Notes for Course on Mathematical Methods of Physics, CMI, Spring 2025

Govind S. Krishnaswami, April 30, 2025

Please let me know at govind@cmi.ac.in of any comments or corrections

Course website http://www.cmi.ac.in/~govind/teaching/math-meth-e25

Contents

1 Complex function theory

Complex variables and complex analysis are useful in many parts of classical physics (e.g., oscillation problems, incompressible irrotational planar flow, Laplace-Fourier transforms in solving differential equations, etc.). Complex numbers enter the very formulation of quantum mechanics (the i in the Schrödinger equation, wave functions being elements of a complex vector space, etc.) and complex analysis finds use in diverse areas of quantum physics (dispersion relations, analytic continuation in complex energy and complex angular momentum in scattering theory (Regge poles), coherent states, Wick rotation and the relation between statistical and quantum mechanics, etc.)

• A Cauchy (1789-1857) and B Riemann (1826-1866) played important roles in the initial development of complex function theory. While Cauchy's approach was analytic, Riemann's had a geometric flavor. Cauchy also did a lot of work on solid mechanics and elasticity theory.

1.1 Some references on complex function theory and applications

1. L V Ahlfors, Complex Analysis

2. Dennery and Krzywicki, Mathematics for Physicists

3. J W Dettman, Applied Complex Variables

4. Byron and Fuller, Mathematics of Classical and Quantum Physics

5. Stone and Goldbart, Mathematics for Physics
6. Ablowitz and Fokas, Complex Variables: Introduction and Applications

1.2 Real numbers

• Real numbers are familiar to us. They are numbers with a decimal expansion e.g., $1.0423946\ldots$ and may be thought of as points along a line.
• In addition to being a **vector space**, the set $\mathbb{R}$ of real numbers is a **field**. We have **commutative operations of addition and multiplication** ($a + b = b + a$, $ab = ba$) with the latter distributing over addition: $a(b + c) = ab + ac$. Zero and one are the additive and multiplicative identities: $a + 0 = a$ and $a\,1 = a$ for every $a \in \mathbb{R}$. Every real number a has an additive inverse $-a$ while nonzero reals have multiplicative inverses $a^{-1} = 1/a$, allowing us to divide by nonzero real numbers.
• The real number field is **totally ordered**: given distinct real numbers $a \neq b$, either $a < b$ or $b < a$ with the transitivity property that if $a < b$ and $b < c$ then $a < c$. Ordered sets come up in physics: the set of events in Minkowski space-time are partially ordered by the property of lying in the causal past or future. Event E' is in the future of E if E' lies in the future lightcone of E so that signals can reach E' from E. It is only partially ordered since a pair of events may not be causally connected at all.
• Real numbers form a **metric space**. The distance between a and b is the absolute value[1] of their difference $d(a, b) = |a-b|$. The metric is a positive definite symmetric function $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfying the triangle inequality:

$$\begin{array}{ll} (1) & d(a,b) \geq 0 \quad \text{with} \quad d(a,b) = 0 \;\text{ iff }\; a = b, \\ (2) & d(a,b) = d(b,a) \quad \text{and} \\ (3) & d(a,c) \leq d(a,b) + d(b,c), \end{array} \qquad (1)$$

for any three reals a, b, c.
• In fact, the real numbers form a **complete metric space**. This means that every **Cauchy sequence** of real numbers converges to a real number. We will explain what a Cauchy sequence is in §1.5.

1.3 Complex numbers

• A complex number is an ordered pair of real numbers $z = (x, y)$. So they may be viewed as points on the two-dimensional Euclidean plane, which is called the complex or Argand plane. Commutative addition and multiplication are defined using those of real numbers:

$$\begin{array}{rcl} (x,y) + (x',y') &=& (x + x', y + y') \quad \text{and} \\ (x,y)(x',y') &=& (xx' - yy', xy' + yx'). \end{array} \qquad (2)$$

[1] The absolute value is an example of a norm on a vector space: a positive definite function satisfying $||\lambda a|| = |\lambda| ||a||$ (for any scalar λ and any vector a) and the triangle inequality $||a + b|| \leq ||a|| + ||b||$ for any vectors a and b.

While $(0,0)$ is the additive identity, $(1,0)$ is the multiplicative identity $(x,y)(1,0) = (x,y)$. So we denote these by 0 and 1. Since $\mathbb{C}$ may be regarded as a two dimensional vector space, it is convenient to have a basis for it. The natural choice of basis elements is $(1,0) = 1$ and $(0,1)$. The latter has the interesting property $(0,1)(0,1) = (-1,0) = -1$. The special element $(0,1)$ is denoted i and we have just seen that $i^2 = -1$. Thus, any complex number can be written as a linear combination $z = (x,y) = x(1,0) + y(0,1) = x + iy$. The real numbers x and y are called the real and imaginary parts: $\Re z = x$ and $\Im z = y$. One checks that the set of complex numbers $\mathbb{C}$ forms a **field**. The multiplicative inverse of any nonzero complex number $z = x + iy$ is $(x - iy)/(x^2 + y^2)$.

• Unlike the reals, the complex numbers are **not an ordered field**: we do not have a way of saying whether z is larger or smaller than z' unless they both happen to be real.

• Nevertheless, by viewing them as points on the plane $\mathbb{R}^2$, there is a natural notion of magnitude. The absolute value of $z = x + iy$ is the Euclidean distance of (x,y) from the origin: $|z| = \sqrt{x^2 + y^2}$. The **absolute value** or **modulus** defines a **norm** on $\mathbb{C}$.

• We may use the absolute value to turn $\mathbb{C}$ into a **metric space** by defining the distance function $d(z,z') = |z - z'| = \sqrt{(x-x')^2 + (y-y')^2}$. This is simply the Euclidean distance between points on the plane. It satisfies the triangle inequality as a consequence of the corresponding property of plane triangles.

