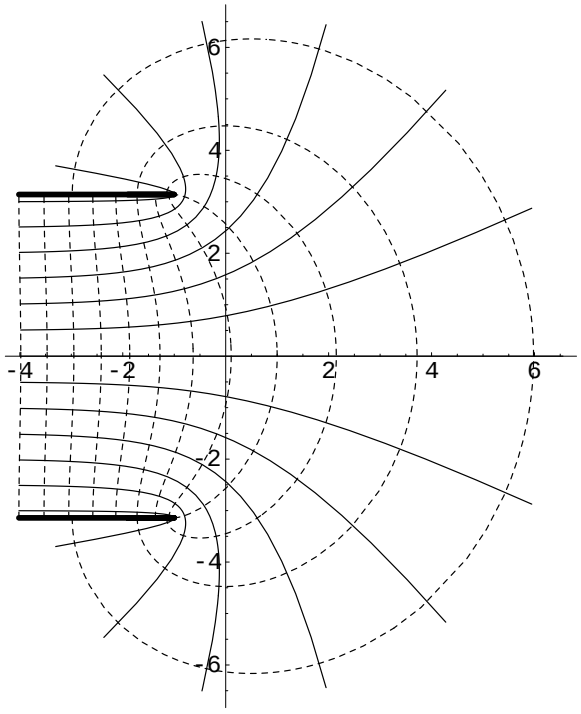


Image of part of the strip $-\pi \leq y \leq \pi$, $-\infty < x < \infty$ under the map $z \rightarrow w = z + e^z$.

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 later) to find the streamlines for the flow past a circle in the z plane, and then use Joukowski's transformation,

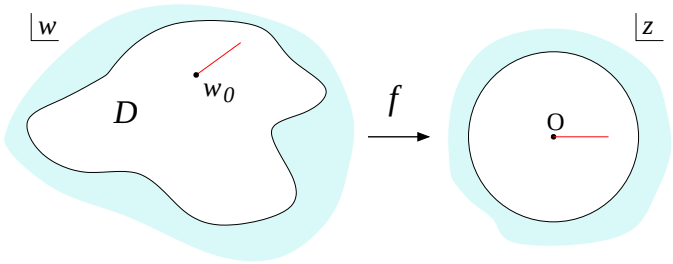
$$w = f(z) = \frac{1}{2}\left(z + \frac{1}{z}\right), \tag{7.35}$$

to map this simple flow to the flow past the aerofoil. The circle must go through the point $z = 1$, where the derivative of f vanishes, and this point becomes the sharp trailing edge of the aerofoil. To see this in action visit the web site: *http://www.math.psu.edu/glasner/Smp51/example1.html* where there is a java applet that lets you explore this map.

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 $\mathbf{C}$ *whose boundary consists of more that one point can be mapped conformally 1-1 and onto the interior of the unit circle. It is possible to chose 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.



The Riemann mapping theorem.

This theorem was "obvious" to Riemann, and for the reason we will give as a physical "proof". This argument is not rigorous, however, and it was many years before a real proof was found.

For the physical proof, observe that in the function

$$-\frac{1}{2\pi}\ln z = -\frac{1}{2\pi}\left\{\ln|z| + i\theta\right\}, \tag{7.36}$$

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 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

$$\nabla^2\phi_1 = -\delta^2(w - w_0), \quad \phi_1 = 0 \quad \text{on} \quad \partial D. \tag{7.37}$$

Now find the ϕ_2 that is harmonically conjugate to ϕ_1. Set

$$\phi_1 + i\phi_2 = \Phi(w) = -\frac{1}{2\pi}\ln(ze^{i\alpha}); \tag{7.38}$$

then we see that the transformation $w \to z$, or

$$z = e^{-i\alpha}e^{-2\pi\Phi(w)}, \tag{7.39}$$

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

$$z = e^{-i\alpha}\frac{w - i}{w + i}. \tag{7.40}$$

We immediately verify that that this works: we have $|z| = 1$ when w is real, and $z = 0$ at $w = i$.

The trouble 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 very jagged. However, unless ∂D is smooth enough to be *locally connected*, the potential ϕ_1 cannot be continuously extended to the boundary.

7.2 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 the earlier part of the course. 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 over 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 art to describe the outline of something), and the integrals themselves are then called *contour integrals*.

7.2.1 The Complex Integral

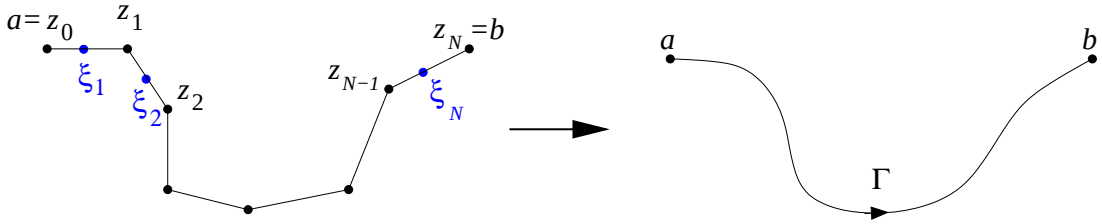
The complex integral

$$\int_\Gamma f(z)dz, \tag{7.41}$$

over a path Γ may be defined by expanding out the real and imaginary parts

$$\int_\Gamma f(z)dz \equiv \int_\Gamma (u+iv)(dx+idy) = \int_\Gamma (udx - vdy) + i\int_\Gamma (vdx + udy). \tag{7.42}$$

and treating the two integrals on the right hand side as standard vector-calculus line-integrals of the form $\int \mathbf{v}\cdot d\mathbf{r}$, with $\mathbf{v} \to (u, -v)$ and $\mathbf{v} \to (v, u)$.



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

$$\sum_{m=1}^{N} f(\xi_m)(z_m - z_{m-1}) \tag{7.43}$$

as N gets large and $\max|z_m - z_{m-1}| \to 0$. 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 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

$$\begin{aligned} \left|\sum_{m=1}^{N} f(\xi_m)(z_m - z_{m-1})\right| &\le \sum_{m=1}^{N} |f(\xi_m)(z_m - z_{m-1})| \\ &= \sum_{m=1}^{N} |f(\xi_m)|\,|(z_m - z_{m-1})|. \end{aligned} \tag{7.44}$$

For sufficiently smooth curves the last sum will converge to the real integral $\int_\Gamma |f(z)|\,|dz|$, and we deduce that

$$\left|\int_\Gamma f(z)\,dz\right| \le \int_\Gamma |f(z)|\,|dz|. \tag{7.45}$$

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

$$\left| \int_\Gamma f(z)\, dz \right| \leq M|\Gamma|. \tag{7.46}$$

We will 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,

$$f(z) = \frac{dg}{dz}, \tag{7.47}$$

then, for Γ a smooth path from $z = a$ to $z = b$, we have

$$\int_\Gamma f(z)dz = g(b) - g(a). \tag{7.48}$$

This follows by approximating $f(\xi_m) \approx (g(z_m) - g(z_{m-1}))/(z_m - z_{m-1})$, and observing that the sum resultant Riemann sum

$$\sum_{m=1}^{N} \Big(g(z_m) - g(z_{m-1}) \Big) \tag{7.49}$$

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 will 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 most of complex analysis, the result was known to Gauss.

7.2.2 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 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 one connected component is therefore a closed curve. We will mostly consider integrals about *simple* closed curves — these being curves that do

not self intersect — or 1-cycles consisting of 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 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 $\Gamma = \partial\Omega$ with f analytic, so $\partial_{\overline{z}} f = 0$, in Ω. We now exploit the fact that $\partial_{\overline{z}} f = 0$ in computing the exterior derivative,

$$df = \partial_z f\,dz + \partial_{\overline{z}} f\,d\overline{z} = \partial_z f\,dz, \tag{7.50}$$

of f, and use Stokes' theorem to deduce that

$$\oint_{\Gamma=\partial\Omega} f(z)dz = \int_\Omega d(f(z)dz) = \int_\Omega \partial_z f dz \wedge dz = 0. \tag{7.51}$$

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

$$\oint_\Gamma \left(\frac{1}{z}\right) dz = \int_0^{2\pi} e^{-i\theta}\, d\left(e^{i\theta}\right) = i\int_0^{2\pi} d\theta = 2\pi i. \tag{7.52}$$

Cauchy's theorem is not applicable because $1/z$ is *singular*, *i.e.* not differentiable, at $z = 0$. The formula (7.52) will hold for Γ any contour homologous to the unit circle in $\mathbf{C} \setminus 0$, the complex plane punctured by the removal of the point $z = 0$. Thus

$$\oint_\Gamma \left(\frac{1}{z}\right) dz = 2\pi i \tag{7.53}$$

for any contour Γ that encloses the origin. We can deduce a rather remarkable formula from this. Writing $\Gamma = \partial\Omega$ with anticlockwise orientation, we have

$$\oint_\Gamma \left(\frac{1}{z}\right) dz = \int_\Omega \partial_{\overline{z}} \left(\frac{1}{z}\right) d\overline{z}dz = 2\pi i \tag{7.54}$$

whenever Ω contains the origin. Since $d\overline{z}dz = 2idxdy$, we can restate this as

$$\partial_{\overline{z}} \left(\frac{1}{z}\right) = \pi\delta^2(x, y). \tag{7.55}$$

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

$$\oint_\Gamma \left(\frac{1}{z^2}\right) dz = \int_0^{2\pi} e^{-2i\theta}\, d\left(e^{i\theta}\right) = i\int_0^{2\pi} e^{-i\theta}\, d\theta = 0. \tag{7.56}$$

The same is true for all higher integer powers:

$$\oint_\Gamma \left(\frac{1}{z^n}\right) dz = 0, \quad n \geq 2. \tag{7.57}$$

We can understand this vanishing in another way by evaluating the integral as

$$\oint_\Gamma \left(\frac{1}{z^n}\right) dz = \oint_\Gamma \frac{d}{dz}\left(-\frac{1}{n-1}\frac{1}{z^{n-1}}\right) dz = \left[-\frac{1}{n-1}\frac{1}{z^{n-1}}\right]_\Gamma = 0, \quad n \neq 1. \tag{7.58}$$

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. Since the origin is within the contour, this angle increases by 2π, and so

$$[\ln z]_\Gamma = [i\arg z]_\Gamma = i\left(\arg e^{2\pi i} - \arg e^{0i}\right) = 2\pi i. \tag{7.59}$$

Exercise: Suppose $f(z)$ is analytic in a simply connected domain D, and $z_0 \in D$. Set $g(z) = \int_{z_0}^{z} f(z)$ 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: *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 (7.55) to argue formally that the general solution is

$$F(\zeta,\bar{\zeta}) = -\frac{1}{\pi}\int_\Omega \frac{f(z,\overline{z})}{z-\zeta} dx\, dy + g(\zeta),$$

where $g(\zeta)$ is an arbitrary analytic function. This result can be shown to be correct by more rigorous reasoning.

7.2.3 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

$$\oint_\Gamma f(z)\, dz = \sum_{\text{poles} \in D} 2\pi i \,(\text{residue at pole}), \tag{7.60}$$

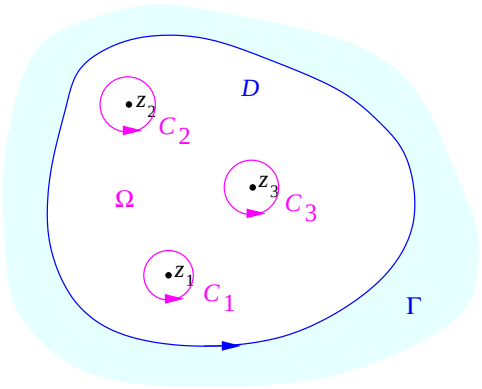
the integral being traversed in a positive (anticlockwise) sense. The words *pole* and *residue* referred to in the theorem mean the following: A pole is place where the function blows up. If, near z_0, the function can be written

$$f(z) = \left\{ \frac{a_N}{(z-z_0)^N} + \cdots + \frac{a_2}{(z-z_0)^2} + \frac{a_1}{(z-z_0)} \right\} g(z), \tag{7.61}$$

where $g(z)$ is analytic and non-zero at z_0, then $f(z)$ has a pole of order N at z_0. If $N = 1$ we have a *simple pole*. If we normalize $g(z)$ so that $g(z_0) = 1$ then the coefficient, a_1, of $1/(z-z_0)$ is the residue of the pole at z_0. The coefficients of the more singular terms do not influence the result of the integral, but N must be finite.

The evaluation of contour integrals therefore boils down to identifying where a complex function blows up, and looking at just how it does it.

We prove the residue theorem by drawing small circles C_i about each singular point z_i in D.



We then assert that

$$\oint_\Gamma f(z)\,dz = \sum_i \oint_{C_i} f(z)\,dz, \tag{7.62}$$

because the 1-cycle

$$\Gamma - \sum_i C_i = \partial\Omega \tag{7.63}$$

is the boundary of a region Ω in which f is analytic, and hence is homologous to zero. If we take the radius R_i of the circle C_i small enough we may replace $g(z)$ by its limit $g(z_i)$, and so set

$$\begin{aligned} f(z) &\to \left\{ \frac{a_1}{(z-z_i)} + \frac{a_2}{(z-z_i)^2} + \cdots \frac{a_N}{(z-z_i)^N} \right\} g(z_i), \\ &= \frac{a_1}{(z-z_i)} + \frac{a_2}{(z-z_i)^2} + \cdots \frac{a_N}{(z-z_i)^N}, \end{aligned} \tag{7.64}$$

on C_i. We the evaluate the integral over C_i by using our previous results. The theorem then follows.

We need to restrict ourselves to contours containing only finitely many poles 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: *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[2] space of D.

a) Suppose that $\varphi_n(z)$, $n = 1, 2, \ldots$, are a complete set of orthonormal functions on the Bergman space. Show that

$$K(\zeta, z) = \sum_{m=1}^{\infty} \varphi_n(\zeta)\overline{\varphi_n(z)}.$$

has the property that

$$g(\zeta) = \iint_D K(\zeta, z) g(z)\, dxdy.$$

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 the previous part 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 D and its boundary, use Stokes' theorem to show that

$$\iint_D \partial_{\overline{z}} g(z, \overline{z}) dxdy = \frac{1}{2i} \oint_C g(z, \overline{z}) dz.$$

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, $C = \partial D$, consists of more than one point. 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

[2]This space is not to be confused with the Bargmann-Fock space of analytic functions on the entirety of $\mathbf{C}$ with inner product

$$\langle f, g \rangle = \int_{\mathbf{C}} e^{-|z|^2} \bar{f} g d^2 z.$$

Bergman and Bargmann are two different people.

$\overline{f(\zeta)} = 0$ 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

$$g(\zeta) = \iint_D K(\zeta, z) g(z)\, dxdy.$$

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.

7.3 Applications

We now know enough about complex variables to work through some interesting applications, including understanding the mechanism by which an aeroplane flies.

7.3.1 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 $\mathbf{R}^2$ becomes

$$\begin{aligned} ds^2 &= dx \otimes dx + dy \otimes dy \\ &= d\overline{z} \otimes dz \\ &= g_{zz} dz \otimes dz + g_{\overline{z}z} d\overline{z} \otimes dz + g_{z\overline{z}} dz \otimes d\overline{z} + g_{\overline{z}\overline{z}} d\overline{z} \otimes d\overline{z}, \end{aligned} \quad (7.65)$$

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

$$\nabla^2 = g^{ij} \partial^2_{ij} = 2(\partial_z \partial_{\overline{z}} + \partial_{\overline{z}} \partial_z). \quad (7.66)$$

It is not safe to assume that $\partial_z \partial_{\overline{z}} f = \partial_{\overline{z}} \partial_z f$ when f has singularities. For example, from

$$\partial_{\overline{z}} \left(\frac{1}{z} \right) = \pi \delta^2(x, y), \tag{7.67}$$

we deduce that

$$\partial_{\overline{z}} \partial_z \ln z = \pi \delta^2(x, y). \tag{7.68}$$

When we evaluate the derivatives in the opposite order, however, we have

$$\partial_z \partial_{\overline{z}} \ln z = 0. \tag{7.69}$$

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

$$\begin{aligned} \nabla^2 \ln |z| &= 2\pi \delta^2(x, y), \\ (\partial_x \partial_y - \partial_y \partial_x)\theta &= 2\pi \delta^2(x, y). \end{aligned} \tag{7.70}$$

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 coordinate 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 coordinates rule, but a quicker route is to observe that

$$\mathbf{v} \cdot d\mathbf{r} = v_x dx + v_y dy = v_z dz + v_{\overline{z}} d\overline{z}. \tag{7.71}$$

Now

$$\partial_{\overline{z}} v_z = \frac{1}{4}(\partial_x v_x + \partial_y v_y) + i\frac{1}{4}(\partial_y v_x - \partial_x v_y). \tag{7.72}$$

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.

7.3.2 Milne-Thomson Circle Theorem

As we mentioned earlier, we can describe an irrotational and incompressible fluid motion either by a velocity potential

$$v_x = \partial_x \phi, \quad v_y = \partial_y \phi, \tag{7.73}$$

where $\mathbf{v}$ is automatically irrotational but incompressibilty requires $\nabla^2 \phi = 0$, or by a stream function

$$v_x = \partial_y \chi, \quad v_y = -\partial_x \chi, \tag{7.74}$$

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 Cauchy-Riemann and is therefore an analytic function of z. We see that

$$2v_z = \frac{d\Phi}{dz}, \tag{7.75}$$

ϕ and χ making equal contributions.

The Milne-Thomson theorem says that if Φ is the complex stream function for a flow in free space, then

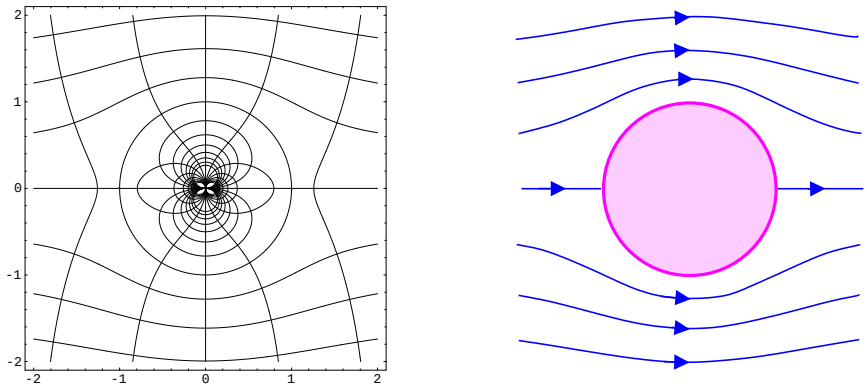
$$\tilde{\Phi} = \Phi(z) + \overline{\Phi}\left(\frac{a^2}{z}\right) \tag{7.76}$$

is the stream function after the cylinder $|z| = a$ 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 $\operatorname{Im} \tilde{\Phi} = \chi = 0$. 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 addional 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

$$\tilde{\Phi}(z) = U\left(z + \frac{a^2}{z}\right).$$

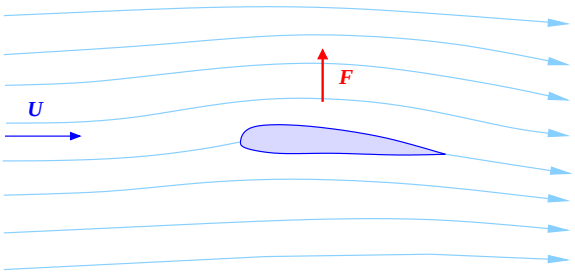
Since v_z is the derivative of this, we see that the perturbing effect of the obstacle on the velocity field falls off as the square of the distance from the cylinder.