• Using polar coordinates on the complex plane, we introduce the **modulus-argument** form of a nonzero complex number: $z = re^{i\theta}$. Here $r = |z| = \sqrt{x^2 + y^2}$ is called the modulus while the argument $\arg z = \theta = \arctan(y/x)$. We note that $\arg z$ is a **multivalued function** of z, it is any angle θ such that $x = r\cos\theta$ and $y = r\sin\theta$. Evidently, any two such angles differ by an integer multiple of 2π. When $z = 0$, the modulus vanishes and the argument is not defined or can be taken to be arbitrary.

• A key distinction between real and complex numbers is that the latter are **algebraically closed** unlike the former. In other words, a polynomial equation with real coefficients $a_0 + a_1 x + \cdots + a_n x^n = 0$ need not have any real roots. On the other hand, by the **Fundamental Theorem of Algebra**, a polynomial of degree n $(a_n \neq 0)$ with complex coefficients $a_0, \cdots, a_n$, is guaranteed to have n complex roots.

• The complex numbers admit an **involution** called **complex conjugation**, that takes $z = x + iy$ to $\bar{z} = x - iy$. The conjugate is also denoted z^*. It is clear that when conjugation is applied twice, one gets the identity: $(z^*)^* = z$ (an involution is such an operation). Conjugation is reflection in the horizontal axis $(x,y) \to (x,-y)$. It reduces to the identity on the real numbers. The squared modulus is expressible as $|z|^2 = x^2 + y^2 = \bar{z}z$.

1.4 Stereographic projection: Riemann sphere and the extended complex plane

• Complex numbers may also be viewed as points on a punctured sphere. Consider the unit sphere S^2 consisting of the points (x_1, x_2, x_3) in 3d Euclidean space $\mathbb{R}^3$ with $x_1^2 + x_2^2 + x_3^2 = 1$. The equatorial x_1-x_2 plane will be viewed as the complex plane $\mathbb{C}$ where we will denote $x_1 = x$ and $x_2 = y$ and put $z = x + iy$. The points $(0,0,1)$

and $(0,0,-1)$ on the unit sphere are called its North and South poles. We will now define a map from $S^2 \setminus \{N\}$ (sphere with North pole removed) to the complex plane $\mathbb{C}$. Its extension to the north pole will allow us to define the point at $z = \infty$ and the extended complex plane.

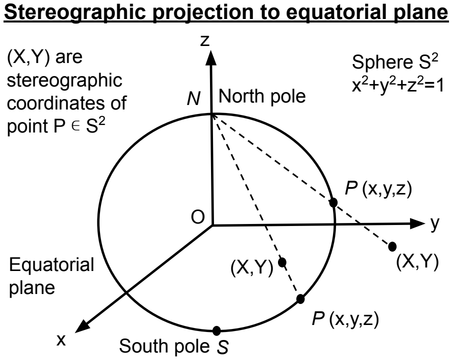

Stereographic projection to equatorial plane
(X,Y) are stereographic coordinates of point P ∈ S²
z
N
North pole
Sphere S²
x²+y²+z²=1
P (x,y,z)
O
y
Equatorial plane
(X,Y)
(X,Y)
P (x,y,z)
x
South pole S

Figure 1: Stereographic coordinates (x, y) of a point $P(x_1, x_2, x_3)$ on the sphere S^2 are given by the point of intersection with the equatorial plane of the line from the North pole through P. In the figure, (x, y, z) is to be read as (x_1, x_2, x_3) while (X, Y) is to be read as (x, y).

• Suppose P is a point on S^2 with Cartesian coordinates (x_1, x_2, x_3). Its image P' under the stereographic projection is the point with coordinates (x, y) on the equatorial plane through which the line passing through N and P passes. Evidently, points in the northern/southern hemisphere ($x_3 > 0$ or $x_3 < 0$) are mapped to points outside/inside the unit circle ($x^2 + y^2 = 1$). The equator ($x_1^2 + x_2^2 = 1$, $x_3 = 0$) is mapped to this unit circle. Explicitly, the map is given by

$$x = \frac{x_1}{1 - x_3} \quad \text{and} \quad y = \frac{x_2}{1 - x_3} \quad \text{or} \quad z = \frac{x_1 + i x_2}{1 - x_3}. \tag{3}$$

The map is clearly invertible, with inverse $\mathbb{C} \to S^2 \setminus \{N\}$ given by

$$x_1 = \frac{2x}{x^2 + y^2 + 1}, \quad x_2 = \frac{2y}{x^2 + y^2 + 1} \quad \text{and} \quad x_3 = \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}. \tag{4}$$

While the South pole is mapped to the origin $z = 0$, the North pole is in a limiting sense, mapped to the 'circle at infinity' in the complex plane. We will identify all the points on this 'circle at infinity' and denote them by $z = \infty$. The set $\mathbb{C} \cup \{\infty\}$ is called the one-point compactification of the complex plane and denoted $\mathbb{C}_\infty$ or $\hat{\mathbb{C}}$. Equations (3) and (4) imply that the North pole $N(0, 0, 1)$ is mapped to the point $z = \infty$ and vice versa if we take the limits $x_3 \to 1$ and $x^2 + y^2 \to \infty$.

• Under the stereographic projection, we may view $z = x + iy$ as a stereographic complex coordinate of a point on the unit sphere. We may develop the stereographic projection from any point on S^2. We need a minimum of two such stereographic projections (say, from N and S) to assign a complex coordinate to all points on S^2.

These lead to two compatible coordinate systems covering all of S^2. When viewed this way, we call S^2 the **Riemann sphere**. More on this when we study manifolds.

• The stereographic projection takes circles on the Riemann sphere with the North pole excluded to circles on the complex plane. Circles passing through N are mapped to straight lines.

• We may define a metric on the extended complex plane using natural notions of distance on the Riemann sphere such as the great circle distance or the chordal distance.