The real and imaginary parts of the function $z + z^{-1}$ provide the streamlines and velocity potentials for irrotational incompressible flow past a unit radius cylinder.

7.3.3 Blasius and Kutta-Joukowski Theorems

We now derive the celebrated result, discovered independently by Kutta (1902) and 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 $\kappa = \oint \mathbf{v} \cdot d\mathbf{r}$ 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 flow can be regarded as being two dimensional.



Flow past an aerofoil.

Begin by recalling how the momentum flux tensor

$$T_{ij} = \rho v_i v_j + g_{ij} P \tag{7.77}$$

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

$$\rho(\partial_t v_i + v^j \partial_j v_i) = -\partial_i P + f_i. \tag{7.78}$$

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 rewrite this using mass conservation,

$$\partial_t \rho + \partial^i(\rho v_i) = 0, \tag{7.79}$$

as

$$\partial_t(\rho v_i) + \partial^j(\rho v_j v_i + \delta_{ij} P) = f_i. \tag{7.80}$$

This 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:

$$f_i = \partial^l T_{li} \equiv g^{kl} \partial_k T_{li}. \tag{7.81}$$

Since we are interested in steady, irrotational motion with constant 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

$$T_{ij} = \rho(v_i v_j - \frac{1}{2} g_{ij}|v|^2). \tag{7.82}$$

Using $v_z = \frac{1}{2}(v_x - i v_y)$ and

$$T_{zz} = \frac{\partial x^i}{\partial z}\frac{\partial x^j}{\partial z} T_{ij} \tag{7.83}$$

together with

$$x \equiv x^1 = \frac{1}{2}(z + \overline{z}), \qquad y \equiv x^2 = \frac{1}{2i}(z - \overline{z}) \tag{7.84}$$

we find

$$T \equiv T_{zz} = \frac{1}{4}(T_{xx} - T_{yy} - 2iT_{xy}) = \rho(v_z)^2. \tag{7.85}$$

This is the only component of T_{ij} we will need to consider. $T_{\overline{z}\overline{z}}$ is simply $\overline{T}$ while $T_{z\overline{z}} = 0 = T_{\overline{z}z}$ because T_{ij} is traceless.

In our complex coordinates, the equation

$$f_i = g^{kl}\partial_k T_{li} \tag{7.86}$$

reads

$$f_z = g^{\overline{z}z}\partial_{\overline{z}} T_{zz} + g^{z\overline{z}}\partial_z T_{\overline{z}z} = 2\partial_{\overline{z}} T. \tag{7.87}$$

We see that in steady flow the net momentum flux $\dot{P}_i$ out of a region Ω is given by

$$\dot{P}_z = \int_\Omega f_z \, dxdy = \frac{1}{2i}\int_\Omega f_z \, d\overline{z}dz = \frac{1}{i}\int_\Omega \partial_{\overline{z}} T \, d\overline{z}dz = \frac{1}{i}\oint_{\partial\Omega} T \, dz. \tag{7.88}$$

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

$$\dot{P}_z = -i\rho \oint_{\partial\Omega} v_z^2 \, dz, \tag{7.89}$$

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

$$F_z = i\rho \oint_{\partial\Omega} v_z^2 \, dz. \tag{7.90}$$

The result (7.90) is often called *Blasius' theorem*.

Evaluating this integral 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 the integral over a path encircling the aerofoil at large distance where the flow field takes the asymptotic form

$$v_z = U_z + \frac{\kappa}{4\pi i}\frac{1}{z} + O(\frac{1}{z^2}). \tag{7.91}$$

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

$$\oint \mathbf{v}\cdot d\mathbf{r} = \oint v_z dz + \oint v_{\overline{z}} \, d\overline{z} = \kappa. \tag{7.92}$$

Substituting v_z in (7.90) 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 with $\kappa/(4\pi i z)$. This gives

$$F_z = i\rho\left(\frac{U_z\kappa}{2\pi i}\right)\oint\frac{dz}{z} = i\rho\kappa U_z \tag{7.93}$$

or

$$\frac{1}{2}(F_x - iF_y) = i\rho\kappa\frac{1}{2}(U_x - iU_y). \tag{7.94}$$

Thus, in conventional coordinates, the reaction force on the body is

$$\begin{aligned} F_x &= \rho\kappa U_y, \\ F_y &= -\rho\kappa U_x. \end{aligned} \tag{7.95}$$

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 leave a pair of *wingtip vortices* of strength κ trailing behind them, and these 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.

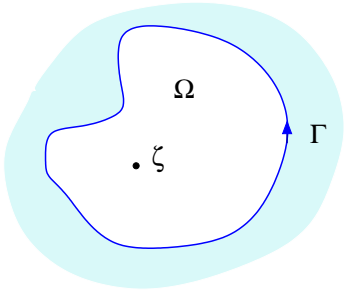
7.4 Applications of Cauchy's Theorem

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.

7.4.1 Cauchy's Integral Formula

If $f(z)$ is analytic within and on the boundary of a simply connected region 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*

$$f(\zeta) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z-\zeta}\, dz, \quad \zeta \in \Omega. \tag{7.96}$$



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

$$f^{(n)}(\zeta) = \frac{n!}{2\pi i} \oint_\Gamma \frac{f(z)}{(z-\zeta)^{n+1}}\, dz. \tag{7.97}$$

This shows that being *once* differentiable (analytic) in a region automatically implies that $f(z)$ is differentiable *arbitrarily many times*!

Exercise: *The generalized Cauchy formula.* Now suppose that we have solved a D-bar problem, 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 (7.55). 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.

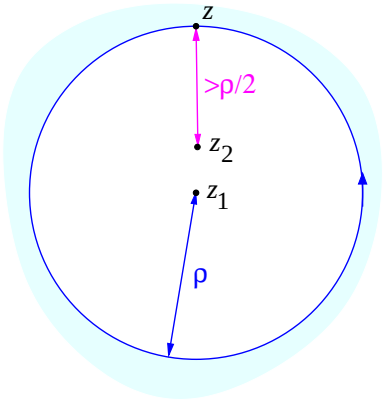
Liouville's Theorem

A dramatic corollary of Cauchy's integral formula is provided by *Liouville's theorem*: *If $f(z)$ is analytic in all of $\mathbf{C}$, 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 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 works. We select any two points, z_1 and z_2, and use Cauchy to write

$$f(z_1) - f(z_2) = \frac{1}{2\pi i} \oint_{\Gamma} \left(\frac{1}{z - z_1} - \frac{1}{z - z_2} \right) f(z) \, dz. \tag{7.98}$$

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 $|z - z_2| > \rho/2$.



Contour for Liouville' theorem.

Then, using $|\int f(z)dz| < \int |f(z)||dz|$, we have

$$\begin{aligned} |f(z_1) - f(z_2)| &= \frac{1}{2\pi}\left|\oint_\Gamma \frac{(z_1 - z_2)}{(z - z_1)(z - z_2)} f(z)\, dz\right| \\ &< \frac{1}{2\pi}\int_0^{2\pi} \frac{|z_1 - z_2|K}{\rho/2}\, d\theta = \frac{2|z_1 - z_2|K}{\rho}. \end{aligned} \tag{7.99}$$

The right hand side can be made arbitrarily small by taking ρ large enough, so we we must have $f(z_1) = f(z_2)$. Since z_1 and z_2 were any pair of points, we deduce that $f(z)$ takes the same value everywhere.

7.4.2 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 will 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

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n \equiv \lim_{N\to\infty} S_N, \tag{7.100}$$

where S_N are the *partial sums*

$$S_n = \sum_{n=0}^{N} a_n (z - z_0)^n. \tag{7.101}$$

Suppose that this limit exists (i.e the series is convergent) for some $z = \zeta$; then the series is *absolutely convergent*[3] for any $|z - z_0| < |\zeta - z_0|$.

To establish the absolute convergence we may assume, without loss of generality, that $z_0 = 0$. Then, convergence of the sum 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

$$|a_n z^n| = |a_n\zeta^n| \left|\frac{z}{\zeta}\right|^n < B \left|\frac{z}{\zeta}\right|^n . \tag{7.102}$$

The sum $\sum |a_n\zeta^n|$ 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 each power series series possesses a *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 will 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 $\sum a_n z^n$:

$$\begin{aligned} R &= \lim_{n\to\infty} \frac{|a_{n-1}|}{|a_n|} \\ &= \lim_{n\to\infty} |a_n|^{1/n}. \end{aligned} \tag{7.103}$$

The proof of these 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 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

[3]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*!

the same bound as in the proof of absolute convergence, we have

$$|a_n z^n| < B\frac{|z^n|}{|\zeta|^n} < B\frac{r^n}{|\zeta|^n} = M_n \tag{7.104}$$

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

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \tag{7.105}$$

and R is defined by $R = \zeta$ for any ζ for which the series converges, then $f(z)$ is analytic in $|z| < R$ and

$$f'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}. \tag{7.106}$$

Morera's Theorem

This is a partial converse of Cauchy's theorem: *If $f(z)$ is defined and continuous in a domain D and $\oint_\Gamma f(z)\,dz = 0$ for all contours that are homologous to zero, then $f(z)$ is analytic in D.* To prove this we set $F(z) = \int_P^z f(\zeta)\,d\zeta$, so (this is the point where we need continuity) $F'(z) = f(z)$. 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 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.

ii) $f_n'(z) \to f'(z)$ uniformly.

We use Morera, to prove i) (appealing to the uniform convergence to justify the interchange the order of summation and integration), and use Cauchy to prove ii).

Taylor's Theorem

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

$$f(\zeta) = f(a) + \sum_{n=0}^{\infty} \frac{(\zeta - a)^n}{n!} f^{(n)}(a), \tag{7.107}$$

meaning that this series converges to $f(\zeta)$ for all ζ such that $|\zeta - a| < \rho$.

We use the identity

$$\frac{1}{z-\zeta} = \frac{1}{z-a} + \frac{(\zeta - a)}{(z-a)^2} + \cdots + \frac{(\zeta - a)^{N-1}}{(z-a)^N} + \frac{(\zeta - a)^N}{(z-a)^N} \frac{1}{z-\zeta}. \tag{7.108}$$

and Cauchy's integral, to write

$$\begin{aligned} f(\zeta) &= \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{(z-\zeta)}\, dz \\ &= \sum_{n=0}^{N-1} \frac{(\zeta - a)^n}{2\pi i} \oint \frac{f(z)}{(z-a)^{n+1}}\, dz + \frac{(\zeta - a)^N}{2\pi i} \oint \frac{f(z)}{(z-a)^N (z-\zeta)}\, dz \\ &= \sum_{n=0}^{N-1} \frac{(\zeta - a)^n}{n!} f^{(n)}(a) + R_N \end{aligned} \tag{7.109}$$

where

$$R_N = \frac{(\zeta - a)^N}{2\pi i} \oint_\Gamma \frac{f(z)}{(z-a)^N (z-\zeta)}\, dz. \tag{7.110}$$

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 $\mathbf{C}$.) Then, estimating the integral using methods similar to those invoked in our proof of Liouville's Theorem, we find that

$$R_N < \frac{r^N}{2\pi} \left(\frac{2\pi\rho M}{\rho^N (\rho - r)} \right). \tag{7.111}$$

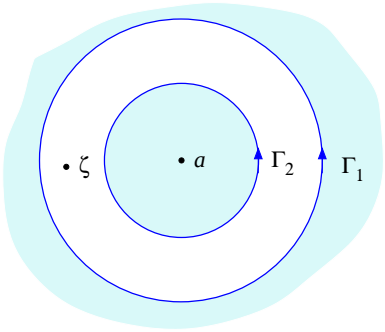
Since $r < \rho$, this tends to zero as $N \to \infty$.

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 is the distance to the nearest singularity.

Laurent Series

Theorem (Laurent): Let Γ_1 and Γ_2 be two anticlockwise circles 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:

$$f(\zeta) = \sum_{n=0}^{\infty} a_n(\zeta - a)^n + \sum_{n=1}^{\infty} b_n(\zeta - a)^{-n}. \tag{7.112}$$



Contours for Laurent's theorem.

The coefficients are given by

$$a_n = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{(z-a)^{n+1}}\,dz, \quad b_n = \frac{1}{2\pi i}\oint_{\Gamma_2} f(z)(z-a)^{n-1}\,dz. \tag{7.113}$$

This is proved by observing that

$$f(\zeta) = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{(z-\zeta)}\,dz - \frac{1}{2\pi i}\oint_{\Gamma_2} \frac{f(z)}{(z-\zeta)}\,dz. \tag{7.114}$$

and using the identities

$$\frac{1}{z-\zeta} = \frac{1}{z-a} + \frac{(\zeta-a)}{(z-a)^2} + \cdots + \frac{(\zeta-a)^{N-1}}{(z-a)^N} + \frac{(\zeta-a)^N}{(z-a)^N}\frac{1}{z-\zeta}. \tag{7.115}$$

and

$$-\frac{1}{z-\zeta} = \frac{1}{\zeta-a} + \frac{(z-a)}{(\zeta-a)^2} + \cdots + \frac{(z-a)^{N-1}}{(\zeta-a)^N} + \frac{(z-a)^N}{(\zeta-a)^N}\frac{1}{\zeta-z} \quad (7.116)$$

Once again we can show that the Remainder terms tend to zero.
Warning: Although the coefficients a_n are given by same integrals as in Taylor's theorem, they are not interpretable as derivatives of f unless $f(z)$ is analytic within the inner circle, when all the b_n are zero.

7.4.3 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 $z_0 \in D$,

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n. \quad (7.117)$$

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 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

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n + \sum_{n=1}^{\infty} b_n (z-a)^{-n}. \quad (7.118)$$

The second term, consisting of negative powers, is called *principal part* of $f(z)$ at $z = a$. It may happen that $b_m \neq 0$ while $b_n = 0$, $n > m$. This 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 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, and so $f(z) = z^m\varphi(z)$, where $\varphi(a) \neq 0$. In the second case f has 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 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.

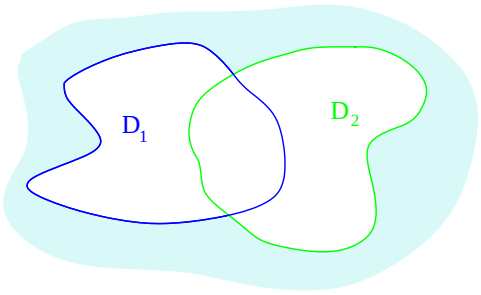
vi) Limit points of zeros II: Suppose that we know that f is analytic in D, except possibly at $z = a$ which is limit point of zeros as in iii), but we also know that f is not identically zero. Then $z = a$ must be singularity of f — but not a pole (or it f would tend to infinity and could not have arbitrarily 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 $a_n = 1/n\pi$.

vii) A limit point of poles or other singularities would be a *non-isolated essential singularity*.

7.4.4 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 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 iii),

vanishes everywhere in D_2.



We can use this result, coupled with the circular domains of convergence of the Taylor series, to extend the range of analytic functions beyond the domain of validity 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 positive α, is defined by its evaluation on a test function $\varphi(x)$ as

$$(x_+^{\alpha-1}, \varphi) = \int_0^\infty x^{\alpha-1}\varphi(x)\,dx. \tag{7.119}$$

The pairing $(x_+^{\alpha-1}, \varphi)$ is an an analytic funtion 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

$$\frac{d}{dx}\left(\frac{x^\alpha}{\alpha}\varphi(x)\right) = x^{\alpha-1}\varphi(x) + \frac{x^\alpha}{\alpha}\varphi'(x). \tag{7.120}$$

We find that

$$\left[\frac{x^\alpha}{\alpha}\varphi(x)\right]_\epsilon^\infty = \int_\epsilon^\infty x^{\alpha-1}\varphi(x)\,dx + \int_\epsilon^\infty \frac{x^\alpha}{\alpha}\varphi'(x)\,dx.$$

The integrated-out part tends to zero as we take ϵ to zero and both of the integrals converge in this limit as well. Consequently

$$I_1(\alpha) \equiv -\frac{1}{\alpha}\int_0^\infty x^\alpha\varphi'(x)\,dx$$

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 continued function possesses a pole at $\alpha = 0$, with residue

$$-\int_0^\infty \varphi'(x)\,dx = \varphi(0).$$

We can repeat the integration by parts, and find that

$$I_2(\alpha) \equiv \frac{1}{\alpha(\alpha+1)} \int_0^\infty x^{\alpha+1} \varphi''(x)\,dx$$

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, and 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

$$I(\alpha) \equiv \int_0^\infty x^{\alpha-1} \phi(x)\,dx$$

is convergent and analytic in the strip $0 < \mathrm{Re}\,(\alpha) < 1$. By the same reasoning as above, $I(\alpha)$ is there equal to

$$-\int_0^\infty \frac{x^\alpha}{\alpha} \phi'(x)\,dx.$$

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

$$\begin{aligned}
-\int_0^\infty \frac{x^\alpha}{\alpha} \phi'(x)\,dx &= \lim_{\epsilon \to 0} \left\{ \int_\epsilon^\infty x^{\alpha-1} \phi(x)\,dx - \left[\frac{x^\alpha}{\alpha} \phi(x) \right]_\epsilon^\infty \right\} \\
&= \lim_{\epsilon \to 0} \left\{ \int_\epsilon^\infty x^{\alpha-1} \phi(x)\,dx + \phi(0) \frac{\epsilon^\alpha}{\alpha} \right\} \\
&= \lim_{\epsilon \to 0} \left\{ \int_\epsilon^\infty x^{\alpha-1} [cd\phi(x) - \phi(0)]\,dx \right\}, \\
&= \int_0^\infty x^{\alpha-1} [\phi(x) - \phi(0)]\,dx.
\end{aligned}$$

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

$$I(\alpha) = \int_0^\infty x^{\alpha-1}[\phi(x) - \phi(0)]\,dx.$$

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

$$\begin{aligned}
I(\alpha) &= \int_0^\infty x^{\alpha-1}[\phi(x) - \phi(0)]\,dx, \qquad -1 < \mathrm{Re}\,(\alpha) < 0 \\
&= \int_0^\infty x^{\beta-1}\left[\frac{\phi(x) - \phi(0)}{x}\right] dx, \quad \text{where } \beta = \alpha + 1, \\
&\to \int_0^\infty x^{\beta-1}\left[\frac{\phi(x) - \phi(0)}{x} - \phi'(0)\right] dx, \quad -1 < \mathrm{Re}\,(\beta) < 0 \\
&= \int_0^\infty x^{\alpha-1}[\phi(x) - \phi(0) - x\phi'(0)]\,dx, \quad -2 < \mathrm{Re}\,(\alpha) < -1,
\end{aligned}$$

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

$$(x_+^{\alpha-1}, \varphi) = \begin{cases} \displaystyle\int_0^\infty x^{\alpha-1}\varphi(x)\,dx, & 0 < \mathrm{Re}\,(\alpha) < \infty \\ \displaystyle\int_0^\infty x^{\alpha-1}[\varphi(x) - \varphi(0)]\,dx, & -1 < \mathrm{Re}\,(\alpha) < 0 \\ \displaystyle\int_0^\infty x^{\alpha-1}[\varphi(x) - \varphi(0) - x\varphi'(0)]\,dx, & -2 < \mathrm{Re}\,(\alpha) < -1. \end{cases}$$

Sit perpetuum — 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.

7.4.5 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? The answer is yes, there is a unique extension provided that the function is well enough behaved near the point. Curiously, the proof of this gives us insight into the wild behaviour of functions near essential singulaities.