1.5 Convergence of a sequence and series

• Ideas of convergence were developed from practical considerations of obtaining solutions of equations (algebraic or differential) by a sequence of approximations. The current definition of convergence came out of thinking about accuracy of approximations, error estimates, tolerance for deviation and the number of iterations to be performed. The French mathematical physicist Augustin-Louis Cauchy (1789-1857) played an important role in developing ideas of convergence of sequences and series. He also did a lot of work in continuum mechanics and elasticity.

• A **sequence** is a map from the set of natural numbers $\{1, 2, 3, \ldots\}$ to any set Ω. The elements of the sequence s are written as $s_1, s_2, s_3, \ldots$ with $s_n \in \Omega$ for each $n = 1, 2, 3, \ldots$. Evidently, the natural numbers are used to index the elements in the sequence. For example, Ω could be the set of complex numbers or a set of functions of a complex variable. Thus, we can have sequences of complex numbers $z_1, z_2, \cdots$ or sequences of functions $f_1(z), f_2(z), \cdots$. For example, we have the sequence of bound state energies E_n of the 1d harmonic oscillator or those of the hydrogen atom. We also have the sequence of bound state wave functions $\psi_n(x)$ of the harmonic oscillator. Sequences also arise elsewhere in physics: the sequence of states x_n of a discrete time dynamical system where n plays the role of time.

• **Convergence of a sequence on a metric space.** If Ω is a metric space (like $\mathbb{C}$ with the Euclidean distance function), then we can introduce the idea of the convergence of a sequence. This plays an important role in complex analysis, not least because a complex analytic function is defined in terms of a convergent sequence of Taylor polynomials. When the sequence represents successive states of a dynamical system, convergence or the lack thereof tells us about the asymptotic behavior of the system.

• **Convergent sequence.** We will say that a sequence s_n on a metric space Ω converges to a point $x \in \Omega$ if, given any tolerance $\epsilon > 0$ there exists a positive integer N such that $d(s_n, x) < \epsilon$ for all $n > N$. In other words, by going far enough down the sequence, its elements can be made as close as we want, to x. We write $\lim_{n\to\infty} s_n = x$ or simply $s_n \to x$. As a consequence of the definition and the axiom $d(x, y) > 0$ if $x \neq y$, a sequence can converge to at most one point.

• **Bounded sequence.** A sequence s_n of complex numbers is bounded in magnitude if there is a positive real number B such that $|s_n| \leq B$ for all $n \geq 1$. A bounded sequence need not be convergent. For instance, the sequence whose elements $e^{in\pi/3}$ go round and round the unit circle is bounded but does not converge. The sequence $(-1)^n + \frac{1}{n}$ whose even elements tend to 1 and whose odd ones tend to -1 is not convergent.

- **Limit point or accumulation point.** In situations such as the example above, it is useful to introduce the idea of a limit point of a sequence. We say that z is a limit point of the sequence of complex numbers z_n if every disc centered at z contains an infinite number of elements of the sequence. In other words, a subsequence of elements converges to z. A convergent sequence has precisely one limit point. However, a bounded sequence can have several or even infinitely many limit points. The sequence $e^{in\pi/3}$ has six limit points while $(-1)^n + \frac{1}{n}$ has two limit points. The even subsequence of the latter converges of 1 while the odd subsequence converges to -1. Limit points are useful in discussing long time behavior (attractors) of dynamical systems such as the Logistic map: $x_{n+1} = rx_n(1 - x_n)$ (for $0 \leq x_n \leq 1$) which models the population of a species with growth rate $0 \leq r \leq 4$ at successive instants of time n.

- A bounded sequence need not have a maximal or minimal element. E.g., the sequence of bound state energies (in electron volts) of the hydrogen atom $-13.6/n^2$ has a minimal element but no maximal element.

- **Supremum and infimum.** A sequence (or set) of real numbers is bounded above by B if all elements are $\leq B$. The smallest such upper bound is called the least upper bound or supremum (abbreviated sup). Similarly, we have the notion of the greatest lower bound or infimum (abbreviated inf) of a sequence (or set) of real numbers that is bounded below. A bounded sequence of real numbers must have a supremum and infimum irrespective of whether it is convergent.

- **Limit supremum and limit infimum.** We can combine the notions of limit points and the supremum to define the limit supremum of a real sequence that is bounded above. The $\limsup$ is the supremum of its limit points. Similarly, the limit infimum ($\liminf$) of a real sequence that is bounded below is the infimum of its limit points. The $\limsup$ and $\liminf$ are only sensitive to the limit points, they do not depend on a finite number of larger or smaller elements. On the other hand, $\sup_n x_n$ and $\inf_n x_n$ depend on these outliers as well.

- **Cauchy sequence.** A sequence s_n on a metric space is a Cauchy sequence if the distance between elements can be made arbitrarily small by going far enough down the sequence. In other words, given $\epsilon > 0$, there is a positive integer N such that $d(s_m, s_n) < \epsilon$ for all $m, n > N$. Although the points may get closer to each other, a Cauchy sequence need not converge to a point in the metric space. For instance the Cauchy sequence $1/n$ does not converge to a point in the open unit interval $(0, 1)$. Evidently, the open interval is missing its 'limit points' 0 and 1. We may include these to get a complete metric space. On the other hand, $s_n = \log n$ is not a Cauchy sequence. This is because although the distance between successive elements goes to zero, the distance between, say, $\log n$ and $\log 5n$ does not go to zero as $n \to \infty$. It is noteworthy that every Cauchy sequence can be shown to be bounded.

- **Complete metric space.** A metric space Ω is called *complete* if every Cauchy sequence in it converges to a point of Ω. Under the Euclidean distance function, the complex numbers are a complete metric space while the rational numbers are not.