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

$$f(\zeta) = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{z-\zeta}\,dz - \frac{1}{2\pi i}\oint_{\Gamma_2} \frac{f(z)}{z-\zeta}\,dz. \tag{7.121}$$

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

$$\tilde{f}(\zeta) = \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{f(z)}{z-\zeta}\,dz \tag{7.122}$$

within Γ_1, we see that $\tilde{f} = f$ in $D \setminus a$, and is analytic in all of D.

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 $\mathbf{C}$.

To see this, define $N_\delta(a) = \{z \in \mathbf{C} : |z-a| < \delta\}$, and $N_\epsilon(\zeta) = \{z \in \mathbf{C} : |z-\zeta| < \epsilon\}$. 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

$$\left|\frac{1}{f(z)-\zeta}\right| < \frac{1}{\epsilon} \tag{7.123}$$

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

$$f(z) = \zeta + \frac{1}{g(z)} \tag{7.124}$$

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 $z = 0$.

7.5 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.

7.5.1 Principle of the Argument

If $f(z)$ is meromorphic in D with $\partial D = \Gamma$, and $f(z) \neq 0$ on Γ, then

$$\frac{1}{2\pi i} \oint_\Gamma \frac{f'(z)}{f(z)}\, dz = N - P \tag{7.125}$$

where N is the number of zero's in D and P is the number of poles. To see this we note that if $f(z) = (z-a)^m \varphi(z)$ where φ is analytic and non-zero near a, then

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{\varphi'(z)}{\varphi(z)} \tag{7.126}$$

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

$$\oint_\Gamma \frac{f'(z)}{f(z)}\, dz = \Delta_\Gamma \ln f(z) = i \Delta_\Gamma \arg f(z), \tag{7.127}$$

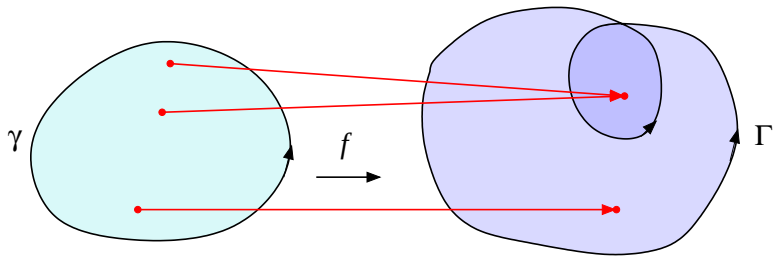
the symbol Δ_Γ denoting the total change in the quantity after we traverse Γ. Thus

$$N - P = \frac{1}{2\pi} \Delta_\Gamma \arg f(z). \tag{7.128}$$

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.



The map is one-to-one where the winding number is one, but two-to-one at points where the image curve winds twice.

To see this we appeal to the principal of the argument as

$$\begin{aligned} \#\text{ of zeros of } (f-a) \text{ within } \gamma &= \frac{1}{2\pi i}\oint_\gamma \frac{f'(z)}{f(z)-a}\,dz, \\ &= \frac{1}{2\pi i}\oint_\Gamma \frac{dw}{w-a}, \\ &= n(\Gamma, a), \end{aligned} \tag{7.129}$$

where $n(\Gamma, a)$ is called the winding number of the image curve Γ about a. It is equal to

$$n(\Gamma, a) = \frac{1}{2\pi}\Delta_\gamma \arg(w-a), \tag{7.130}$$

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.

7.5.2 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 Rouchés 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 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 $|z| = R$, provided R is sufficiently large. To prove this note that we can take R sufficiently big that

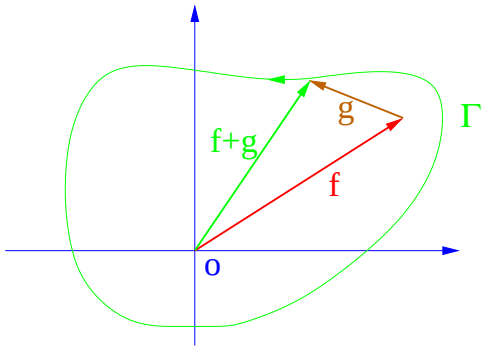
$$\begin{aligned} |a_n z^n| &= |a_n| R^n \\ &> |a_{n-1}| R^{n-1} + |a_{n-2}| R^{n-2} \cdots + |a_0| \\ &> |a_{n-a} z^{n-1} + a_{n-2} z^{n-2} \cdots + a_0|, \end{aligned} \tag{7.131}$$

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

$$\begin{aligned} \#\text{ of zeros of } f+g &= n(\Gamma, 0) \\ &= \frac{1}{2\pi} \Delta_\gamma \arg(f+g) \\ &= \frac{1}{2\pi i} \Delta_\gamma \ln(f+g) \\ &= \frac{1}{2\pi i} \Delta_\gamma \ln f + \frac{1}{2\pi i} \Delta_\gamma \ln(1+g/f) \\ &= \frac{1}{2\pi} \Delta_\gamma \arg f + \frac{1}{2\pi} \Delta_\gamma \arg(1+g/f). \end{aligned} \tag{7.132}$$

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.)



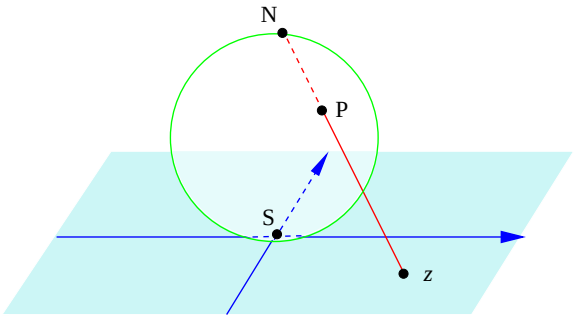
The curve Γ is the image of γ under the map $f + g$. If $|g| < |f|$, then, as z traverses γ, $f + g$ winds about the origin the same number of times that f does.

7.6 Analytic Functions and Topology

7.6.1 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 $\mathbf{C}$. 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.



Stereographic mapping of the complex plane to the 2-Sphere.

By the phrase a *neighbourhood* of z, we mean any 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 $\mathbf{C} \cup \infty$ 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 $z \to \zeta = 1/z$ to bring ∞ to the origin, and study the behaviour of the function there. Thus the polynomial

$$f(z) = a_0 + a_1 z + \cdots + a_N z^N \tag{7.133}$$

becomes

$$f(\zeta) = a_0 + a_1 \zeta^{-1} + \cdots + a_N \zeta^{-N}, \tag{7.134}$$

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

$$\frac{dz}{z} = \zeta\, d\left(\frac{1}{\zeta}\right) = -\frac{d\zeta}{\zeta}, \tag{7.135}$$

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

$$\begin{aligned} f(z) &= \cdots + a_{-2}z^{-2} + a_{-1}z^{-1} + a_0 + a_1 z + a_2 z^2 + \cdots + A_N z^N \\ &= \cdots + a_{-2}\zeta^{2} + a_{-1}\zeta + a_0 + a_1\zeta^{-1} + a_2\zeta^{-2} + \cdots + A_N\zeta^{-N} \end{aligned} \tag{7.136}$$

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

$$w = \left(\frac{z - z_0}{z - z_\infty}\right)\left(\frac{z_1 - z_\infty}{z_1 - z_0}\right) \tag{7.137}$$

maps

$$\begin{aligned} z_0 &\to 0, \\ z_1 &\to 1, \\ z_\infty &\to \infty, \end{aligned} \tag{7.138}$$

for example. Using this language, the Möbius maps

$$w = \frac{az + b}{cz + d} \tag{7.139}$$

become one-to-one maps of $S^2 \to S^2$. They are the only such one-to-one maps. When the matrix

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

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 argument 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

$$F(z) = f(z) - \sum_{n=1}^{N} \sum_{m=1}^{m_n} \frac{b_{n,m}}{(z-z_n)^m} \tag{7.140}$$

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

$$f(z) = \sum_{n=1}^{N} \sum_{m=1}^{m_n} \frac{b_{n,m}}{(z-z_n)^m} + C, \tag{7.141}$$

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 $z \to f(z)$ given by the rational function

$$f(z) = \frac{P(z)}{Q(z)} = \frac{a_n z^n + a_{n-1} z^{n-1} + \cdots a_0}{b_n z^n + b_{n-1} z^{n-1} + \cdots b_0} \tag{7.142}$$

wraps the Riemann sphere n times around the target S^2. In other words, it is a n-to-one map.

7.6.2 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 necessitates 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 are needed to make fractional powers single valued. We define the power z^α for for non-integral α by setting

$$z^\alpha = \exp\{\alpha \ln z\} = |z|^\alpha e^{i\alpha\theta}, \tag{7.143}$$

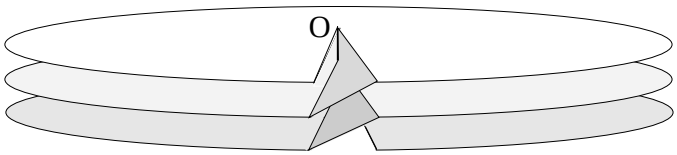
where $z = |z|e^{i\theta}$. For the square root $z^{1/2}$ we get

$$z^{1/2} = \sqrt{|z|}e^{i\theta/2}, \tag{7.144}$$

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!).

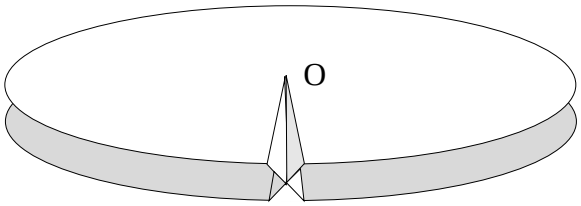
We can get away without cuts if we imagine the functions being maps *from* some set other than the complex plane. The new set is called a *Riemann surface*. It consists of a number of copies of the complex plane, one for each possible value of our "multivalued function". The map from this new surface 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:



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:

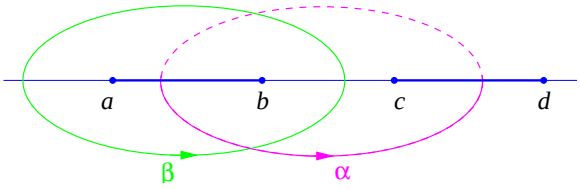


Part of the Riemann surface for $\sqrt{z}$*. Two copies of* $\mathbf{C}$ *are cross-connected. Circling the origin once takes you to the lower level. A second cicuit brings you back to the upper level.*

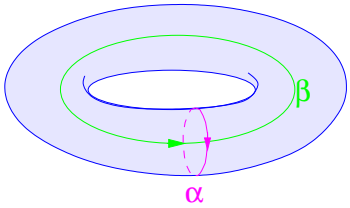
In both this and the previous drawing, we have shown the cross-connections being made rather abruptly along the cuts. This is not necessary —there is no singularity in the function at the cut — but it is often a convenient way to think about the structure of the surface. For example, the surface for $\sqrt{(z-a)(z-b)}$ 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.

Riemann surfaces often have interesting topology. As we have seen, the complex numbers, with the point at infinity included, have the topology of a sphere. The $\sqrt{(z-a)(z-b)}$ 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 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, the following figures 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 how the curves in the two-sheeted plane manage to intersect in only one point, just as they do when drawn on the torus.



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.



The 1-cycles α and β on the torus.

That the topology of the twice-cut plane is that of a torus has important consequences. This is because the *elliptic integral*

$$w = I^{-1}(z) = \int_{z_0}^{z} \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}} \tag{7.145}$$

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 two numbers $\omega_{1,2}$ are given by

$$\begin{aligned}\omega_1 &= \oint_\alpha \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}},\\ \omega_2 &= \oint_\beta \frac{dt}{\sqrt{(t-a)(t-b)(t-c)(t-d)}},\end{aligned} \tag{7.146}$$

and are called the *periods* of the *elliptic function* $z = I(w)$. The object $I(w)$ is a genuine function because the original z is uniquely determined by w. It is *doubly periodic* because

$$I(w + n\omega_1 + m\omega_2) = I(w), \quad n, m \in \mathbf{Z}. \tag{7.147}$$

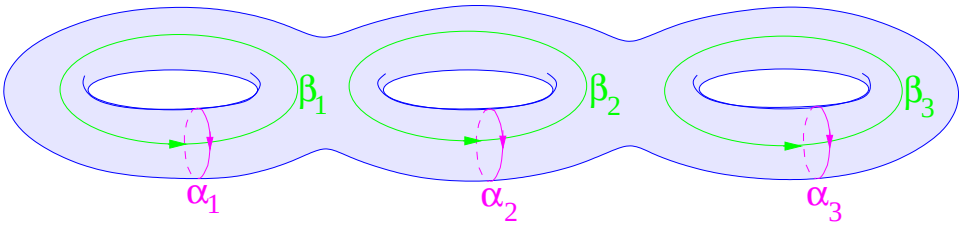
The inverse "function" $w = I^{-1}(z)$ is not a genuine function of z, however, because w increases by ω_1 or ω_2 each time z goes around a curve deformable into α or β, respectively. The periods are complicated functions of a, b, c, d.

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.

$$\text{“}dw\text{”} = \frac{dz}{\sqrt{(z-a)(z-b)(z-c)(z-d)}} \in H^1(T^2) \tag{7.148}$$

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.

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 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 $\chi = 2(1 - g)$.



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 $H_2(M)$.

For example, the figure 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

$$\chi = \sum_{p=0}^{2} (-1)^p \dim H_p(M) = 1 - 6 + 1 = -4, \tag{7.149}$$

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

$$z^3 + 3w^2 z + w + 3 = 0, \tag{7.150}$$

for example, we can think of z being a three-sheeted function of w defined by solving this cubic. Alternatively we can consider w to be the two-sheeted function of z obtained by solving the quadratic equation

$$w^2 + \frac{1}{3z} w + \frac{(3+z^3)}{3z} = 0. \tag{7.151}$$

In each case the branch points will be located where two or more roots coincide. The roots of (7.151), for example, coincide when

$$1 - 12z(3+z^3) = 0. \tag{7.152}$$

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. A generic (*i.e.* non-singular) curve

$$\sum_{r,s} a_{rs} z^r w^s = 0 \tag{7.153}$$

has genus

$$g = \frac{1}{2}(d-1)(d-2), \tag{7.154}$$

where $d = \max(r+s)$ 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 that *any* finite genus Riemann surface is the complex algebraic curve associated with some two-variable polynomial.

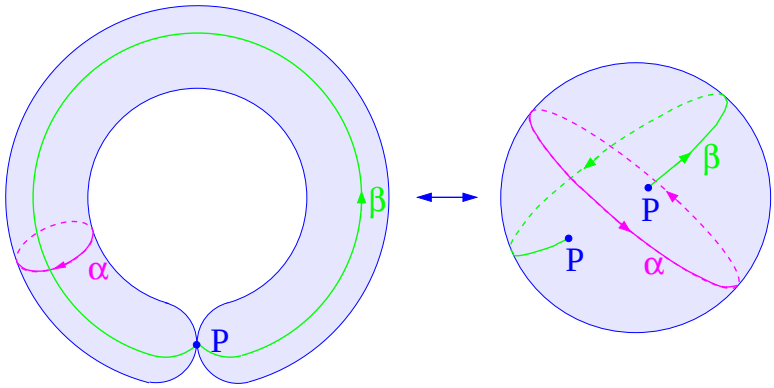
The "non-singular" condition above is important. A curve $P(z,w) = 0$ is said to be *singular* at $\mathrm{P} = (z_0, w_0)$ if all three of

$$P(z,w), \quad \frac{\partial P}{\partial z}, \quad \frac{\partial P}{\partial w}$$

vanish at P. If the curve has a singular point then then it degenerates and ceases to be a manifold. For example, we have seen that the curve

$$w^2 = (z-a)(z-b)(z-c)(z-d) \tag{7.155}$$

describes a torus when a, b, c, d are all distinct. If we allow b to coincide with c then the point $\mathrm{P} = (w_0, z_0) = (0, b)$ becomes a singular. If we look back at the figure of the twice-cut plane, we see that as b approaches c we can 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. This set is equivalent to a two-sphere with two points identified.



A degenerate torus is topologically the same as a sphere with two points identified.

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 coordinates that make it look like a region in $\mathbf{R}^2$. If we further let a coincide with $b = c$, then the two identified points on the sphere collide, and what is left is an surface that is homeomorphic to a sphere but with a singularity at P that prevents it from being diffeomorphic to the Riemann sphere.

7.6.3 Conformal Coordinates

Let's look back to some of our earlier work on differential geometry, and see how it looks from a complex variable point of view. Suppose we have a two-dimensional curved Riemann manifold with metric

$$`ds^2 = g_{ij}\, dx^i \otimes dx^j.$$

In two dimensions it is always possible to select what are called *conformal coordinates* x, y in which the metric tensor is diagonal, $g_{ij} = e^\sigma \delta_{ij}$, and so

$$ds^2 = e^\sigma (dx \otimes dx + dy \otimes dy).$$

The e^σ is called the *scale factor* or *conformal factor*. We won't try to prove this, but simply explore some of the consequences.

Firstly, $g^{ij}/\sqrt{g} = \delta_{ij}$, the conformal factor having cancelled. If you look back at its definition, you will see that this means that when the Hodge "$\star$"

operator acts on one forms, the result is independent of the metric. If ω is a one-form

$$\omega = p\,dx + q\,dy,$$

then

$$\star\omega = -q\,dx + p\,dy.$$

Note that, on one-forms,

$$\star\star = -1.$$

In complex coordinates $z = x + iy$, $\overline{z} = x - iy$ we have

$$\omega = \frac{1}{2}(p - iq)\,dz + \frac{1}{2}(p + iq)\,d\overline{z}.$$

Let us focus on the dz part,

$$A = \frac{1}{2}(p - iq)\,dz = \frac{1}{2}(p - iq)(dx + idy).$$

Then

$$\star A = \frac{1}{2}(p - iq)(dy - idx) = -iA.$$

Similarly, if

$$B = \frac{1}{2}(p + iq)\,d\overline{z},$$

then

$$\star B = iB.$$

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 decomposition of the identity into the sum of two projection operators

$$\begin{aligned} \mathbf{I} &= \frac{1}{2}(1 + i\star) + \frac{1}{2}(1 - i\star), \\ &= \quad P \quad + \quad \overline{P}, \end{aligned}$$

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 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 it means that $(p - iq)dz$ and $(p + iq)d\overline{z}$ are separately closed, and therefore $p - iq$ a holomorphic function.

The Jacobean torus

Suppose that M is a Riemann surface of genus g with α_i, β_i ,$i = 1, \ldots, g$, representative generators of $H_1(m)$. Suppose further that a and b are closed 1-forms, then, by cutting open the surface along the curves α_i, β_i we can show that

$$\int_M a \wedge b = \sum_{i=1}^{g} \left\{ \int_{\alpha_i} a \int_{\beta_i} b - \int_{\beta_i} a \int_{\alpha_i} b \right\}. \tag{7.156}$$

Applying de Rham's theorem to our genus-g surface we know that there must be $2g$ independent closed 1-forms forming a basis of $H^1(M)$. By applying the operator P we can assemble thse into g holomorphic closed 1-forms ω_i. Suppose that ω is is such a closed holomorphic 1-form, then its Hodge inner-product norm is

$$\begin{aligned} \|\omega\|^2 = \int_M \omega \star \overline{\omega} &= \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega \int_{\beta_i} \star\overline{\omega} - \int_{\beta_i} \omega \int_{\alpha_i} \star\overline{\omega} \right\} \\ &= i \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega \int_{\beta_i} \overline{\omega} - \int_{\beta_i} \omega \int_{\alpha_i} \overline{\omega} \right\} \\ &= \sum_{i=1}^{g} \left\{ A_i \overline{B}_i - B_i \overline{A}_i \right\}, \end{aligned} \tag{7.157}$$

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. We see, therefore, that if all the A_i are zero then $\|\omega\| = 0$ and so $\omega = 0$.