- **Convergent series.** Given a sequence $s_1, s_2, \ldots$, of real or complex numbers, we have the sequence of partial sums $S_n = \sum_{k=1}^n s_k$. If the sequence of partial sums $S_1, S_2, \ldots$ forms a convergent sequence, then we say that the sum of the sequence (or simply, series) is convergent and write

$$\lim_{n\to\infty} \sum_{k=1}^{n} s_k = \sum_{k=1}^{\infty} s_k. \tag{5}$$

If the partial sums do not converge, then we say that the series is divergent.

- For any fixed $|z| < 1$, the **geometric series** $\sum_{k=0}^{\infty} z^k$ is convergent, as we deduce from the limit of the sequence of partial sums: $\sum_{k=0}^{n} z^k = (1 - z^{n+1})/(1 - z)$. It converges to $1/(1 - z)$ since $z^{n+1} \to 0$ for any $|z| < 1$. It is a prototype for a convergent series.
- Sometimes, we may determine whether a series is convergent by comparing its terms with those of a geometric series, where the ratio of successive term is a constant.
- **Ratio test for (absolute) convergence.** A useful test for convergence is the ratio test. A series $\sum_k a_k$ is convergent if the absolute value of the ratio of successive terms has a limit that is less than one: $\lim_{n\to\infty} |a_{n+1}/a_n| < 1$. It diverges if this limit exceeds one. The test is inconclusive if the limit equals unity. According to the ratio test, the geometric series is convergent for $|z| < 1$. A refined version of the ratio test applies even when the above limits do not exist. It states that the series converges if $\limsup_n |a_{n+1}/a_n| < 1$ and diverges if $\liminf_n |a_{n+1}/a_n| > 1$. In fact, the ratio test is a test for absolute convergence, a concept that we now introduce.
- **Absolute convergence.** A series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent if the series of absolute values $\sum_{k=1}^{\infty} |a_k|$ converges. Absolute convergence is a stronger condition that convergence. An absolutely convergent series is automatically convergent. On the other hand, a series $\sum_k a_k$ may converge due to cancellations among terms, which may fail to happen in the series of absolute values. For instance, the alternating harmonic series $\sum_k (-1)^{k+1}/k$ converges to $\log 2$ although not absolutely. The harmonic series $\sum_{k=1}^{\infty}(1/k)$ is logarithmically divergent ($\sum_{k=1}^{n}(1/k) \sim \log n + \gamma$). The series $\sum_k (-1)^{k+1}/k^2$ is absolutely convergent as is the geometric series $\sum_{k=1}^{\infty} z^k$ for $|z| < 1$.
- A series is called **conditionally convergent** if it converges, though not absolutely. For example, the series $1 - 1 + \frac{1}{2} - \frac{1}{2} + \frac{1}{3} - \frac{1}{3} + \cdots$ converges to zero but the sum of the absolute values is divergent (it is twice the harmonic series). In an absolutely convergent series, we may rearrange the terms without affecting its sum. However, according to Riemann's theorem, this is not the case for conditionally convergent series. Terms in a conditionally convergent series may be rearranged (permuted) so that the sum converges to any desired value (including $\pm\infty$) or even not converge at all.
- **Comparison test for absolute convergence.** If $\sum_k d_k$ is absolutely convergent and $|c_k| \le |d_k|$ for all sufficiently large k, then the series $\sum_k c_k$ is absolutely convergent. We say that the series $\sum_k d_k$ eventually dominates the series $\sum_k c_k$.
- **Cauchy's n^{th} root test for absolute convergence.** A series $\sum_n a_n$ is absolutely convergent if $\limsup |a_n|^{1/n} < 1$ and it diverges if $\limsup |a_n|^{1/n} > 1$. Roughly, if $\limsup |a_n|^{1/n} < 1$, then for all but a finite number of terms in the sum, $|a_n| < R^n$

for some positive $R < 1$. Then the series converges by comparison with the geometric series $\sum_n R^n$.

1.6 Convergence in the space of bounded continuous functions

• We now wish to go from discussing the convergence of sequences of numbers to that of sequences of functions. For this we need a suitable metric space of functions, which is furnished by bounded continuous functions.

• **Continuous functions.** Let us begin by recalling that a real-valued function $f(x)$ of a real variable x is continuous at x_0 if given any $\epsilon > 0$ there exists a δ such that $|f(x) - f(x_0)| < \epsilon$ for all x with $|x - x_0| < \delta$. In other words, for a continuous function, we can make $f(x)$ arbitrarily close to its value at x_0 by choosing x sufficiently close to x_0. The Heaviside step function $\theta(x)$ which is equal to one for $x \geq 0$ and zero otherwise is continuous everywhere except at $x = 0$. It is the indicator or characteristic function of the positive real line. The characteristic function of a subset of $\mathbb{R}$ is the one that is equal to one on the subset and zero on the complement. The characteristic function of the rational numbers is nowhere continuous. Polynomials are continuous everywhere. On the other hand, the function $f(x)$ which vanishes at $x = 0$ and is equal to $\sin(1/x)$ elsewhere is discontinuous at $x = 0$. It cannot be made continuous by any other choice of $f(0)$.

• **Supremum norm.** The space of bounded complex-valued functions of a real variable in a finite interval $f : [a, b] \to \mathbb{C}$ is a metric space with distance function given by the supremum norm

$$d(f, g) = \sup_{x \in [a,b]} |f(x) - g(x)| \tag{6}$$

The sup norm measures the largest difference in heights between the graphs of the functions.

• There are other interesting distance functions such as the one based on the L^2 norm:

$$d_{L^2}(f, g) = \left[\int_a^b |f(x) - g(x)|^2 dx \right]^{1/2} . \tag{7}$$

• **Uniform convergence.** We may now define the notion of a uniformly convergent sequence of functions. A sequence of functions $f_n : [a, b] \to \mathbb{C}$ is uniformly convergent to the function f if given any $\epsilon > 0$ there exists a positive integer $N > 0$ such that $|f_n(x) - f(x)| < \epsilon$ for each $n > N$ and any $x \in [a, b]$. In other words $f_n(x)$ must come within ϵ of $f(x)$ for all values of x, it is in this sense that the convergence is uniform. If the value of N depended on x, then the convergence would be called nonuniform or pointwise, rather than uniform. When a sequence converges nonuniformly, the rate of convergence differs at distinct values of x.

• Uniform convergence is the same as convergence in the supremum norm since $|f_n(x) - f(x)| < \epsilon$ for all x if and only if $\sup |f_n(x) < f(x)| < \epsilon$.

• **Completeness and approximation by polynomials or Fourier series.** This space of bounded continuous functions is in fact a complete metric space under the sup norm. Thus, any Cauchy sequence of bounded continuous functions must converge to a continuous function. It turns out that polynomials (as well as Fourier series) form a dense subset of this space of bounded continuous functions on an interval. Thus, we may realize any continuous function as the limit of a sequence of polynomials or Fourier series. For instance, we may use Legendre polynomials for this purpose.

1.7 Power series

• A **power series** around the point $z_0 \in \mathbb{C}$ is one of the form $\sum_{k=0}^{\infty} a_k (z-z_0)^k$ for constant complex coefficients a_k.

• **Radius of convergence.** By the Cauchy $n^{\rm th}$ root test, a power series converges absolutely if $\limsup_n |a_n|^{1/n}|z-z_0| < 1$. Thus, a power series converges in a disk of radius

$$R = [\limsup_{n\to\infty} |a_n|^{1/n}]^{-1} \tag{8}$$

around z_0. This is the radius of convergence of the series. Henceforth, unless otherwise stated, we will consider power series around the origin and take $z_0 = 0$.

• The **geometric series** $\sum_{n=0}^{\infty} z^n$ is a simple example of a power series. It converges (absolutely) for $|z| < 1$ to $1/(1-z)$.

• We will soon see that $\sum_{n=1}^{\infty} nz^n$, $\sum_{n=1}^{\infty} n^2 z^n$ and in fact $\sum_{n=1}^{\infty} n^k z^n$ for $k = 1, 2, 3, \ldots$ are all absolutely convergent for $|z| < 1$.

• The **exponential series** $\sum_{n\geq 0} z^n/n!$ has an infinite radius of convergence. It defines the exponential function e^z. For any fixed z, the ratio of magnitudes of successive terms $|z^{n+1}/(n+1)!|/|z^n/n!| = |z|/(n+1)$ tends to zero as $n \to \infty$. So the exponential series converges for any z.

• **Divergent series.** On the other hand, the series $\sum_n n! z^n$ has zero radius of convergence. The absolute ratio of successive terms is $(n+1)|z|$, which tends to ∞ as $n \to \infty$ for any fixed $z \neq 0$. Thus, this series is divergent for any $z \neq 0$. Such series arise as perturbation series in quantum field theory. The $n!$ arises because there are that many Feynman diagrams.

• **Term-by-term differentiation** allows us to obtain a (possibly new) convergent power series from a familiar convergent one. Suppose $\sum_n a_n z^n$ has radius of convergence $R = [\limsup_{n\to\infty} |a_n|^{1/n}]^{-1}$. Then the differentiated series $\sum_n n a_n z^{n-1}$ has the radius of convergence

$$R' = [\limsup_{n\to\infty} |(n+1)a_{n+1}|^{1/n}]^{-1}. \tag{9}$$

Now $(n+1)^{1/n} \to n^{1/n} \to 1$ and so

$$\frac{1}{R'} = \limsup_{n\to\infty} |a_{n+1}|^{1/n} = \limsup_{n\to\infty} |a_{n+1}|^{\frac{1}{n+1}\frac{n+1}{n}} = \limsup_{n\to\infty} |a_{n+1}|^{\frac{1}{n+1}} = \frac{1}{R}. \tag{10}$$

So the term-by-term differentiated series has the same radius of convergence as the original series.

• **Binomial series.** For example, differentiating $1/(1-z)$ we get

$$\frac{1}{(1-z)^2} = \sum_{k=1}^{\infty} kz^{k-1} = \sum_{k\geq 0}(k+1)z^k. \tag{11}$$

Splitting into two series, $\sum kz^k + \sum z^k$, we see that $\sum_k kz^k = (1-z)^{-2} - (1-z)$. As anticipated, this series is absolutely convergent with radius of convergence unity.

Differentiating again,

$$\frac{1}{(1-z)^3} = \frac{1}{2}\sum_{k\geq 2} k(k-1)z^{k-2} = \sum_{k\geq 0}\frac{(k+2)!}{2}\frac{z^k}{k!} = \sum_{k\geq 0} 3\cdot 4\cdots(k+2)\frac{z^k}{k!}. \tag{12}$$

More generally,

$$\frac{1}{(1-z)^n} = \sum_{r=0}^{\infty} n(n+1)\cdots(n+r-1)\frac{z^r}{r!}, \tag{13}$$

which we recognize as a generalization of the binomial theorem.

• **Logarithm.** The logarithm is defined for $|z| < 1$ via the absolutely convergent series

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} + \cdots \tag{14}$$

Differentiating term by term,

$$\frac{d}{dz}\log(1+z) = 1 - z + z^2 + \cdots = \frac{1}{1+z} \tag{15}$$

Defining $w = 1+z$ we arrive at a result familiar from the calculus of one real variable: $d\log w/dw = 1/w$ at least for $|w-1| < 1$.