Let $A_{ij} = \int_{\alpha_i} \omega_j$. We will show that 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

$$0 = A_{ij}\lambda_j = \int_{\alpha_i} (\omega_j \lambda_j), \tag{7.158}$$

but, by (7.157) this implies that $\omega_j \lambda_j = 0$, contrary to the linear independence of the ω_i. We can therefore solve the equations

$$A_{ij}\lambda^{jk} = \delta_{ik} \tag{7.159}$$

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

$$\begin{aligned} 0 = \int_M \omega_m \wedge \omega_n &= \sum_{i=1}^{g} \left\{ \int_{\alpha_i} \omega_m \int_{\beta_i} \omega_n - \int_{\beta_i} \omega_m \int_{\alpha_i} \omega_n \right\} \\ &= \sum_{i=1}^{g} \{\delta_{im}\tau_{in} - \tau_{im}\delta_{in}\} \\ &= \tau_{mn} - \tau_{nm}. \end{aligned} \tag{7.160}$$

The matrix τ_{ij} is therefore symmetric. A similar compuation shows that

$$\|\lambda_i \omega_i\|^2 = 2\overline{\lambda}_i (\operatorname{Im} \tau_{ij}) \lambda_j \tag{7.161}$$

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*. It parameterises the shape of the Riemann surface.

Chapter 8

Complex Analysis II

In this chapter we will apply what we have learned of complex variables.

8.1 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

$$\oint_\Gamma f(z)\,dz = \sum_{\text{poles} \in D} 2\pi i\,(\text{residue at pole}).$$

8.1.1 Tricks of the Trade

The effective application of the residue theorem is something of an *art*, but there are useful classes of integrals which you should recognize.

Rational Trigonometric Expressions

Integrals of the form

$$\int_0^{2\pi} F(\cos\theta, \sin\theta)\,d\theta \tag{8.1}$$

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

$$I = \int_0^{2\pi} \frac{d\theta}{a + b\cos\theta} = \frac{2}{i}\oint_{|z|=1} \frac{dz}{bz^2 + 2az + b} = \frac{2}{ib}\oint \frac{dz}{(z-\alpha)(z-\beta)}. \quad (8.2)$$

Since $\alpha\beta = 1$, only one pole is within the contour. This is at

$$\alpha = (-a + \sqrt{a^2 - b^2})/b. \quad (8.3)$$

The residue is

$$\frac{2}{ib}\frac{1}{\alpha - \beta} = \frac{1}{i}\frac{1}{\sqrt{a^2 - b^2}}. \quad (8.4)$$

Therefore, the integral is given by

$$I = \frac{2\pi}{\sqrt{a^2 - b^2}}. \quad (8.5)$$

These integrals are, of course, also do-able by the "t" substitution $t = \tan(\theta/2)$, whence

$$\sin\theta = \frac{2t}{1+t^2}, \quad \cos\theta = \frac{1-t^2}{1+t^2}, \quad d\theta = \frac{2dt}{1+t^2}, \quad (8.6)$$

followed by a partial fraction decomposition. The labour is perhaps slightly less using the contour method.

Rational Functions

Integrals of the form

$$\int_{-\infty}^{\infty} R(x)\,dx, \quad (8.7)$$

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

$$I = 2\pi i \left(\sum \text{Residues of poles in upper half-plane}\right). \quad (8.8)$$

We could also use a rectangle in the lower half-plane with the result

$$I = -2\pi i \left(\sum \text{Residues of poles in lower half-plane} \right), \tag{8.9}$$

This must give the same answer.

For example, let n be a positive integer and consider

$$I = \int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^n}. \tag{8.10}$$

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

$$\begin{aligned} \frac{1}{(1+z^2)^n} &= \frac{1}{[(i+\zeta)^2+1]^n} = \frac{1}{(2i\zeta)^n}\left(1 - \frac{i\zeta}{2}\right)^{-n} \\ &= \frac{1}{(2i\zeta)^n}\left(1 + n\left(\frac{i\zeta}{2}\right) + \frac{n(n+1)}{2!}\left(\frac{i\zeta}{2}\right)^2 + \cdots\right). \end{aligned} \tag{8.11}$$

The coefficient of ζ^{-1} is

$$\frac{1}{(2i)^n}\frac{n(n+1)\cdots(2n-2)}{(n-1)!}\left(\frac{i}{2}\right)^{n-1} = \frac{1}{2^{2n-1}i}\frac{(2n-2)!}{((n-1)!)^2}. \tag{8.12}$$

The integral is therefore

$$I = \frac{\pi}{2^{2n-2}}\frac{(2n-2)!}{((n-1)!)^2}. \tag{8.13}$$

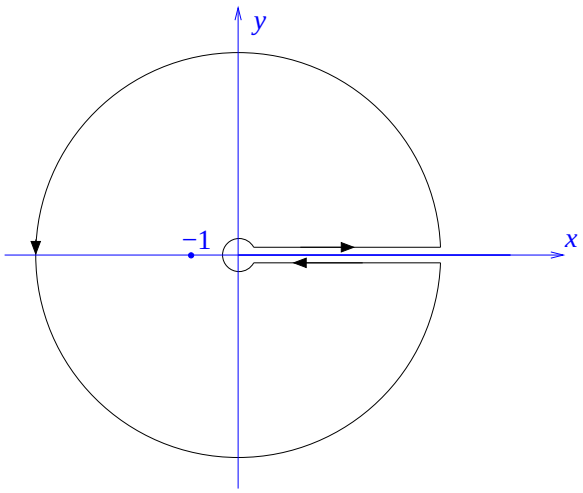
These integrals can also be done by partial fractions.

8.1.2 Branch-cut integrals

Integrals of the form

$$I = \int_0^{\infty} x^{\alpha-1} R(x) dx, \tag{8.14}$$

where $R(x)$ is rational, can be evaluated by integration round a slotted circle (or "key-hole") contour.



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

$$I = \int_0^\infty \frac{x^{\alpha-1}}{1+x} dx, \quad 0 < \operatorname{Re}\alpha < 1. \tag{8.15}$$

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

$$z^{\alpha-1} = \exp[(\alpha-1)(\ln|z| + i\theta)], \tag{8.16}$$

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

$$\oint_\Gamma \frac{z^{\alpha-1}}{1+z} dz = 2\pi i e^{\pi i(\alpha-1)}. \tag{8.17}$$

The integral decomposes as

$$\oint_\Gamma \frac{z^{\alpha-1}}{1+z} dz = \oint_{|z|=\Lambda} \frac{z^{\alpha-1}}{1+z} dz + (1 - e^{2\pi i(\alpha-1)}) \int_\epsilon^\Lambda \frac{x^{\alpha-1}}{1+x} dx - \oint_{|z|=\epsilon} \frac{z^{\alpha-1}}{1+z} dz. \tag{8.18}$$

As we send Λ off to infinity we can ignore the "1" in the denominator compared to the z, and so estimate

$$\left|\oint_{|z|=\Lambda} \frac{z^{\alpha-1}}{1+z} dz\right| \to \left|\oint_{|z|=\Lambda} z^{\alpha-2} dz\right| \leq 2\pi\Lambda \times \Lambda^{\mathrm{Re}\,(\alpha)-2}. \tag{8.19}$$

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

$$-e^{\pi i\alpha} 2\pi i = \left(1 - e^{2\pi i(\alpha-1)}\right) I. \tag{8.20}$$

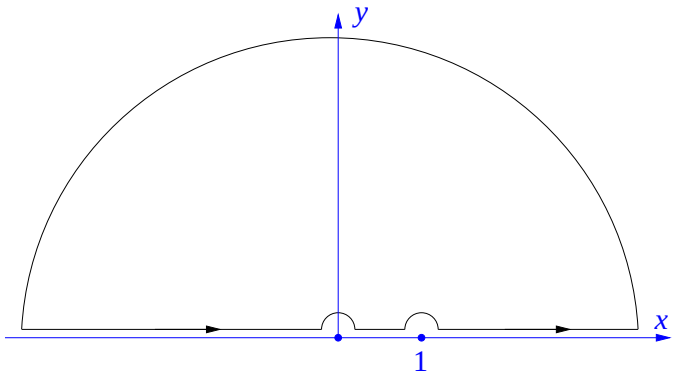
We conclude that

$$I = \frac{2\pi i}{(e^{\pi i\alpha} - e^{-\pi i\alpha})} = \frac{\pi}{\sin \pi\alpha}. \tag{8.21}$$

Exercise: 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: 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 $z = 1$



The contour Γ_1.

to get

$$0 = \oint_\Gamma \frac{z^{a-1}}{z-1}\, dz = P\int_0^\infty \frac{x^{a-1}}{x-1}\, dx - i\pi + (\cos \pi a + i \sin \pi a) \int_0^\infty \frac{x^{a-1}}{x+1}\, dx.$$

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

$$P\int_0^\infty \frac{x^{a-1}}{1-x}dx = \pi \cot \pi a, \quad 0 < \operatorname{Re} a < 1.$$

8.1.3 Jordan's Lemma

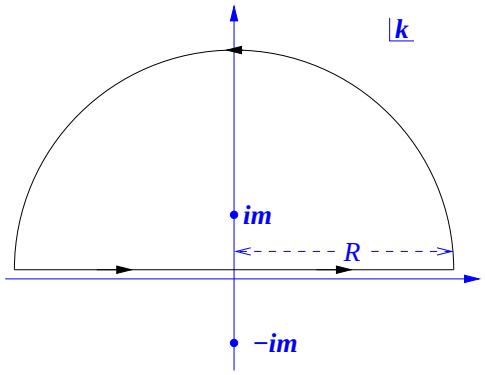
We often need to evaluate Fourier integrals

$$I(k) = \int_{-\infty}^{\infty} e^{ikx} R(x)\, dx \tag{8.22}$$

with $R(x)$ a rational function. For example, the Green function for the operator $-\partial_x^2 + m^2$ is given by

$$G(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \frac{e^{ikx}}{k^2 + m^2}. \tag{8.23}$$

Suppose $x \in \mathbf{R}$ 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.



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

$$G(x) = \frac{1}{2m}e^{-mx}, \quad x > 0. \tag{8.24}$$

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

$$G(x) = \frac{1}{2m}e^{+mx}, \quad x < 0. \tag{8.25}$$

We can combine the two results as

$$G(x) = \frac{1}{2m}e^{-m|x|}. \tag{8.26}$$

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

$$\int_\Gamma e^{i\lambda z} f(z)\, dz \to 0, \quad \text{as} \quad R \to \infty. \tag{8.27}$$

To establish this, we assume that R is large enough that $|f| < \epsilon$ on the contour, and make a simple estimate

$$\begin{aligned}\left|\int_\Gamma e^{i\lambda z} f(z)\, dz\right| &< 2R\epsilon \int_0^{\pi/2} e^{-\lambda R \sin\theta}\, d\theta \\ &< 2R\epsilon \int_0^{\pi/2} e^{-2\lambda R\theta/\pi}\, d\theta \\ &= \frac{\pi\epsilon}{\lambda}(1 - e^{-\lambda R}) < \frac{\pi\epsilon}{\lambda}. \end{aligned} \tag{8.28}$$

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

$$I(\alpha) = \int_{-\infty}^{\infty} \frac{\sin(\alpha x)}{x}\, dx.$$

We have

$$I(\alpha) = \mathrm{Im}\left\{\int_{-\infty}^{\infty} \frac{\exp i\alpha z}{z}\, dz\right\}.$$

If we take $\alpha > 0$, we can close in the upper half-plane, but our contour must exclude the pole at $z = 0$. Therefore

$$0 = \int_{|z|=R} \frac{\exp i\alpha z}{z}\, dz - \int_{|z|=\epsilon} \frac{\exp i\alpha z}{z}\, dz + \int_{-R}^{-\epsilon} \frac{\exp i\alpha x}{x}\, dx + \int_{\epsilon}^{R} \frac{\exp i\alpha x}{x}\, dx.$$

As $R \to \infty$, we can ignore the big semicircle, the rest, after letting $\epsilon \to 0$, gives

$$0 = -i\pi + P\int_{-\infty}^{\infty} \frac{e^{i\alpha x}}{x}\, dx.$$

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

$$\int_{-\infty}^{\infty} \frac{\sin\alpha x}{x}\, dx = \pi, \quad P\int_{-\infty}^{\infty} \frac{\cos\alpha x}{x}\, dx = 0, \qquad \alpha > 0.$$

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

$$\int_{-\infty}^{\infty} \frac{\sin \alpha x}{x}\, dx = \pi \operatorname{sgn} \alpha.$$

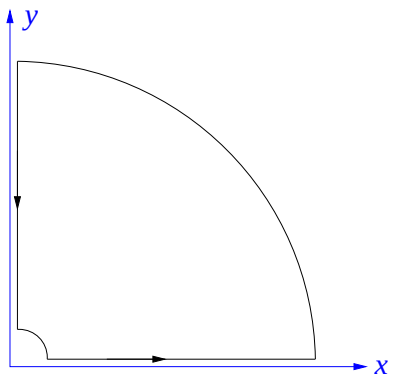
This identity is called *Dirichlet's discontinuous integral*.

We can interpret this calculation as giving the Fourier transform of the distribution $P(1/x)$ as

$$P \int_{-\infty}^{\infty} \frac{e^{i\omega x}}{x}\, dx = i\pi \operatorname{sgn} \omega.$$

This will be of use later in the chapter.

Example:



Quadrant contour.

Evaluate the integral

$$\oint_C e^{iz} z^{a-1}\, dz$$

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, deduce that

$$0 = \int_0^{\infty} e^{ix} x^{a-1}\, dx - i \int_0^{\infty} e^{-y} y^{a-1} e^{(a-1)\frac{\pi}{2} i}\, dy, \quad 0 < a < 1.$$

Take real and imaginary parts to find

$$\begin{aligned} \int_0^{\infty} x^{a-1} \cos x\, dx &= \Gamma(a) \cos\left(\frac{\pi}{2} a\right), \quad 0 < a < 1, \\ \int_0^{\infty} x^{a-1} \sin x\, dx &= \Gamma(a) \sin\left(\frac{\pi}{2} a\right), \quad 0 < a < 1, \end{aligned}$$

where

$$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y}\, dy$$

is the Euler Gamma function.

Example: Fresnel integrals. Integrals of the form

$$C(t) = \int_0^t \cos(\pi x^2/2)\, dx, \tag{8.29}$$

$$S(t) = \int_0^t \sin(\pi x^2/2)\, dx, \tag{8.30}$$

occur in the theory of diffraction and are called *Fresnel integrals* after Augustin Fresnel. They are naturally combined as

$$C(t) + iS(t) = \int_0^t e^{i\pi x^2/2}\, dx. \tag{8.31}$$

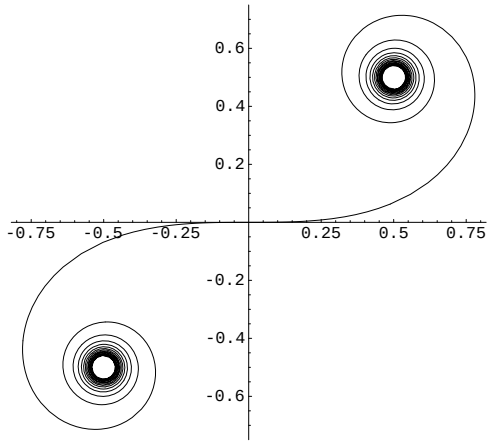
The limit as $t \to \infty$ exists and is finite. Even though the integrand does not tend to zero at infinity, its rapid oscillation for large x is just sufficient to ensure convergence[1].

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 $t \to \infty$.

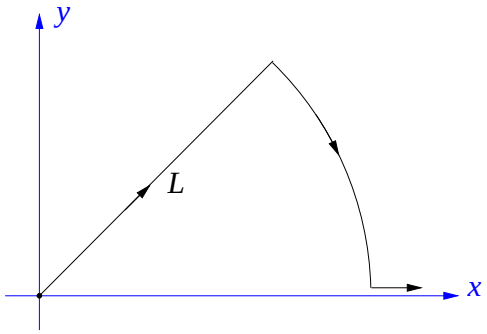


The Cornu spiral $C(t) + iS(t)$ for t in the range $-8 < t < 8$. The spiral in the first quadrant corresponds to positive values of t.

We can evaluate the limiting value

$$C(\infty) + iS(\infty) = \int_0^\infty e^{i\pi x^2/2}\, dx \tag{8.32}$$

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.



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

$$\int_0^{e^{i\pi/4}\infty} e^{i\pi z^2/2}\,dz = e^{i\pi/4}\int_0^{\infty} e^{-\pi s^2/2}\,ds = \frac{1}{\sqrt{2}}e^{i\pi/4} = \frac{1}{2}(1+i). \tag{8.33}$$

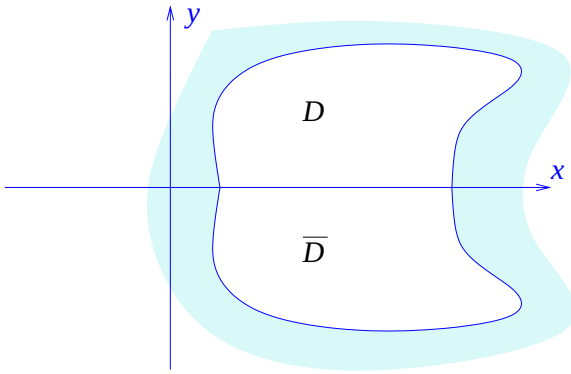
The figure shows how $C(t) + iS(t)$ orbits the limiting point $0.5 + 0.5i$ and slowly spirals in towards it. Taking real and imaginary parts we have

$$\int_0^{\infty} \cos\left(\frac{\pi x^2}{2}\right)\,dx = \int_0^{\infty} \sin\left(\frac{\pi x^2}{2}\right)\,dx = \frac{1}{2}. \tag{8.34}$$

8.2 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*

$$g(z) = \begin{cases} f(z), & z \in D, \\ \overline{f(\overline{z})}, & z \in \overline{D}, \\ \text{either}, & z \in \mathbf{R}. \end{cases}$$



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 $\overline{D}$:

$$\oint_C \overline{f(\overline{z})}\,dz,$$

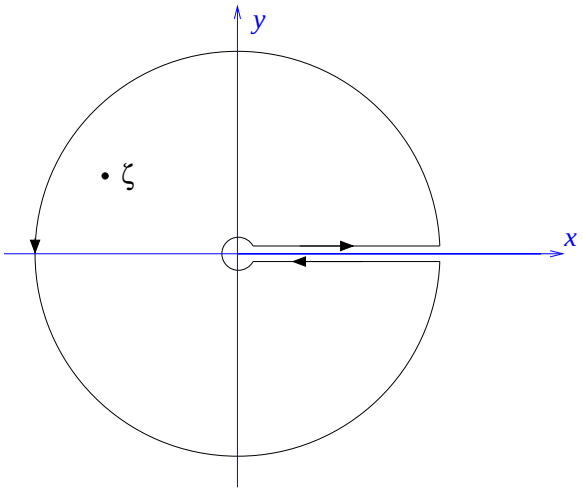
where $C = \{\overline{\eta(t)}\}$ is the image of $\overline{C} = \{\eta(t)\} \subset D$ under reflection in the real axis. We can rewrite this as

$$\oint_C \overline{f(\overline{z})}\, dz = \oint \overline{f(\eta)}\, \frac{d\bar{\eta}}{dt} dt = \overline{\oint f(\eta) \frac{d\eta}{dt} dt} = \overline{\oint_{\overline{C}} f(\eta)\, dz = 0}.$$

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 the figure



The contour Γ *for the dispersion relation.* .

we find

$$f(\zeta) = \frac{1}{\pi} \int_0^\infty \frac{\mathrm{Im}\, f(x + i\epsilon)}{x - \zeta}\, dx, \tag{8.35}$$

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

$$f(\zeta) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z-\zeta}\, dz, \tag{8.36}$$

and

$$f(b) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z-b}\, dz, \tag{8.37}$$

for some convenient point b within the contour. We then subtract to get

$$f(\zeta) = f(b) + \frac{(\zeta-b)}{2\pi i} \int_\Gamma \frac{f(z)}{(z-b)(z-\zeta)}\, dz. \tag{8.38}$$

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

$$f(\zeta) = f(b) + \frac{(\zeta-b)}{\pi} \int_0^\infty \frac{\mathrm{Im}\, f(x+i\epsilon)}{(x-b)(x-\zeta)}\, dx. \tag{8.39}$$

This is called a *once-subtracted* dispersion relation.