• By composing the power series around the point $z=0$, we can show that $e^{\log(1+z)} = 1+z$ at least for $|z| < 1$. In fact, the first few terms are

$$\begin{aligned} e^{\log(1+z)} &= 1 + \log(1+z) + \frac{\log(1+z)^2}{2!} + \frac{\log(1+z)^3}{3!} + \cdots \\ &= 1 + z + \left(-\frac{1}{2} + \frac{1}{2}\right)z^2 + \left(\frac{1}{3} - \frac{2}{2!} + \frac{1}{3!}\right)z^3 + \cdots \\ &= 1 + z + \mathcal{O}(z^4). \end{aligned} \tag{16}$$

We have exploited absolute convergence to rearrange the terms. Proceeding this way, one must show that the coefficients of all higher powers of z vanish.

1.8 Laurent series

Laurent series with finitely many negative powers in a punctured disk. A Laurent series around $z = 0$ with finitely many negative exponents is a power series with nonzero radius of convergence (R) plus a polynomial (of degree N) in $1/z$. Such a Laurent series may be written as

$$f(z) = \sum_{n=-N}^{\infty} a_n z^n. \tag{17}$$

This series converges in the punctured disk $0 < |z| < R$ (we remove the origin from an open disk of radius R). The nonnegative integer N is called the degree of the pole of f at $z = 0$. The coefficient of $1/z$, i.e., a_{-1} is called the residue of the pole at $z = 0$. The significance of the residue will become clear when we discuss Cauchy's formula for contour integrals. If $N = 1$, i.e., $f(z) = a_{-1}/z + a_0 + a_1 z + \cdots$, then we say that f has a simple pole at $z = 0$, if $N = 2$, it has a double pole, etc.

• More generally, a Laurent series around $z = z_0$ may be written as $\sum_{n \geq -N} a_n(z - z_0)^n$. The sum of the terms with strictly negative powers of $(z - z_0)$ is called the singular part:

$$\text{singular part of } f \text{ at } z_0 = \frac{a_{-N}}{(z - z_0)^N} + \cdots + \frac{a_{-1}}{(z - z_0)}. \tag{18}$$

The singular part is sometimes called the 'principal part' or the 'pole part'.

• The convergent Taylor series in a Laurent series is called the regular part:

$$\text{regular part of } f \text{ at } z_0 = \sum_{n \geq 0} a_n (z - z_0)^n. \tag{19}$$

Doubly infinite Laurent series in an annulus. More generally, we may consider Laurent series with a pole of possibly infinite order, i.e., doubly infinite series of the form

$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n. \tag{20}$$

The strictly negative powers ($n < 0$) comprise the singular part while the regular part is given by the power series with nonnegative exponents ($n \geq 0$). The regular part is assumed to have a positive radius of convergence $R > 0$. The singular part can be viewed as a power series in $1/z$ and is assumed to converge for $|1/z| < 1/r$. Hence the singular part converges for $|z| > r$. Furthermore, if $r < R$, the doubly infinite Laurent series converges for z in the annulus $r < |z| < R$. It could of course happen that $r \to 0$ or $R \to \infty$, in which case the domain of convergence becomes a punctured disk or the exterior of a disk.

• For example, the Laurent series for the function $e^{1/z}$:

$$\sum_{0}^{\infty} \frac{z^{-n}}{n!} = 1 + \frac{1}{z} + \frac{1}{2z^2} + \cdots \tag{21}$$

converges in the punctured complex plane $|z| > 0$ ($R \to \infty$ and $r \to 0$). It has a pole of infinite order at $z = 0$.

1.9 Analytic functions defined by convergent power series

• **Analytic function.** A function $f(z)$ of a complex variable is called analytic at the point z_0 if it is equal to an absolutely convergent power series $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ within the circle of convergence of the series. When this is the case, the coefficients are uniquely determined and given by the so-called **Taylor coefficients**

$$a_n = \frac{1}{n!} \frac{d^n f(z)}{dz^n}\bigg|_{z=z_0}. \tag{22}$$

The series is called the **Taylor series** of f around the point z_0.

• In a similar fashion, an **antianalytic function** is one that may be represented by an absolutely convergent power series in $\bar{z} - \bar{z}_0$ for any point $\bar{z}_0$ in its domain of antianalyticity. For example, suppose $f(z)$ is analytic in the open connected domain D. Then, around any point $z_0 \in D$, it may be represented by an absolutely convergent power series $f(z) = \sum_{n\geq 0} a_n (z - z_0)^n$. We may use this to construct an antianalytic function $g(\bar{z})$ in the domain $\bar{D}$, which is the reflection of D in the real axis (complex conjugate domain). Indeed, consider the complex conjugate function $g(\bar{z}) = (f(z))^*$. It admits an absolutely convergent series representation around any point $\bar{z}_0 \in \bar{D}$, given by $g(\bar{z}) = \sum_{n\geq 0} \bar{a}_n (\bar{z} - \bar{z}_0)$. The condition for absolute convergence of this conjugate series is met since $|\bar{a}_n(\bar{z} - \bar{z}_0)| = |a_n(z - z_0)|$.

• **Polynomials.** The simplest examples of analytic functions are **polynomials** in z: $f(z) = \sum_{k=0}^{n} a_k z^k$. The degree of the polynomial is n. Such a polynomial can be rewritten as a polynomial (of the same degree) in $z - z_0$ for any z_0. Thus, a polynomial defines a function that is analytic everywhere in $\mathbb{C}$.

• **Entire function.** A function that is analytic at every $z_0 \in \mathbb{C}$ is called an entire function. Our previous remark shows that polynomials are entire. What is more, polynomials have an infinite radius of convergence. More generally, a power series with an infinite radius of convergence defines an entire function. The exponential series $e^z = \sum_{n=0}^{\infty} z^n/n!$, which has an infinite radius of convergence, defines an entire function.

• **Sum, product and quotient of analytic functions.** The sum and product of two functions analytic at z_0 is again analytic. The radius of convergence of a product of two convergent series is at least as big as the smaller of the two radii of convergence. On the other hand, the quotient of two analytic functions f/g is analytic at z_0 provided $g(z_0) \neq 0$. We may then write $g(z) = g(z_0) + h(z)$ and

$$g(z)^{-1} = g(z_0)^{-1}(1 + h/g(z_0))^{-1} = g(z_0)^{-1}\left[1 - \frac{h}{g(z_0)} + \frac{h^2}{g(z_0)^2} + \cdots\right]. \tag{23}$$

Here $h(z) = g(z) - g(z_0)$ vanishes at z_0, is analytic and admits a convergent power series expansion about z_0. We multiply out these terms and then multiply by the series for $f(z)$ to obtain a convergent power series for f/g in powers of $z - z_0$.

• **Derivative of an analytic function is analytic.** We have already noted that term-by-term differentiation of a convergent series gives us another convergent series with the same radius of convergence. Thus, the derivative of an analytic function is automatically analytic. In particular, a complex analytic function is infinitely differentiable. It is worth bearing in mind that complex analyticity is a much stronger condition than differentiability. A once differentiable function need not be twice differentiable. For example, the function of one real variable given by $f(x) = x^2 \,\mathrm{sgn}\,(x)$ where $\mathrm{sgn}\,(x)$ is the signum function (equal to the sign of its argument) that vanishes at $x = 0$, has a first derivative $f'(x) = 2x \,\mathrm{sgn}\,(x)$ that is continuous but not differentiable at $x = 0$. In other words, the second derivative $f''(x) = 2 \,\mathrm{sgn}\,(x)$ is discontinuous at $x = 0$.

• We will use the word domain for a nonempty open connected subset Ω of the complex plane. It is connected if it comes in one piece: given any two points $p, q \in \Omega$ there must be a continuous curve γ (thought of as parameterized by time t) that joins them, i.e., $\gamma : [0, 1] \to \Omega$ with $\gamma(0) = p$ and $\gamma(1) = q$.

• **Regularity classes: C^k, C^∞, C^ω.** It is useful to have notation for various regularity classes of functions in some domain Ω. A C^0 function is one that is continuous. A C^1 function is one whose first partial derivatives exist (for a function of two variables, these are $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$) and are continuous. A C^2 function is one whose second partial derivatives exist and are continuous. A C^k function is k times continuously differentiable. A C^∞ function is one for which all derivatives exist and are continuous, such a function is called smooth. An analytic function (one that admits a convergent power series representation in $z = x + iy$) is said to be of class C^ω. If we denote the space of C^k functions in Ω by $C^k(\Omega)$, then we have the strict inclusions

$$C^\omega(\Omega) \subset C^\infty(\Omega) \subset \cdots \subset C^2(\Omega) \subset C^1(\Omega) \subset C^0(\Omega). \qquad (24)$$

A smooth (C^∞) function need not be analytic (C^ω). For example, $f(x, y) = x - iy$ has continuous partial derivatives of all orders and is smooth. However, it cannot be expressed as a power series in $x + iy$. Another example comes from the calculus of one real variable. The function defined by $f(x) = e^{-1/x^2}$ for $x \neq 0$ and $f(0) = 0$, is a smooth function. Plot its graph! It has continuous derivatives of all orders. Moreover, we find that all its derivatives vanish at $x = 0$. Thus, its Taylor coefficients are all zero and the corresponding Taylor series $T(x) = 0$ does not agree with the function $f(x)$. Thus, f does not admit a Taylor series that converges to the function. We say that f is not real analytic.

1.10 Zeros and isolated singularities of an analytic function

• **Zeros of an analytic function.** Suppose f is analytic in a domain Ω. We will say that f has a zero at $z_0 \in \Omega$ if $f(z_0) = 0$. A more refined notion is a zero of order n. We will say that f has a zero of order n if its first $n - 1$ derivatives vanish at z_0 while its n^{th} derivative is nonvanishing:

$$f(z_0) = f'(z_0) = \cdots = f^{(n-1)}(z_0) = 0 \quad \text{but} \quad f^{(n)}(z_0) \neq 0. \qquad (25)$$

Note that if f has a zero of order n, then it cannot be identically zero. Since f is assumed analytic, this means its first n Taylor coefficients around z_0 must vanish and its convergent power series representation must be of the form

$$f(z) = a_n(z-z_0)^n + a_{n+1}(z-z_0)^{n+1} + \cdots \tag{26}$$

with $a_n \neq 0$. We may thus factorize f as

$$f(z) = (z-z_0)^n g(z) \quad \text{where} \quad g(z) = \sum_{k=0}^{\infty} a_{n+k}(z-z_0)^k \tag{27}$$

is also analytic at z_0. What is more, since $a_n \neq 0$, $g(z_0) \neq 0$. By continuity, g must be nonvanishing in an open neighborhood of z_0. Since $(z-z_0)^n$ is also nonvanishing away from z_0, we conclude that f must be nonvanishing in a punctured neighborhood of z_0. Thus, we have shown that the zeros of an analytic function must be isolated in the domain of analyticity Ω if f is not identically zero in Ω. In particular, a nonconstant analytic function cannot have an accumulation point of zeros within the domain of analyticity, such an accumulation can only occur at a boundary point. Such a point is an example of a singular point of an analytic function.

• A point z_0 where a function is analytic is called a regular point. A point where it fails to be analytic is called a singular point.

• When the power series representing an analytic function ceases to be convergent, the function could have a singularity. Suppose f admits a convergent power series expansion around z_1 with radius of convergence R. Then there must be a point z_0 on the circle $|z-z_1| = R$ at which the power series diverges. Such a point z_0 cannot be a regular point, it is a singular point of f. Thus, the radius of convergence is the distance to the nearest singularity. For example, the geometric series $1 + z + z^2 + \cdots$ diverges when we put $z = 1$, which is a singularity of the corresponding analytic function $1/(1-z)$. As this example indicates, not all points on the boundary of the disk of convergence may be singular points.