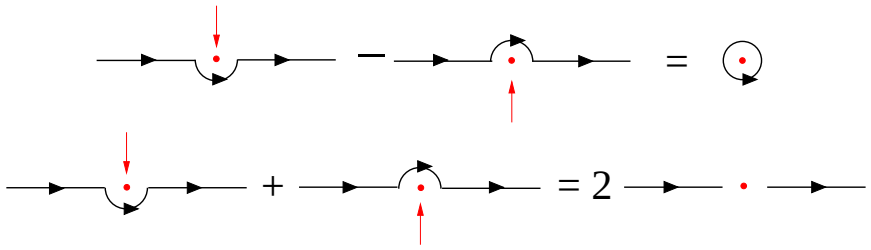
The dispersion relations derived above apply when ζ lies within the contour. In physics applications we often need $f(\zeta)$ for ζ real and positive. What happens as ζ approaches the axis, and we attempt to divide by zero in such an integral, is summarized by the *Plemelj formulæ*: If $f(\zeta)$ is defined by

$$f(\zeta) = \frac{1}{\pi} \int_\Gamma \frac{\rho(z)}{z-\zeta}\, dz,$$

where Γ has a segment lying on the real axis, then, if x lies in this segment,

$$\begin{aligned} \frac{1}{2}(f(x+i\epsilon) - f(x-i\epsilon)) &= i\rho(x) \\ \frac{1}{2}(f(x+i\epsilon) + f(x-i\epsilon)) &= \frac{P}{\pi} \int_\Gamma \frac{\rho(x')}{x'-x}\, dx'. \end{aligned}$$

As usual, the "P" means that we delete an infinitesimal segment of the contour lying symmetrically about the pole.



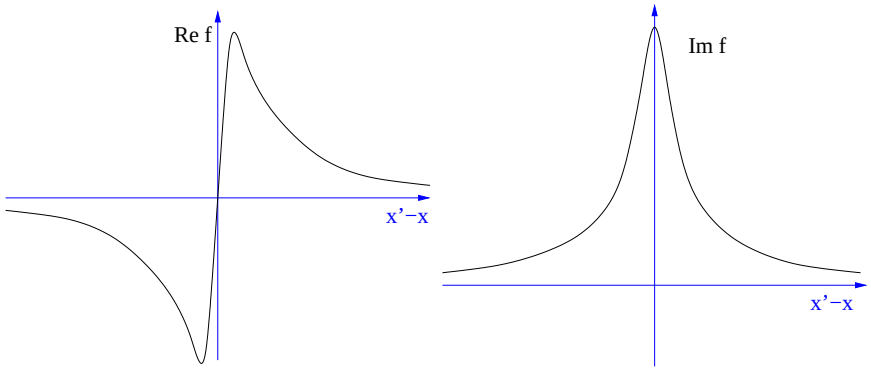
Origin of the Plemelj formulae.

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 are commonly encoded in physics papers via the "$i\epsilon$" cabala

$$\frac{1}{x' - x \pm i\epsilon} = P\left(\frac{1}{x' - x}\right) \mp i\pi\delta(x' - x).$$

A limit $\epsilon \to 0$ is always to be understood in this formula.



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:

$$\frac{1}{x' - (x \pm i\epsilon)} = \frac{x - x'}{(x' - x)^2 + \epsilon^2} \pm \frac{i\epsilon}{(x' - x)^2 + \epsilon^2}.$$

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 $\epsilon \to 0$.

8.2.1 Kramers-Kronig Relations

Causality is the usual source of analyticity in physical applications. If $G(t)$ is a response function

$$\phi_{\text{response}}(t) = \int_{-\infty}^{\infty} G(t-t') f_{\text{cause}}(t')\, dt' \tag{8.40}$$

then for no effect to anticipate its cause we must have $G(t) = 0$ for $t < 0$. The Fourier transform

$$G(\omega) = \int_{-\infty}^{\infty} e^{i\omega t} G(t)\, dt, \tag{8.41}$$

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

$$\ddot{x} + 2\gamma\dot{x} + (\Omega^2 + \gamma^2)x = F(t), \tag{8.42}$$

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

$$G(t,t') = \begin{cases} \Omega^{-1} e^{-\gamma(t-t')} \sin\Omega(t-t') & t > t' \\ 0, & t < t' \end{cases} \tag{8.43}$$

and so

$$x(t) = \frac{1}{\Omega} \int_{-\infty}^{t} e^{-\gamma(t-t')} \sin\Omega(t-t')\, F(t')\, dt'.$$

Because the integral extends only from 0 to $+\infty$, the Fourier transform of $G(t,0)$,

$$\tilde{G}(\omega) \equiv \frac{1}{\Omega} \int_0^{\infty} e^{i\omega t} e^{-\gamma t} \sin\Omega t\, dt,$$

is nicely convergent when $\mathrm{Im}\,\omega > 0$, as evidenced by

$$\tilde{G}(\omega) = -\frac{1}{(\omega + i\gamma)^2 - \Omega^2}$$

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

$$\varphi(\mathbf{x}, t) = e^{in(\omega)\mathbf{k}\cdot\mathbf{x} - i\omega t}.$$

We can decompose n into its real and imaginary parts

$$\begin{aligned} n(\omega) &= n_R(\omega) + in_I(\omega) \\ &= n_R(\omega) + \frac{i}{2|k|}\gamma(\omega) \end{aligned}$$

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[3].

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

$$n(\omega)_{\text{physical}} = \lim_{\epsilon \to 0} n(\omega + i\epsilon)$$

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

$$\begin{aligned} \gamma(-\omega + i\epsilon) &= -\gamma(\omega + i\epsilon), \\ n_R(-\omega + i\epsilon) &= +n_R(\omega + i\epsilon) \end{aligned} \tag{8.44}$$

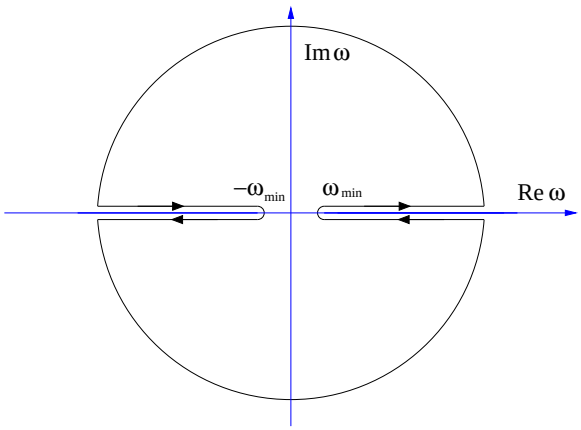
[2]If a pole in a response function manages to sneak into the upper half plane, then the system will be unstable to exponentially growing oscillations. This may happen, for example, when we design an electronic circuit containing a feedback loop. Such poles, and the resultant instabilities, can be detected by applying the principle of the argument from the last chapter. This method leads to the *Nyquist stability criterion.*

[3]For a dilute medium of incoherent scatterers, such as the air molecules responsible for Rayleigh scattering, $\gamma = N\sigma_{tot}$, where N is the density of scatterers and σ_{tot} is the total scattering cross section of a single scatterer.

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.



Contour for the $n - 1$ dispersion relation.

By applying the dispersion-relation strategy, we find

$$n(\omega) = 1 + \frac{1}{\pi}\int_{-\infty}^{\omega_{\min}} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' + \frac{1}{\pi}\int_{\omega_{\min}}^{\infty} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega'$$

for ω within the contour. Using Plemelj we can now take ω onto the real axis to get

$$\begin{aligned} n_R(\omega) &= 1 + \frac{P}{\pi}\int_{-\infty}^{\omega_{\min}} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' + \frac{P}{\pi}\int_{\omega_{\min}}^{\infty} \frac{n_I(\omega')}{\omega' - \omega}\, d\omega' \\ &= 1 + \frac{P}{\pi}\int_{\omega_{\min}^2}^{\infty} \frac{n_I(\omega')}{\omega'^2 - \omega^2}\, d\omega'^2, \end{aligned}$$

$$= \quad 1 + \frac{c}{\pi} P \int_{\omega_{\min}}^{\infty} \frac{\gamma(\omega')}{\omega'^2 - \omega^2}\, d\omega'.$$

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[4].

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

$$n_R = 1 + \frac{c}{\pi} P \int_{\omega_{min}^2}^{\infty} \left(\frac{\omega'^2}{\omega^2}\right) \frac{\gamma(\omega')}{\omega'^2 - \omega^2}\, d\omega',$$

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"

$$\epsilon(\omega) = \epsilon_1 + i\epsilon_2.$$

Again $\epsilon \to 1$ as $|\omega| \to \infty$, and, proceeding as before, we deduce that

$$\epsilon_1(\omega) = 1 + \frac{P}{\pi} \int_{\omega_{min}^2}^{\infty} \frac{\epsilon_2(\omega')}{\omega'^2 - \omega^2}\, d\omega'^2.$$

8.2.2 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

$$f(z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{f(x)}{x - z}\, dx$$

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

[4]H. A. Kramers, *Nature*, **117** (1926) 775; R. de L. Kronig, J. Opt. Soc. Am. **12** (1926) 547

reason the integral must give zero when z is taken in the lower half-plane. Using Plemelj we deduce that on the real axis,

$$f(x) = \frac{P}{\pi i}\int_{-\infty}^{\infty} \frac{f(x')}{x'-x}\,dx',$$

and we can derive Kramers-Kronig in this way even if n_I never vanishes so we cannot use Schwarz.

This result motivates the definition of the *Hilbert transform*, $\mathcal{H}\psi$, of a function $\psi(x)$, as

$$(\mathcal{H}\psi)(x) = \frac{P}{\pi}\int_{-\infty}^{\infty} \frac{\psi(x')}{x-x'}\,dx'.$$

Note the interchange of x, x' in the denominator compared to the previous formula. This is to make the Hilbert transform into a convolution integral. The motivating result shows that a function that is the boundary value of a function analytic and tending to zero 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 in the lower half-plane will be an eigenvector with eigenvalue $+i$. The Hilbert transform of a constant is zero[5].

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

$$\begin{aligned} f_I(x) &= (\mathcal{H}f_R)(x), \\ f_R(x) &= (\mathcal{H}^{-1}f_I)(x) = -(\mathcal{H}f_I)(x). \end{aligned}$$

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

$$Z(t) = X_R(t) + iX_I(t) = A(t)e^{i\phi(t)}$$

analytic in the upper half-plane. This complex function is the *analytic signal*[6]. The real quantity $A(t)$ is then known as the *instantaneous amplitude*, or *envelope*, while $\phi(t)$ is the *instantaneous phase* and

$$\omega_{\rm IF}(t) = \dot{\phi}(t)$$

[5]A function analytic in the *entire* complex plane and tending to zero at infinity must vanish identically by Liouville's theorem.

[6]D. Gabor, J. Inst. Elec. Eng. (Part 3), **93** (1946) 429-457.

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: Use the formula given earlier in this chapter for the Fourier transform of $P(1/x)$, combined with the convolution theorem for Fourier transforms, to show that analytic signal is derived from the original real signal by suppressing all negative frequency components (those proportional to $e^{-i\omega t}$ with $\omega > 0$) and multiplying the remaining positive-frequency amplitudes by two. Confirm, by investigating the convergence properties of the integral, that the resulting Fourier representation of the analytic signal does indeed give a function that is is analytic in the upper half plane.

8.3 Partial-Fraction and Product Expansions

In this section we will study other useful representations of functions which devolve from their analyticity properties.

8.3.1 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

$$\frac{1}{2\pi i}\oint_{\Gamma_n}\frac{f(z')}{z'-z}\,dz' = f(z) + \sum_{j=1}^{n}\frac{r_j}{z_j - z}.$$

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

$$f(z) - f(0) = \frac{z}{2\pi i}\oint_{\Gamma_n}\frac{f(z')}{z'(z'-z)}\,dz' + \sum_{j=1}^{n} r_j\left(\frac{1}{z-z_j} + \frac{1}{z_j}\right).$$

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

$$f(z) = f(0) + \sum_{j=1}^{\infty} r_j \left(\frac{1}{z - z_j} + \frac{1}{z_j} \right)$$

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

$$\operatorname{cosec} z - \frac{1}{z} = {\sum_{n=-\infty}^{\infty}}' (-1)^n \left(\frac{1}{z - n\pi} + \frac{1}{n\pi} \right).$$

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

$$\operatorname{cosec} z = \frac{1}{z} + 2z \sum_{1}^{\infty} (-1)^n \frac{1}{z^2 - n^2\pi^2}.$$

Example 2): A similar method gives

$$\cot z = \frac{1}{z} + {\sum_{n=-\infty}^{\infty}}' \left(\frac{1}{z - n\pi} + \frac{1}{n\pi} \right).$$

We can pair terms together to writen this as

$$\begin{aligned} \cot z &= \frac{1}{z} + \sum_{n=1}^{\infty} \left(\frac{1}{z - n\pi} + \frac{1}{z + n\pi} \right), \\ &= \frac{1}{z} + \sum_{n=1}^{\infty} \frac{2z}{z^2 - n^2\pi^2} \end{aligned}$$

or

$$\cot z = \lim_{N \to \infty} \sum_{n=-N}^{N} \frac{1}{z - n\pi}.$$

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: 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: By differentiating the partial fraction expansion for $\cot z$, show that, for k an integer ≥ 1, and $\mathrm{Im}\, z > 0$, 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: The *Bernoulli numbers* are defined by

$$\frac{x}{e^x - 1} = 1 + B_1 x + \sum_{n=1}^{\infty}B_{2k}\frac{x^{2k}}{(2k)!}.$$

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

$$x\cot x = ix + \frac{2ix}{e^{2ix} - 1} = 1 - \sum_{n=1}^{\infty}(-1)^{k+1}B_{2k}\frac{2^{2k}x^{2k}}{(2k)!}.$$

By expanding $1/(x^2 - n^2\pi^2)$ as a power series in x and comparing coefficients, deduce that, for positive integer k,

$$\sum_{n=1}^{\infty}\frac{1}{n^{2k}} = (-1)^{k+1}\pi^{2k}\frac{2^{2k-1}}{(2k)!}B_{2k}.$$

8.3.2 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

$$\frac{d}{dz}\ln g(z) = \left.\frac{g'(z)}{g(z)}\right|_{z=0} + \sum_{j=1}^{\infty}\left(\frac{1}{z - z_j} + \frac{1}{z_j}\right).$$

Integrating up we have

$$\ln g(z) = \ln g(0) + cz + \sum_{j=1}^{\infty} \left(\ln(1 - z/z_j) + \frac{z}{z_j} \right),$$

where $c = g'(0)/g(0)$. We now re-exponentiate to get

$$g(z) = g(0) e^{cz} \prod_{j=1}^{\infty} \left(1 - \frac{z}{z_j}\right) e^{z/z_j}.$$

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

$$\frac{\sin z}{z} = \prod_{n=1}^{\infty} \left(1 - \frac{z}{n\pi}\right) e^{z/n\pi} \left(1 + \frac{z}{n\pi}\right) e^{-z/n\pi}.$$

Thus

$$\sin z = z \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2\pi^2}\right).$$

Convergence of Infinite Products

Although not directly relevant to the material above, it is worth pointing out the following: Let

$$p_N = \prod_{n=1}^{N} (1 + a_n), \quad a_n > 0.$$

then

$$1 + \sum_{n=1}^{N} a_n < p_N < \exp\left\{ \sum_{n=1}^{N} a_n \right\}.$$

The infinite sum and product therefore converge or diverge together. If

$$P = \prod_{n=1}^{\infty} (1 + |a_n|),$$

converges, we say that

$$p = \prod_{n=1}^{\infty} (1 + a_n),$$

converges absolutely. As with sums, absolute convergence implies convergence, but not vice-versa.

Exercise: 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

$$\prod_{n=2}^{\infty}\left(1-\frac{1}{n^2}\right)=\frac{1}{2}.$$

8.4 Wiener-Hopf Equations

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 integral equations*. It is, however, easier to discuss Wiener-Hopf *sum equations*, which are their discrete analogue. In this case analyticity in the upper or lower half-plane is replaced by analyticity within or without the unit circle.

8.4.1 Wiener-Hopf Sum Equations

Consider the infinite system of equations

$$y_n=\sum_{m=-\infty}^{\infty}a_{n-m}x_m,\quad -\infty<n<\infty \tag{8.45}$$

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

$$\begin{aligned}A(\theta) &= \sum_{n=-\infty}^{\infty}a_n e^{in\theta},\\ Y(\theta) &= \sum_{n=-\infty}^{\infty}y_n e^{in\theta},\end{aligned} \tag{8.46}$$

then the systems of equations is, in principle at least, easy to solve. We simply introduce the function

$$X(\theta)=\sum_{n=-\infty}^{\infty}x_n e^{in\theta}, \tag{8.47}$$

and (8.45) becomes

$$Y(\theta) = A(\theta)X(\theta). \tag{8.48}$$

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 semi-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

$$y(s) = \int_{-\infty}^{\infty} A(s-t)y(t)\,dt, \quad -\infty < s < \infty. \tag{8.49}$$

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

$$\begin{aligned} A(z) &= \sum_{n=-\infty}^{\infty} a_n z^n, \\ X(z) &= \sum_{n=-\infty}^{\infty} x_n z^n, \\ Y(z) &= \sum_{n=-\infty}^{\infty} y_n z^n. \end{aligned} \tag{8.50}$$

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*

$$y(s) = \int_{0}^{\infty} A(s-t)y(t)\,dt, \quad 0 \le s < \infty \tag{8.51}$$

is the *Wiener-Hopf sum equation*

$$y_n = \sum_{m=0}^{\infty} a_{n-m}x_m, \quad 0 \le n < \infty. \tag{8.52}$$

This requires a more sophisticated approach. If you look back at our earlier discussion of why Wiener-Hopf integral equations are hard, you will see that there we claim 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 now explain how this works.

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 $N = n(\Gamma, A)$ where Γ is the unit circle. In other words, on how many times the function $A(z)$ circles 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

$$A(z) = \lambda f_+(z) z^N [f_-(z)]^{-1}, \tag{8.53}$$

where

$$\begin{aligned} f_+(z) &= 1 + \sum_{n=1}^{\infty} f_n^{(+)} z^n, \\ f_-(z) &= 1 + \sum_{n=1}^{\infty} f_{-n}^{(-)} z^{-n}. \end{aligned} \tag{8.54}$$

Here we demand that $f_+(z)$ be analytic and non-zero for $|z| < 1 + \epsilon$, and that $f_-(1/z)$ 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 $f_\pm(z)$ 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.

We now introduce the notation $[F(z)]_+$ and $[F(z)]_-$, meaning that we expand F 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 (8.52) in the form

$$[Y_+(z) + Y_-(z)] f_-(z) = \lambda z^N f_+(z) X_+. \tag{8.55}$$

If $N \geq 0$, and we break this equation into its positive and negative powers, we find

$$\begin{aligned} [Y_+ f_-]_+ &= \lambda z^N f_+(z) X_+, \\ [Y_+ f_-]_- &= -Y_- f_-(z). \end{aligned} \tag{8.56}$$

From the first of these equations we can read off the desired x_n as the positive power Laurent coefficients of

$$X_+(z) = [Y_+ f_-]_+ (\lambda z^N f_+(z))^{-1}. \tag{8.57}$$

As a byproduct, the second gives coefficient y_{-n} of $Y_-(z)$. Observe that there is a condition on Y_+ for this to work: the power series expansion of $\lambda z^N f_+(z) X_+$ starts with z^N, and so for a solution to exist the first N terms of $(Y_+ f_-)_+$ as a power series in z must be zero. In other words the given vector y_n must satisfy N consistency conditions. Another way of expressing this is to observe that the range of the operator A represented by the matrix a_{n-m} falls short of the being the entire space of possible y_n by N dimensions. This means that the null space of $A^\dagger$ is N dimensional:

$$\dim\left[\mathrm{Ker}\, A^\dagger\right] = N.$$

When $N < 0$, on the other hand, we have

$$\begin{aligned} [Y_+(z) f_-(z)]_+ &= [\lambda z^{-|N|} f_+(z) X_+(z)]_+ \\ [Y_+(z) f_-(z)]_- &= -Y_-(z) f_-(z) + [\lambda z^{-|N|} f_+(z) X_+(z)]_- . \end{aligned} \tag{8.58}$$

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 $Z_+(x) = z^n [f_+(z)]^{-1}$ 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

$$\boxed{\mathrm{Index}\,(A) \stackrel{\mathrm{def}}{=} \dim\left(\mathrm{Ker}\, A\right) - \dim\left(\mathrm{Ker}\, A^\dagger\right) = -N}$$

This connection between a topological quantity – in the present case the winding number — and the difference of the dimension of the null-spaces of an operator and its adjoint 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

$$A(z) = C \frac{\prod_n (z - a_n)}{\prod_m (z - b_m)}, \tag{8.59}$$

we simply take

$$f_+(z) = \frac{\prod_{|a_n|>0} (1 - z/a_n)}{\prod_{|b_m|>0} (1 - z/b_m)}, \tag{8.60}$$

where the products are over the linear factors corresponding to poles and zeros outside the unit circle, and

$$f_-(z) = \frac{\prod_{|b_m|<0} (1 - b_m/z)}{\prod_{|a_n|<0} (1 - a_n/z)}, \tag{8.61}$$

containing the linear factors corresponding to poles and zeros inside the unit circle. The constant λ and the power z^N in equation (8.53) are the factors that we have extracted from the right-hand sides of (8.60) and (8.61), respectively, in order to leave 1's as the first term in each linear factor.

More generally, we take the logarithm of

$$z^{-N}A(z) = \lambda f_+(z)(f_-(z))^{-1} \tag{8.62}$$

to get

$$\ln[z^{-N}A(z)] = \ln[\lambda f_+(z)] - \ln[f_-(z)], \tag{8.63}$$

where we desire $\ln[\lambda f_+(z)]$ to be the boundary value of a function analytic within the unit circle, and $\ln[f_-(z)]$ the boundary value of function analytic outside the unit circle and with $f_-(z)$ 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

$$F(z) = \frac{1}{2\pi i}\oint_{|z|=1} \frac{\ln[\zeta^{-N}A(\zeta)]}{\zeta - z}\, d\zeta \tag{8.64}$$

provides us with the desired factorization. This function $F(z)$ is everywhere analytic except for a branch cut along the unit circle, and its branches, F_+ within and F_- without the circle, differ by $\ln[z^{-N}A(z)]$. We therefore have

$$\begin{aligned} \lambda f_+(z) &= e^{F_+(z)}, \\ f_-(z) &= e^{F_-(z)}. \end{aligned} \tag{8.65}$$

The expression for F as an integral shows that $F(z) \sim const./z$ as $|z|$ goes to infinity and so guarantees that $f_-(z)$ 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

$$\mathbf{A} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & a_0 & a_1 & a_2 & \cdots \\ \cdots & a_{-1} & a_0 & a_1 & \cdots \\ \cdots & a_{-2} & a_{-1} & a_0 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{8.66}$$

into the product of an upper triangular matrix

$$\mathbf{U} = \lambda \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & 1 & f_1^{(+)} & f_2^{(+)} & \cdots \\ \cdots & 0 & 1 & f_1^{(+)} & \cdots \\ \cdots & 0 & 0 & 1 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{8.67}$$

a lower triangular matrix $\mathbf{L}$, where

$$\mathbf{L}^{-1} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 0 & 0 & \cdots \\ \cdots & f_{-1}^{(-)} & 1 & 0 & \cdots \\ \cdots & f_{-2}^{(-)} & f_{-1}^{(-)} & 1 & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{8.68}$$

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*

$$\mathbf{A} = \mathbf{L}\mathbf{\Lambda}^N\mathbf{U}, \tag{8.69}$$

is an infinite-dimensional example of the Gauss-Bruhat (or generalized LU) decomposition of a matrix. The finite dimensional Gauss-Bruhat decomposition factorizes a matrix $\mathbf{A} \in GL(n)$ as

$$\mathbf{A} = \mathbf{L}\mathbf{\Pi}\mathbf{U}, \tag{8.70}$$

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.

Chapter 9

Special Functions II

In this chapter we will apply complex analytic methods to some of the special functions of mathematical physics. The standard text in this field remains the venerable *Course of Modern Analysis* of E. T. Whittaker and G. N. Watson.

9.1 The Gamma Function

As an illustration of much what has gone before we will discuss the properties of Euler's "Gamma Function", $\Gamma(z)$. You probably have some acquaintance with this creature. The usual definition is

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \quad \operatorname{Re} z > 0, \quad \text{(definition A)}. \tag{9.1}$$

An integration by parts, based on

$$\frac{d}{dt}\left(t^z e^{-t}\right) = z t^{z-1} e^{-t} - t^z e^{-t}, \tag{9.2}$$

shows that

$$\left[t^z e^{-t}\right]_0^\infty = z \int_0^\infty t^{z-1} e^{-t}\, dt - \int_0^\infty t^z e^{-t}\, dt. \tag{9.3}$$

The integrated out part vanishes at both limits, provided the real part of z is greater than zero. Thus

$$\Gamma(z+1) = z\Gamma(z). \tag{9.4}$$

Since $\Gamma(1) = 1$, we deduce that

$$\Gamma(n) = (n-1)!, \quad n = 1, 2, 3, \cdots. \tag{9.5}$$

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

$$\Gamma(z) = \frac{\Gamma(z+n)}{z(z+1)\cdots(z+n-1)}. \tag{9.6}$$

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

$$\frac{d}{dt}(t^x e^{-t}) = xt^{x-1}e^{-t} - t^x e^{-t} \tag{9.7}$$

we have

$$\left[t^x e^{-t}\right]_\epsilon^\infty = x\int_\epsilon^\infty dt\, t^{x-1}e^{-t} - \int_\epsilon^\infty dt\, t^x e^{-t}. \tag{9.8}$$

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

$$-\frac{1}{x}\epsilon^x = \int_\epsilon^\infty dt\, t^{x-1}e^{-t} - \frac{1}{x}\int_\epsilon^\infty dt\, t^x e^{-t}. \tag{9.9}$$

Since, for this range of x,

$$-\frac{1}{x}\epsilon^x = \int_\epsilon^\infty dt\, t^{x-1}, \tag{9.10}$$

we can rewrite (9.9) as

$$\frac{1}{x}\int_\epsilon^\infty dt\, t^x e^{-t} = \int_\epsilon^\infty dt\, t^{x-1}\left(e^{-t}-1\right). \tag{9.11}$$

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

$$\frac{1}{x}\Gamma(x+1) = \int_0^\infty dt\, t^{x-1}\left(e^{-t}-1\right). \tag{9.12}$$

Since the left-hand side is equal to $\Gamma(x)$, we have shown that

$$\Gamma(x) = \int_0^\infty dt\, t^{x-1}\left(e^{-t}-1\right), \qquad -1 < \mathrm{Re}\, x < 0. \tag{9.13}$$

Similarly, if $-2 < \mathrm{Re}\, x < -1$, we can show that

$$\Gamma(x) = \int_0^\infty dt\, t^{x-1} \left(e^{-t} - 1 + t \right). \tag{9.14}$$

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,

$$B(a,b) \stackrel{\mathrm{def}}{=} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 (1-t)^{a-1} t^{b-1}\, dt \tag{9.15}$$

(which, as the *Veneziano formula*, was the original inspiration for string theory) and

$$\Gamma(z)\Gamma(1-z) = \pi \mathrm{cosec}\, \pi z. \tag{9.16}$$

Let's prove these. Set $t = y^2$, x^2 so

$$\begin{aligned} \Gamma(a)\Gamma(b) &= 4 \int_0^\infty y^{2a-1} e^{-y^2}\, dy \int_0^\infty x^{2b-1} e^{-x^2}\, dx \\ &= 4 \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} x^{2b-1} y^{2a-1}\, dx dy \\ &= 2 \int_0^\infty e^{-r^2} (r^2)^{a+b-1}\, d(r^2) \int_0^{\pi/2} \sin^{2a-1}\theta \cos^{2b-1}\theta\, d\theta. \end{aligned}$$

At this point we can put $\sin^2\theta = t$ to get the Beta function identity. If, on the other hand we put $a = 1-z$, $b = z$ we get

$$\Gamma(z)\Gamma(1-z) = 2 \int_0^\infty e^{-r^2}\, d(r^2) \int_0^{\pi/2} \cot^{2z-1}\theta\, d\theta = 2 \int_0^{\pi/2} \cot^{2z-1}\theta\, d\theta. \tag{9.17}$$

Now set $\cot\theta = \zeta$ when the last integral becomes one of our earlier examples:

$$2 \int_0^\infty \frac{\zeta^{2z-1}}{\zeta^2+1}\, d\zeta = \pi \mathrm{cosec}\, \pi z, \quad 0 < z < 1. \tag{9.18}$$

Although this integral has a restriction on the range of z, the result

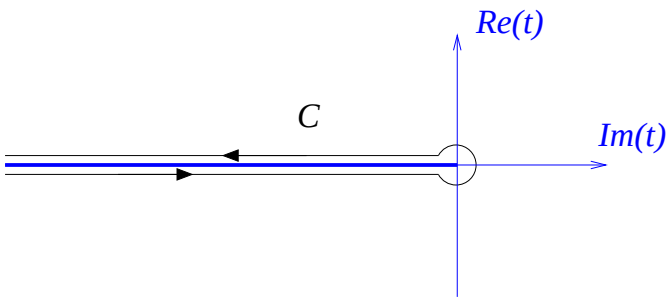
$$\Gamma(z)\Gamma(1-z) = \pi \mathrm{cosec}\, \pi z \tag{9.19}$$

can be analytically continued 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

$$\Gamma(1/2) = \sqrt{\pi}. \tag{9.20}$$

The integral in definition A is only convergent for $\mathrm{Re}\, z > 0$. A more powerful definition, involving an integral which converges for all z, is

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\int_C \frac{e^t}{t^z}\, dt. \quad \text{(definition B)} \tag{9.21}$$



Definition "B" contour.

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

$$\frac{1}{\Gamma(0)} = 0; \qquad \frac{1}{\Gamma(n)} = \frac{1}{(n-1)!}, \quad n > 0. \tag{9.22}$$

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

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\left[\frac{e^t}{(z-1)t^{z-1}}\right]_{-\infty-i\epsilon}^{-\infty+i\epsilon} + \frac{1}{(z-1)2\pi i}\int_C \frac{e^t}{t^{z-1}}\, dt = \frac{1}{(z-1)\Gamma(z-1)}. \tag{9.23}$$

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 $\Gamma(1-z)$

$$\frac{1}{\Gamma(1-z)} = \frac{1}{2\pi i}\int_C e^t t^{z-1}\,dt. \tag{9.24}$$

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

$$\begin{aligned}\frac{1}{\Gamma(1-z)} &= \frac{1}{2\pi i}\int_C e^t t^{z-1}\,dt = -\frac{1}{2\pi i}(2i\sin\pi(z-1))\int_0^\infty t^{z-1}e^{-t}\,dt \\ &= \frac{1}{\pi}\sin\pi z\int_0^\infty t^{z-1}e^{-t}\,dt. \end{aligned}\tag{9.25}$$

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

$$g(z) = g(0)e^{cz}\prod_1^\infty\left\{\left(1-\frac{z}{z_j}\right)e^{z/z_j}\right\} \tag{9.26}$$

for entire functions.

We need to recall the definition of Euler-Mascheroni constant $\gamma = -\Gamma'(1) = .5772157\ldots$, and that $\Gamma(1) = 1$. Then

$$\frac{1}{\Gamma(z)} = ze^{\gamma z}\prod_1^\infty\left\{\left(1+\frac{z}{n}\right)e^{-z/n}\right\}. \tag{9.27}$$

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: 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

$$\pi^2\mathrm{cosec}^2\pi z = \sum_{n=-\infty}^\infty \frac{1}{(z+n)^2}.)$$

9.2 Linear Differential Equations

9.2.1 Monodromy

Consider the linear differential equation

$$Ly \equiv y'' + p(z)y' + q(z)y = 0, \tag{9.28}$$

where p and q are meromorphic. Recall that the point $z = a$ is a *regular singular point* of the equation iff p or q is singular there, but

$$(z-a)p(z), \quad (z-a)^2q(z) \tag{9.29}$$

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 is 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

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \tag{9.30}$$

for some constants a, b, c, d. By a suitable redefinition of the y_i we may either diagonalise the *monodromy* matrix to find

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{9.31}$$

or, if the eigenvalues coincide and the matrix is not diagonalizable, reduce it to a Jordan form

$$\begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{9.32}$$

These equations are satisfied, in the diagonalizable case, by functions of the form

$$y_1 = (z-a)^{\alpha_1}\varphi_1(z), \quad y_2 = (z-a)^{\alpha_2}\varphi_2(z), \tag{9.33}$$

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

$$y_1 = (z-a)^{\alpha}\left[\varphi_1(z) + \frac{1}{2\pi i \lambda}\ln(z-a)\varphi_2(z)\right], \quad y_2 = (z-a)^{\alpha}\varphi_2(z), \tag{9.34}$$

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.

9.2.2 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

$$F(a,b;c;z) \;=\; 1 + \frac{a.b}{1.c}z + \frac{a(a+1)b(b+1)}{2!c(c+1)}z^2 +$$

$$+\frac{a(a+1)(a+2)b(b+1)(b+2)}{3!c(c+1)(c+2)}z^3+\cdots.$$

$$= \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_0^\infty \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)\Gamma(1+n)}z^n. \tag{9.35}$$

For general values of a, b, c, this converges for $|z| < 1$, the singularity restricting the convergence being a branch cut at $z = 1$.
Examples:

$$(1+z)^n = F(-n, b; b; -z), \tag{9.36}$$

$$\ln(1+z) = zF(1, 1; 2; -z), \tag{9.37}$$

$$z^{-1}\sin^{-1} z = F\left(\frac{1}{2}, \frac{1}{2}; \frac{3}{2}; z^2\right), \tag{9.38}$$

$$e^z = \lim_{b\to\infty} F(1, b; 1/b; z/b), \tag{9.39}$$

$$P_n(z) = F\left(-n, n+1; 1; \frac{1-z}{2}\right), \tag{9.40}$$

where in the last line P_n is the Legendre polynomial.

For future reference, we note that expanding the right hand side as a powers series in z and integrating term by term, shows that

$$F(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 (1-tz)^{-a}t^{b-1}(1-t)^{c-b-1}dt. \tag{9.41}$$

We may set $z = 1$ in this to get

$$F(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}. \tag{9.42}$$

The hypergeometric function is a solution of the second-order ODE

$$z(1-z)y'' + [c-(a+b+1)z]y' - aby = 0 \tag{9.43}$$

which has regular singular points at $z = 0, 1, \infty$. If $1 - c$ is not an integer, the general solution is

$$y = AF(a, b; c; z) + Bz^{1-c}F(b-c+1, a-c+1; 2-c; z). \tag{9.44}$$

The hypergeometric equation is a particular case of the general *Fuchsian equation* with three[1] regular singularities at $z = z_1, z_2, z_3$,

$$y'' + P(z)y' + Q(z)y = 0, \tag{9.45}$$

[1]The equation with *two* regular singularities is

$$y'' + p(z)y' + q(z)y = 0$$

where

$$\begin{aligned} P(z) &= \left(\frac{1-\alpha-\alpha'}{z-z_1}+\frac{1-\beta-\beta'}{z-z_2}+\frac{1-\gamma-\gamma'}{z-z_3}\right) \\ Q(z) &= \frac{1}{(z-z_1)(z-z_2)(z-z_3)}\times \\ &\quad \left(\frac{(z_1-z_2)(z_1-z_3)\alpha\alpha'}{z-z_1}+\frac{(z_2-z_3)(z_2-z_1)\beta\beta'}{z-z_2}+\frac{(z_3-z_1)(z_3-z_2)\gamma\gamma'}{z-z_3}\right), \end{aligned} \tag{9.46}$$

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 equations 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

$$r(r-1)+(1-\alpha-\alpha')r+\alpha\alpha'=0, \tag{9.47}$$

which has roots $r=\alpha,\alpha'$, so Riemann's equation has solutions which behave like $(z-z_1)^\alpha$ and $(z-z_1)^{\alpha'}$ near z_1, like $(z-z_2)^\beta$ and $(z-z_2)^{\beta'}$ near z_2, and similarly for z_3. A solution of Riemann's equations is traditionally denoted by the Riemann "P" symbol

$$y=P\left\{\begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array}\right\} \tag{9.48}$$

where the six quantities $\alpha,\beta,\gamma,\alpha',\beta',\gamma'$, are called the *exponents* of the so-

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

$$y=A\left(\frac{z-z_1}{z-z_2}\right)^\alpha+B\left(\frac{z-z_1}{z-z_2}\right)^{\alpha'}.$$

lution. A particular solution is

$$y = \left(\frac{z-z_1}{z-z_2}\right)^{\alpha} \left(\frac{z-z_3}{z-z_2}\right)^{\gamma} F\left(\alpha+\beta+\gamma, \alpha+\beta'+\gamma; 1+\alpha-\alpha'; \frac{(z-z_1)(z_3-z_2)}{(z-z_2)(z_3-z_1)}\right). \tag{9.49}$$

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. Clearly, 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

$$(z-z_1)^r (z-z_2)^s (z-z_3)^t P \left\{ \begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array} \right\} = P \left\{ \begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha+r & \beta+s & \gamma+t & z \\ \alpha'+r & \beta'+s & \gamma'+t & \end{array} \right\} \tag{9.50}$$

provided $r+s+t=0$. Also Riemann's equation retains its form under Möbius maps, only the location of the singular points changing. We therefore deduce that

$$P \left\{ \begin{array}{cccc} z_1 & z_2 & z_3 & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{array} \right\} = P \left\{ \begin{array}{cccc} z_1' & z_2' & z_3' & \\ \alpha & \beta & \gamma & z' \\ \alpha' & \beta' & \gamma' & \end{array} \right\} \tag{9.51}$$

where

$$z' = \frac{az+b}{cz+d}, \quad z_1' = \frac{az_1+b}{cz_1+d}, \quad z_2' = \frac{az_2+b}{cz_2+d}, \quad z_3' = \frac{az_3+b}{cz_3+d}. \tag{9.52}$$

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

$$F(a,b;c;z) = P \left\{ \begin{array}{cccc} 0 & \infty & 1 & \\ 0 & a & 0 & z \\ 1-c & b & c-a-b & \end{array} \right\}. \tag{9.53}$$

Using this observation and a suitable Möbius map we see that

$$F(a,b;a+b-c;1-z)$$

[2]The interchange $\beta \leftrightarrow \beta'$ leaves the hypergeometric function invariant, and so does not give a new solution.

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

$$F(a,b;c;1)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \tag{9.54}$$

together with the trick of substituting $z=0$ and $z=1$, to determine the coefficients and show that

$$\begin{aligned} F(a,b;c;x) &= \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}F(a,b;a+b-c;1-z) \\ &\quad +\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}(1-z)^{c-a-b}F(c-b,c-a;c-a-b+1;1-z). \end{aligned} \tag{9.55}$$

9.3 Solving ODE's via Contour integrals

Our task in this section is to understand the origin of contour integral solutions such as the expression

$$F(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1(1-tz)^{-a}t^{b-1}(1-t)^{c-b-1}dt,$$

we have previously seen for the hypergeometric equation.

We are given a differential operator

$$L_z=\partial_{zz}^2+p(z)\partial_z+q(z)$$

and seek a solution of $L_zu=0$ as an integral

$$u(z)=\int_\Gamma F(z,t)\,dt.$$

If we can find an F such that

$$L_zF=\frac{\partial Q}{\partial t},$$

for some function $Q(z,t)$ then

$$L_z u = \int_\Gamma L_z F(z,t)\,dt = \int_\Gamma \left(\frac{\partial Q}{\partial t}\right)\,dt = [Q]_\Gamma\,.$$

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, then we have succeeded.

Example: Consider Legendre's equation

$$L_z u \equiv (1-z^2)\frac{d^2u}{dz^2} - 2z\frac{du}{dz} + \nu(\nu+1)u = 0.$$

The identity

$$L_z\left\{\frac{(t^2-1)^\nu}{(t-z)^{\nu+1}}\right\} = (\nu+1)\frac{d}{dt}\left\{\frac{(t^2-1)^{\nu+1}}{(t-z)^{\nu+2}}\right\}$$

shows that

$$u(z) = \frac{1}{2\pi i}\int_\Gamma \left\{\frac{(t^2-1)^\nu}{(t-z)^{\nu+1}}\right\}\,dt$$

will be a solution of Legendre's equation provided that

$$\left[\frac{(t^2-1)^{\nu+1}}{(t-z)^{\nu+2}}\right]_\Gamma = 0.$$

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 aquires a phase of $e^{2\pi i(\nu+2)}$. The net phase 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

$$u(z) = \frac{1}{n!}\frac{d^n}{dz^n}(z^2-1)^n,$$

which is (up to factor) Rodriguez' formula for the Legendre polynomials.

It is hard to find a suitable F in one fell swoop. (The identity exploited in the above 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

$$u(z) = \int_a^b K(z,t)v(t)\,dt.$$

Suppose that $L_z\,K(z,t) = M_t\,K(z,t)$, where M_t is differential operator in t which does not involve z. The operator M_t will have have a formal adjoint $M_t^\dagger$ such that

$$\int_a^b v(M_t K)\,dt - \int_a^b K(M_t^\dagger v)\,dt = \left[Q(K,v)\right]_a^b.$$

(This is Lagrange's identity from last semester.) 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

$$L_z u = \frac{d^2u}{dz^2} - z\frac{du}{dz} + \nu u = 0,$$

by using the kernel $K(z,t) = e^{-zt}$. We have $L_z K(z,t) = M_t K(z,t)$ where

$$M_t = t^2 - t\frac{\partial}{\partial t} + \nu,$$

so

$$M_t^\dagger = t^2 + \frac{\partial}{\partial t}t + \nu = t^2 + (\nu+1) + t\frac{\partial}{\partial t}.$$

The equation $M_t^\dagger v = 0$ has solution

$$v(t) = t^{-(\nu+1)} e^{-\frac{1}{2}t^2},$$

and so

$$u = \int_\Gamma t^{-(1+\nu)} e^{-(zt+\frac{1}{2}t^2)}\,dt,$$

for some suitable Γ.

9.3.1 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

$$zy'' + (c - z)y' - ay = 0,$$

with solution

$$\Phi(a, c; z) = \frac{\Gamma(c)}{\Gamma(a)} \sum_{n=0}^{\infty} \frac{\Gamma(a + n)}{\Gamma(c + n)\Gamma(n + 1)} z^n.$$

The second solution, when c is not an integer, is

$$z^{1-c}\Phi(a - c + 1, 2 - c; z).$$

We see that

$$\Phi(a, c; z) = \lim_{b\to\infty} F(a, b; c; z/b).$$

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*

$$\operatorname{erf}(z) = \int_0^z e^{-t^2}\, dt = z\Phi\left(\frac{1}{2}, \frac{3}{2}; -z^2\right)$$

and the *Laguerre polynomials*

$$L_n^m = \frac{\Gamma(n + m + 1)}{\Gamma(n + 1)\Gamma(m + 1)}\Phi(-n, m + 1; z).$$

Bessel's equation involves

$$L_z = \partial_{zz}^2 + \frac{1}{z}\partial_z + \left(1 - \frac{\nu^2}{z^2}\right).$$

Experience shows that a useful kernel is

$$K(z, t) = \left(\frac{z}{2}\right)^{\nu} \exp\left(t - \frac{z^2}{4t}\right).$$

Then

$$L_z K(z,t) = \left(\partial_t - \frac{\nu+1}{t}\right) K(z,t)$$

so M is a first order operator, which is simpler to deal with than the original second order L_z. In this case

$$M^\dagger = \left(-\partial_t - \frac{\nu+1}{t}\right)$$

and we need a v such that

$$M^\dagger v = -\left(\partial_t + \frac{\nu+1}{t}\right) v = 0.$$

Clearly $v = t^{-\nu-1}$ will work. The integrated out part is

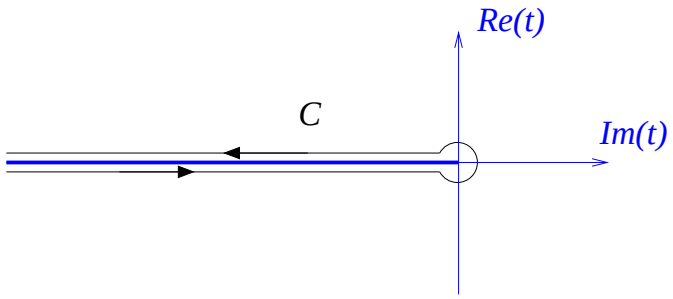
$$[Q(K,v)]_a^b = \left[t^{-\nu-1} \exp\left(t - \frac{z^2}{4t}\right)\right]_a^b.$$

We see that

$$J_\nu(z) = \frac{1}{2\pi i}\left(\frac{z}{2}\right)^\nu \int_C t^{-\nu-1} e^{\left(t - \frac{z^2}{4t}\right)}\, dt.$$

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.



This works because Q is zero at both ends of the contour.

A cosmetic rewrite $t = uz/2$ gives

$$J_\nu(z) = \frac{1}{2\pi i}\int_C u^{-\nu-1} e^{\frac{z}{2}\left(u - \frac{1}{u}\right)}\, du.$$

For ν an integer, there is no discontinuity across the cut, so we can ignore it and take C to be the unit circle. From

$$J_n(z) = \frac{1}{2\pi i}\int_C u^{-n-1} e^{\frac{z}{2}\left(u-\frac{1}{u}\right)}\, du.$$

we get the usual generating function

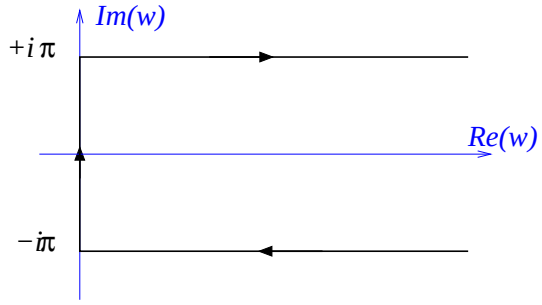
$$e^{\frac{z}{2}\left(u-\frac{1}{u}\right)} = \sum_{-\infty}^{\infty} J_n(z) u^n.$$

When ν is not an integer, we see why we need a branch cut integral.

If we set $u = e^w$ we get

$$J_\nu(z) = \frac{1}{2\pi i}\int_{C'} dw\, e^{z\sinh w - \nu w},$$

where C' starts goes from $\infty - i\pi$ to $-i\pi$, to $+i\pi$ to $\infty + i\pi$.



If we set $w = t \pm i\pi$ on the horizontals and $w = i\theta$ on the vertical part, we can rewrite this as

$$J_\nu(z) = \frac{1}{\pi}\int_0^\pi \cos(\nu\theta - z\sin\theta)\, d\theta - \frac{\sin\nu\pi}{\pi}\int_0^\infty e^{-\nu t - z\sinh t}\, dt.$$

All these are standard formulae for the Bessel function whose origin would be hard to understand without the contour solutions trick.

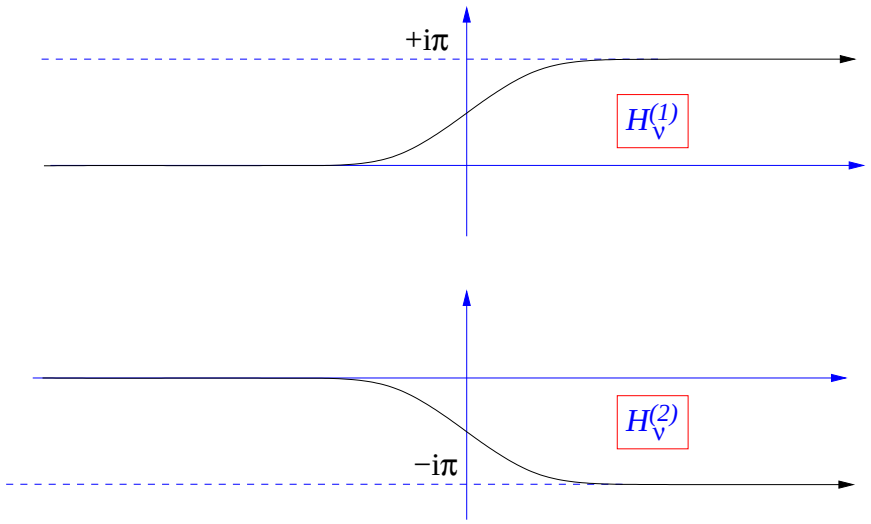
When ν becomes an integer, the functions $J_\nu(z)$ and $J_{-\nu(z)}$ are no longer independent. In order to have a pair of functions that retain their independence even as ν becomes a whole number, it is traditional to define

$$\begin{aligned} Y_\nu(z) &\stackrel{def}{=} \frac{J_\nu(z)\cos\nu\pi - J_{-\nu}(z)}{\sin\nu\pi} \\ &= \frac{\cot\nu\pi}{\pi}\int_0^\pi \cos(\nu\theta - z\sin\theta)\, d\theta - \operatorname{cosec}\nu\pi\pi\int_0^\pi \cos(\nu\theta + z\sin\theta)\, d\theta \\ &\quad -\frac{\cos\nu\pi}{\pi}\int_0^\infty e^{-\nu t - z\sinh t}\, dt - \frac{1}{\pi}\int_0^\infty e^{\nu t - z\sinh t}\, dt. \end{aligned}$$

These functions are real for positive real z and oscillate as slowly decaying sines and cosines.

It is often convenient to decompose these real functions into functions that behave as $e^{\pm iz}$, and so we define the *Hankel functions* by

$$\begin{aligned} H_\nu^{(1)}(z) &= \frac{1}{i\pi}\int_{-\infty}^{\infty+i\pi} e^{z\sinh w-\nu w}\,dw, \qquad |\arg z|<\pi/2 \\ H_\nu^{(2)}(z) &= -\frac{1}{i\pi}\int_{-\infty}^{\infty-i\pi} e^{z\sinh w-\nu w}\,dw, \qquad |\arg z|<\pi/2. \end{aligned}$$



Contours defining $H_\nu^{(1)}(z)$ and $H_\nu^{(2)}(z)$.

Then

$$\begin{aligned} \frac{1}{2}(H_\nu^{(1)}(z)+H_\nu^{(2)}(z)) &= J_\nu(z), \\ \frac{1}{2}(H_\nu^{(1)}(z)-H_\nu^{(2)}(z)) &= Y_\nu(z). \end{aligned} \tag{9.56}$$

9.4 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

$$Z(\lambda) = \int_{-\infty}^{\infty} e^{-x^2-\lambda x^4}\, dx.$$

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

$$\begin{aligned} Z(\lambda) &= \int_{-\infty}^{\infty} e^{-x^2-\lambda x^4}\, dx \\ &= \int_{-\infty}^{\infty} e^{-x^2} \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n x^{4n}}{n!}\, dx \\ &\stackrel{?}{=} \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!} \int_{-\infty}^{\infty} e^{-x^2} x^{4n}\, dx \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!}\, \Gamma(2n+1/2). \end{aligned}$$

Something has clearly gone wrong here, because $\Gamma(2n+1/2) \sim (2n)! \sim 4^n (n!)^2$, and so the radius of convergence of the 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 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

$$Z(\lambda) \sim \sum_{n=0}^{\infty} (-1)^n \frac{\lambda^n}{n!} \Gamma(2n+1/2)$$

where

$$Z(\lambda) \sim \sum_{n=0}^{\infty} a_n \lambda^n$$

is shorthand for the more explicit

$$Z(\lambda) = \sum_{n=0}^{N} a_n \lambda^n + O\left(\lambda^{N+1}\right), \quad N = 1, 2, 3, \ldots.$$

The "big O" notation

$$Z(\lambda) - \sum_{n=0}^{N} a_n \lambda^n = O(\lambda^{N+1})$$

as $\lambda \to 0$, means that

$$\lim_{\lambda \to 0} \left\{ \frac{|Z(\lambda) - \sum_0^N a_n \lambda^n|}{|\lambda^{N+1}|} \right\} = K < \infty.$$

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:

$$F(x) = \sum_{n=0}^{N} b_n x^{-n} + O\left(x^{-N-1}\right), \quad N = 1, 2, 3 \ldots. \tag{9.57}$$

In this case

$$F(x) - \sum_{n=0}^{N} b_n x^{-n} = O\left(x^{-N-1}\right) \tag{9.58}$$

means that

$$\lim_{x \to \infty} \left\{ \frac{|F(x) - \sum_0^N b_n x^{-n}|}{|x^{-N-1}|} \right\} = K < \infty. \tag{9.59}$$

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.

9.4.1 Stirling's Approximation for $n!$

We start from the integral representation of the Gamma function

$$\Gamma(z+1) = \int_0^\infty e^{-t}t^z\,dt$$

Set $t = z\zeta$, so

$$\Gamma(z+1) = z^{z+1}\int_0^\infty e^{zf(\zeta)}\,d\zeta,$$

where

$$f(\zeta) = \ln\zeta - \zeta.$$

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

$$f(\zeta) = -1 - \frac{1}{2}(\zeta-1)^2 + \cdots.$$

So

$$\begin{aligned}\Gamma(z+1) &= z^{z+1}e^{-z}\int_0^\infty e^{-\frac{z}{2}(\zeta-1)^2+\cdots}\,d\zeta \\ &\approx z^{z+1}e^{-z}\int_{-\infty}^\infty e^{-\frac{z}{2}(\zeta-1)^2}\,d\zeta \\ &= z^{z+1}e^{-z}\sqrt{\frac{2\pi}{z}} \\ &= \sqrt{2\pi}z^{z+1/2}e^{-z}. \end{aligned} \tag{9.60}$$

By keeping more of the terms represented by the dots, and expanding them as

$$e^{-\frac{z}{2}(\zeta-1)^2+\cdots} = e^{-\frac{z}{2}(\zeta-1)^2}\left[1+a_1(\zeta-1)+a_2(\zeta-1)^2+\cdots\right], \tag{9.61}$$

we would find, on doing the integral, that

$$\Gamma(z+1) \approx \sqrt{2\pi}z^{z+1/2}e^{-z}\left[1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{24888320z^4}+O\left(\frac{1}{z^5}\right)\right]. \tag{9.62}$$

Since $\Gamma(n+1)=n!$ we also have

$$n! \approx \sqrt{2\pi}n^{n+1/2}e^{-n}\left[1+\frac{1}{12n}+\cdots\right].$$

We make contact with our discusion of asymptotic series by rewriting the expansion as

$$\frac{\Gamma(z+1)}{\sqrt{2\pi}z^{z+1/2}e^{-z}} \sim 1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{24888320z^4}+\ldots \tag{9.63}$$

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.

9.4.2 Airy Functions

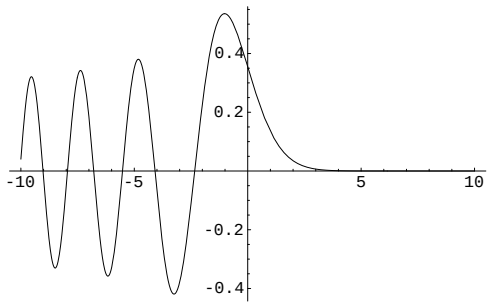
A more sophisticated treatment is needed for this problem, and we will meet with Stokes' phenomenon. Airy's equation is

$$y''-zy=0.$$

On the real axis this becomes

$$-y'' + xy = 0,$$

which we can think of 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 have a turning point at $x = 0$. 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.



The Airy function, $\mathrm{Ai}\,(x)$.

We will look for contour integral solutions to Airy's equation of the form

$$y(x) = \int_a^b e^{xt} f(t)\, dt.$$

Denoting the Airy differential operator by $L_x \equiv \partial_x^2 - x$, we have

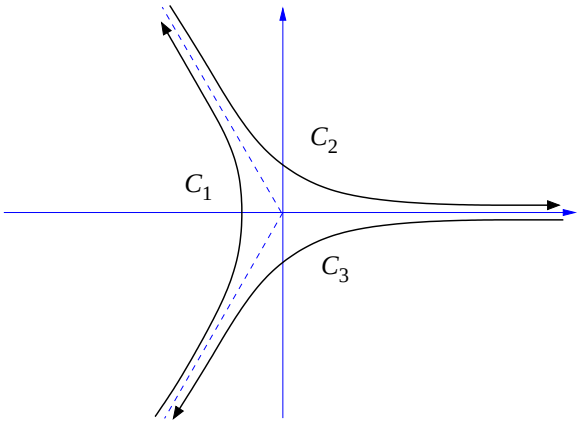
$$\begin{aligned} L_x\, y &= \int_a^b (t^2 - x) e^{xt} f(t)\, dt = \int_a^b f(t) \left\{ t^2 - \frac{d}{dt} \right\} e^{xt}\, dt. \\ &= \left[-e^{xt} f(t) \right]_a^b + \int_a^b \left(\left\{ t^2 + \frac{d}{dt} \right\} f(t) \right) e^{xt}\, dt. \end{aligned}$$

Thus $f(t) = e^{-\frac{1}{3}t^3}$ and

$$y(x) = \int_a^b e^{xt - \frac{1}{3}t^3}\, dt.$$

The contour must end at points where the integrated-out term, $\left[e^{xt - \frac{1}{3}t^3} \right]_a^b$, vanishes. There are therefore three possible contours, which end at any two of

$$+\infty, \quad \infty\, e^{2\pi i/3}, \quad \infty\, e^{-2\pi i/3}.$$



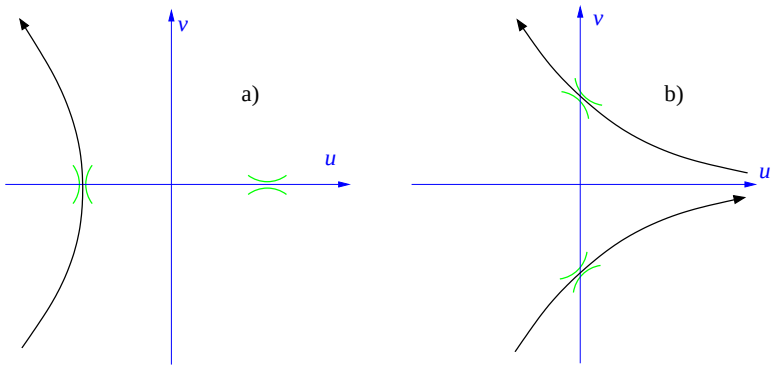
Contours providing solutions of Airy's equation.

Of course $y_{C_1} + y_{C_2} + y_{C_3} = 0$, so only two are linearly independent. The Airy function itself is defined by

$$\mathrm{Ai}\,(z) = \frac{1}{2\pi i}\int_{C_1} e^{xt-\frac{1}{3}t^3}\,dt = \frac{1}{\pi}\int_0^\infty \cos\left(xs + \frac{1}{3}s^3\right)\,ds$$

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.



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

$$f(\xi) = -\frac{2}{3}x^{3/2} + \xi^2\sqrt{x} - \frac{1}{3}\xi^3$$

while near $t = +\sqrt{x}$ we write $t = \zeta + \sqrt{x}$ and find

$$f(\zeta) = -\frac{2}{3}x^{3/2} - \zeta^2\sqrt{x} - \frac{1}{3}\zeta^3$$

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

$$\mathrm{Im}\,(xt - t^3/3) = v(x - u^2 + v^3/3),$$

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

$$u^2 - \frac{1}{3}v^2 = x.$$

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

$$\mathrm{Ai}\,(x) = \frac{1}{2\pi} e^{-\frac{2}{3}x^{3/2}} \int_{-\infty}^{\infty} e^{-\sqrt{x}s^2 + \cdots}\, ds \sim \frac{1}{2\sqrt{\pi}} x^{-1/4} e^{-\frac{2}{3}x^{3/2}}.$$

ii) $x \ll 0$: The stationary points are now at $\pm i\sqrt{|x|}$. Setting $t = \xi \pm i\sqrt{|x|}$ find that

$$f(x) = \mp i\frac{2}{3}|x|^{3/2} \mp i\xi^2\sqrt{|x|}.$$

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

$$\begin{aligned} \frac{1}{2\pi i} e^{3\pi i/4} e^{-i\frac{2}{3}|x|^{3/2}} \int_{-\infty}^{\infty} e^{-\sqrt{x}s^2}\, ds &= \frac{1}{2i\sqrt{\pi}} e^{3\pi i/4} |x|^{-1/4} e^{-i\frac{2}{3}|x|^{3/2}} \\ &= -\frac{1}{2i\sqrt{\pi}} e^{-i\pi/4} |x|^{-1/4} e^{-i\frac{2}{3}|x|^{3/2}} \end{aligned} \tag{9.64}$$

Near $t = -i\sqrt{|x|}$we set $\xi = e^{2\pi i/3}s$ and get

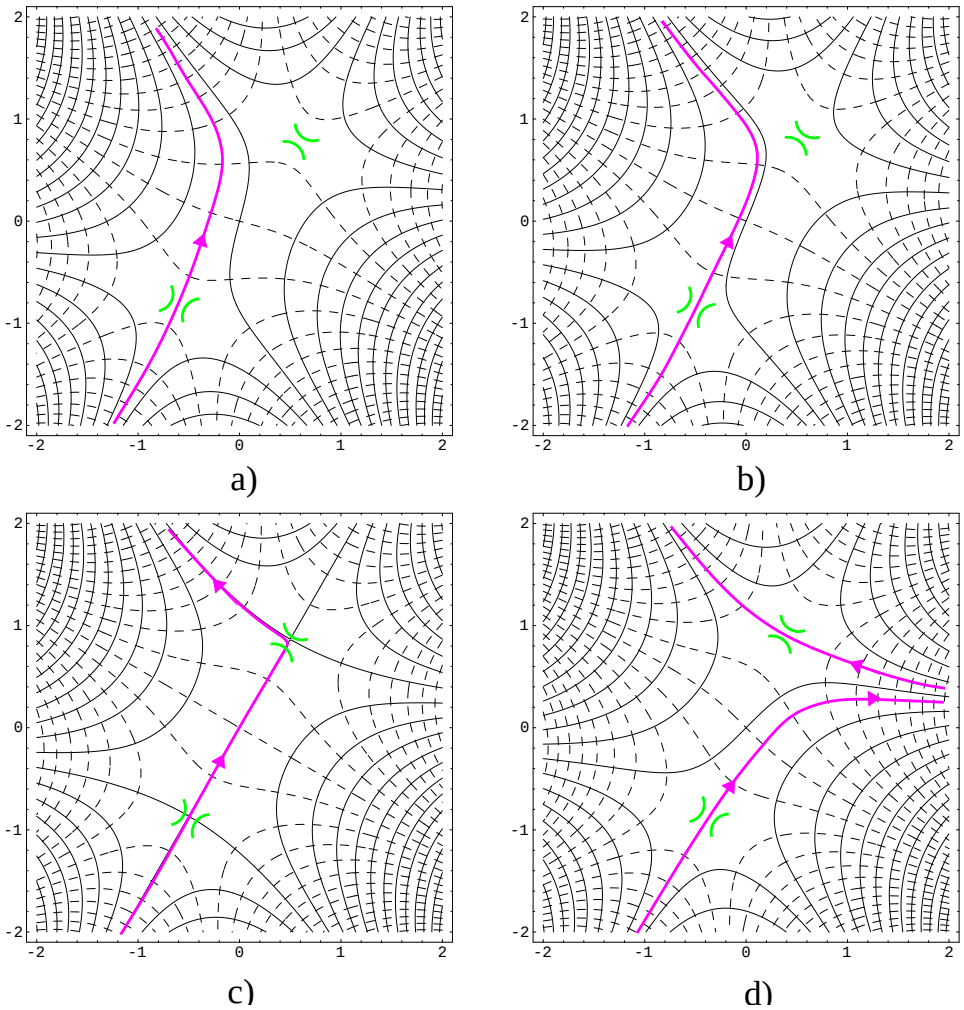
$$\frac{1}{2i\pi} e^{\pi i/4} e^{i\frac{2}{3}|x|^{3/2}} \int_{-\infty}^{\infty} e^{-\sqrt{x}s^2}\, ds = \frac{1}{2i\sqrt{\pi}} e^{\pi i/4} |x|^{-1/4} e^{i\frac{2}{3}|x|^{3/2}}$$

The sum of these two contributions is

$$\mathrm{Ai}\,(x) \sim \frac{1}{\sqrt{\pi}|x|^{1/4}} \sin\left(\frac{2}{3}|x|^3/2 + \frac{\pi}{4}\right).$$

The fruit of our labours is therefore

$$\begin{aligned} \mathrm{Ai}\,(x) &\sim \frac{1}{2\sqrt{\pi}} x^{-1/4} e^{-\frac{2}{3}x^{3/2}} \left[1 + O\left(\frac{1}{x}\right)\right], \quad x > 0, \\ &\sim \frac{1}{\sqrt{\pi}|x|^{1/4}} \sin\left(\frac{2}{3}|x|^3/2 + \frac{\pi}{4}\right) \left[1 + O\left(\frac{1}{x}\right)\right], \quad x < 0. \end{aligned}$$



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)* $9\pi/12$.

Suppose that we allow x to become complex $x \to z = |z|e^{i\theta}$, with $-\pi < \theta < \pi$. Then the figure above 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

$$\begin{aligned}\mathrm{Ai}\,(-z) &\sim \pi^{-1/2} z^{-1/4} \left[\sin(\rho + \pi/4) \sum_{n=0}^{\infty} (-1)^n c_{2n} \rho^{-2n} \right. \\ &\quad \left. - \cos(\rho + \pi/4) \sum_{n=0}^{\infty} (-1)^n c_{2n+1} \rho^{-2n-1} \right] \qquad (9.65)\end{aligned}$$

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*.

Exercise: 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

$$\begin{aligned}J_\nu(x) &\sim \sqrt{\frac{2}{\pi x}} \left[\cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) - \frac{4\nu^2 - 1}{8x} \sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) + \cdots \right], \\ Y_\nu(x) &\sim \sqrt{\frac{2}{\pi x}} \left[\sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) + \frac{4\nu^2 - 1}{8x} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) + \cdots \right].\end{aligned}$$

9.5 Elliptic Functions

The subject of elliptic functions goes back to remarkable identities of Fagano (1750) and Euler (1761). Euler's formula is

$$\int_0^u \frac{dx}{\sqrt{1-x^4}} + \int_0^v \frac{dy}{\sqrt{1-y^4}} = \int_0^r \frac{dz}{\sqrt{1-z^4}},$$

where $0 \le u, v \le 1$, and

$$r = \frac{u\sqrt{1-v^4} + v\sqrt{1-u^4}}{1+u^2v^2}.$$

This looks mysterious, but perhaps so does

$$\int_0^u \frac{dx}{\sqrt{1-x^2}} + \int_0^v \frac{dy}{\sqrt{1-y^2}} = \int_0^r \frac{dz}{\sqrt{1-z^2}},$$

where

$$r = u\sqrt{1-v^2} + v\sqrt{1-u^2},$$

until you realize that the latter formula is merely

$$\sin(a+b) = \sin a \cos b + \cos a \sin b$$

in disguise. To see this set

$$u = \sin a, \quad v = \sin b$$

and remember the integral formula for the inverse trig function

$$a = \sin^{-1} u = \int_0^u \frac{dx}{\sqrt{1-x^2}}.$$

The Fagano-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

$$I_4 = \int_0^x \frac{dt}{\sqrt{(t-a_1)(t-a_2)(t-a_3)(t-a_4)}}$$

$$I_3 = \int_0^x \frac{dt}{\sqrt{(t-a_1)(t-a_2)(t-a_3)}}.$$

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

$$t \to t' = \frac{at+b}{ct+d}, \quad dt' = (ad-bc)\frac{dt}{(ct+d)^2}.$$

Indeed, we can use a suitable Möbius map to send any three of the four points 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).

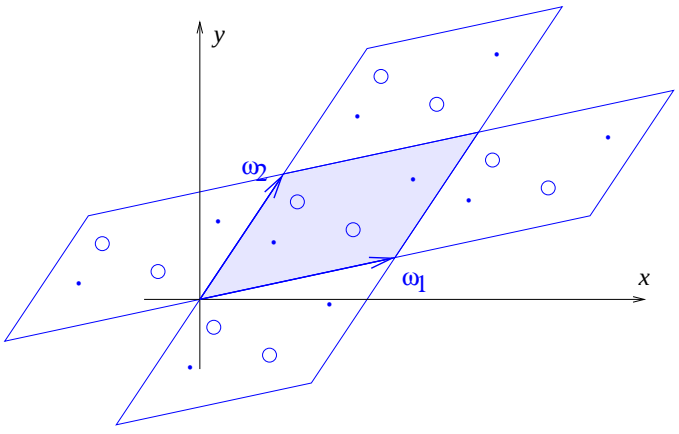
For the general theory, the simplest elliptic function is the Weierstrass $\mathcal{P}$. This is defined by first selecting two linearly independent *periods* ω_1, ω_2, and setting

$$\mathcal{P}(z) = \frac{1}{z^2} + \sum_{m,n\neq 0} \left\{ \frac{1}{(z-m\omega_1-n\omega_2)^2} - \frac{1}{(m\omega_1+n\omega_2)^2} \right\}.$$

The sum is over all non-negative integers m, n, positive and negative. Helped by the counterterm, the sum is absolutely convergent. We can therefore rearrange the terms to prove double periodicity

$$\mathcal{P}(z+m\omega_1+n\omega_2) = \mathcal{P}(z)$$

The function is therefore 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.



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 $\mathcal{P}$ has a second order pole at the origin. It also obeys

$$\begin{aligned} \lim_{|z|\to 0}\left(\mathcal{P}(z) - \frac{1}{z^2}\right) &= 0 \\ \mathcal{P}(z) &= \mathcal{P}(-z) \\ \mathcal{P}'(z) &= -\mathcal{P}'(-z) \end{aligned}$$

The property that makes $\mathcal{P}$ useful for evaluating integrals is

$$\left(\mathcal{P}'(z)\right)^2 = 4\mathcal{P}^3(z) - g_2\mathcal{P}(z) - g_3$$

where

$$g_2 = 60 \sum_{m,n\neq 0} \frac{1}{(m\omega_1 + n\omega_2)^4}, \quad g_3 = 140 \sum_{m,n\neq 0} \frac{1}{(m\omega_1 + n\omega_2)^6}.$$

This is proved by observing that the difference of the left hand and right hand sides is zero at $z = 0$, has no poles or other singularities, and being therefore continuous and periodic is automatically bounded. It is therefore identically zero by Liouville's theorem.

From the symmetry and periodicity of $\mathcal{P}$ we see that $\mathcal{P}'(z) = 0$ at $e_1 = \mathcal{P}(\omega_1/2)$, $e_2 = \mathcal{P}(\omega_2/2)$, and $e_3 = \mathcal{P}((\omega_1+\omega_2)/2)$. Now $\mathcal{P}'$ 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

$$4\mathcal{P}^3(z) - g_2\mathcal{P}(z) - g_3 = 4(\mathcal{P} - e_1)(\mathcal{P} - e_2)(\mathcal{P} - e_3).$$

We note that the coefficient of $\mathcal{P}^2$ 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, $(\mathcal{P}(z) - e_1)$ 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 $(\mathcal{P}(z) - e_1)$ has a double pole at 0 as its only singularity. Thus $(\mathcal{P} - e_1)$ 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*

$$\Delta = 16(e_1 - e_2)^2(e_2 - e_3)^2(e_1 - e_3)^2 = g_2^3 - 27g_3^2,$$

is never zero.

We use $\mathcal{P}$ to write

$$z = \mathcal{P}^{-1}(u) = \int_\infty^u \frac{dt}{2\sqrt{(t - e_1)(t - e_2)(t - e_3)}} = \int_\infty^u \frac{dt}{\sqrt{4t^3 - g_2t - g_3}}.$$

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 $x = \mathcal{P}(z)$, $y = \mathcal{P}'(z)$ move on the *elliptic curve*

$$y^2 = 4x^3 - g_2x - g_3$$

which should be thought of as a set in CP^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-Fagano relation with which we began this section, lies in the (not immediatley obvious) fact that any elliptic function having the same periods as $\mathcal{P}(z)$ can be expressed as a rational function of $\mathcal{P}(z)$ and $\mathcal{P}'(z)$. 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

$$F(x, y) = \sum a_{n,m}x^n y^m$$

is a polynomial in x and y. We can eliminate $\mathcal{P}'(z)$ 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 is because 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 preserve the orientation of the basis vectors then we must restrict ourselves to maps whose determinant $ad - bc$ is unity. The set of such transforms constitute the the group $SL(2, \mathbf{Z})$. Clearly $\mathcal{P}$ is invariant under this group, as are g_2 and g_3 and Δ. Now define $\omega_2/\omega_1 = \tau$, and write

$$g_2(\omega_1, \omega_2) = \frac{1}{\omega_1^4}, \tilde{g}_2(\tau), \quad g_3(\omega_1, \omega_2) = \frac{1}{\omega_1^6}, \tilde{g}_3(\tau). \quad \Delta(\omega_1, \omega_2) = \frac{1}{\omega_1^{12}} \tilde{\Delta}(\tau),$$

and also

$$J(\tau) = \frac{\tilde{g}_2^3}{\tilde{g}_2^3 - 27\tilde{g}_3^2} = \frac{\tilde{g}_2^3}{\tilde{\Delta}}.$$

Because the denominator is never zero when $\operatorname{Im} \tau > 0$, 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

$$\tau \to \frac{a\tau + b}{c\tau + d}.$$

with

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z}).$$

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* $PSL(2, \mathbf{Z}) \equiv SL(2, \mathbf{Z})/\{I, -I\}$.

Taking into accound the change in the prefactors we have

$$\begin{aligned}
\tilde{g}_2\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^4\tilde{g}_3(\tau), \\
\tilde{g}_3\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^6\tilde{g}_3(\tau), \\
\tilde{\Delta}\left(\frac{a\tau+b}{c\tau+d}\right) &= (c\tau+d)^{12}\tilde{\Delta}(\tau). \qquad (9.66)
\end{aligned}$$

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

$$\tilde{\Delta}(\tau) = (2\pi)^{12}q^2\prod_{n=1}^{\infty}\left(1-q^{2n}\right)^{24}.$$

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

$$\begin{aligned}
\tilde{g}_2(\tau) &= (2\pi)^4\left[\frac{1}{12}+20\sum_{n=1}^{\infty}\sigma_3(n)q^{2n}\right], \\
\tilde{g}_3(\tau) &= (2\pi)^6\left[\frac{1}{216}-\frac{7}{3}\sum_{n=1}^{\infty}\sigma_5(n)q^{2n}\right]. \qquad (9.67)
\end{aligned}$$

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

$$J\left(\frac{a\tau+b}{c\tau+d}\right) = J(\tau),$$

so $J(\tau)$ is a *modular invariant*. One can show that if $J(\tau_1) = J(\tau_2)$, then

$$\tau_2 = \frac{a\tau_1+b}{c\tau_1+d}$$

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 $J(\tau)$. Thus $J(\tau)$ is a rather special object.

This $J(\tau)$ is the function referred to in the footnote about the properties of 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

$$1728J(\tau) = q^{-2} + 744 + 196884q^2 + 21493760q^4 + 864299970q^6 + \cdots.$$

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. The remaining integer coefficiants, however, are completely determined by these properties. A number theory interpretation of these integers seemed lacking until John McKay and others observed that that

$$\begin{aligned} 1 &= 1 \\ 196884 &= 1 + 196883 \\ 21493760 &= 1 + 196883 + 21296786 \\ 864299970 &= 2 \times 1 + 2 \times 196883 + 21296786 + 842609326, \end{aligned}$$

where "1" and the large integers on the right-hand side are the dimensions of the smallest irreducible representations of the Monster group. This "Monstrous Moonshine" was originally mysterious and almost unbelievable, ("moonshine" = "fanatstic 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 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).