• Analytic functions can have singularities of various sorts. We will consider isolated singularities for now. An isolated singularity z_0 is one where the function is analytic in a punctured disc $0 < |z-z_0| < R$ of nonzero radius $R > 0$ around the point. Isolated singularities can be removable, poles of finite order or essential singularities. We discuss each case.

• **Removable singularity.** Consider the function $f(z) = \frac{\sin z}{z}$. It is defined everywhere except at $z = 0$. However, the function has the limiting value unity as $z \to 0$. We say that f has a removable singularity at $z = 0$, the singularity may be removed by defining $f(0) = 1$. Once this is done, f defines an analytic function at $z = 0$. In fact, f is an entire function and admits a Taylor expansion with infinite radius of convergence

$$f(z) = \frac{1}{z}\left(z - \frac{1}{3!}z^3 + \frac{1}{5!}z^5 - \cdots\right) = 1 - \frac{1}{3!}z^2 + \frac{1}{5!}z^4 - \cdots . \tag{28}$$

• Generally speaking, we will say that f has a removable singularity at $z = z_0$ if by suitably defining the value of f at z_0, we can make f analytic at z_0.

• **Simple pole.** A function analytic in a punctured disk around $z = z_0$ is said to have a simple pole at $z = z_0$ if it may be expressed as a Laurent series

$$f(z) = \frac{a_{-1}}{(z - z_0)} + \sum_{n=0}^{\infty} a_n (z - z_0)^n \tag{29}$$

in a punctured disk ($0 < |z - z_0| < R$) of some radius $R > 0$ around z_0. The residue at the pole is a_{-1}.

• Notice that the function cannot be redefined at z_0 to make it analytic at z_0. However, by multiplying f by $(z - z_0)$, we get a function $g = (z - z_0)f$ that is analytic at z_0. We will say that a simple pole is neither a removable singularity nor an essential singularity.

• For example, consider $f = 1/(z(z-1))$. On the face of it, f has simple poles at $z = 0$ and $z = 1$. In fact, we may write

$$f = \frac{1}{z-1} - \frac{1}{z}. \tag{30}$$

The first term is the regular part around $z = 0$ while the second is the singular part around $z = 0$. The two terms reverse roles around $z = 1$. We may also read off the residues:

$$\text{Res}_{z=0} f(z) = -1 \quad \text{and} \quad \text{Res}_{z=1} f(z) = 1. \tag{31}$$

These statements are confirmed by the Laurent series around $z = 0$ and $z = 1$:

$$\begin{aligned} f(z) &= -\frac{1}{z} - (z + z^2 + z^3 + \cdots) \quad \text{and} \\ f(z) &= \frac{1}{1-z} - ((z-1) - (z-1)^2 + (z-1)^3 + \cdots). \end{aligned} \tag{32}$$

• Formulae for the residue at a simple pole. (i) If f has a simple pole at $z = z_0$, then the residue is expressible as the following limit:

$$\text{Res}_{z=z_0} f(z) = a_{-1} = \lim_{z \to z_0} [(z - z_0) f(z)]. \tag{33}$$

If the pole is not simple, then this limit would diverge. (ii) Suppose $f(z) = g(z)/h(z)$ is a quotient of functions that are analytic at z_0. For example, $g(z)$ and $h(z)$ could be polynomials. If h has a simple zero at z_0 and $g(z_0) \neq 0$, then near z_0, $h(z) \approx h'(z_0)(z - z_0)$. It follows that f has a simple pole at z_0 with residue given by

$$\text{Res}_{z=z_0} \frac{g(z)}{h(z)} = \frac{g(z_0)}{h'(z_0)}. \tag{34}$$

• **Multiple pole.** A multiple pole is defined in a similar way. Consider a function $f(z)$ analytic in a punctured disk around z_0. It has a pole of order $N = 1, 2, 3, \ldots$ if the function may be represented as a Laurent series

$$f(z) = \frac{a_{-N}}{(z - z_0)^N} + \cdots + \frac{a_{-1}}{(z - z_0)} + \sum_{n \geq 0} a_n (z - z_0)^n. \tag{35}$$

in a punctured disk $0 < |z - z_0| < R$ of some positive radius $R > 0$. Even if f has a pole of order $N > 1$, the residue of f at z_0 is defined as a_{-1}. A formula for this residue that generalizes the one for a simple pole (33) is given by

$$\text{Res}_{z=z_0} f(z) = a_{-1} = \lim_{z \to z_0} \frac{1}{(N-1)!} \frac{d^{N-1}}{dz^{N-1}} [(z - z_0)^N f(z)]. \tag{36}$$

The multiplication by $(z - z_0)^N$ and repeated differentiation serve to get rid of contributions to the limit from poles of order greater than one in the Laurent series.

• If f has a pole of order N at $z = z_0$, then $(z - z_0)^N f$ is regular at z_0. As with a simple pole, we say that a multiple pole is neither a removable nor essential singularity.

• **Isolated essential singularity.** An analytic function has an isolated essential singularity at z_0 if it has a pole of infinite order at z_0. In more detail, suppose f is analytic in a punctured disk around z_0 and admits the Laurent series representation

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n \tag{37}$$

with a_n nonzero for infinitely many $n < 0$. Then we say that f has an isolated essential singularity at z_0. If f has an essential singularity at z_0, then we cannot make it analytic at z_0 upon multiplication by $(z - z_0)^N$ for any positive integer N. This is the sense in which the singularity is essential.

• An isolated essential singularity can arise at a point of accumulation of values (say zeros) of an analytic function.

• The function $e^{1/z}$ has an essential singularity at $z = 0$. Its Laurent expansion

$$e^{1/z} = \sum_{n=0}^{\infty} \frac{z^{-n}}{n!} = 1 + \frac{1}{z} + \frac{1}{2z^2} + \cdots \tag{38}$$

exhibits a pole of 'infinite order' at $z = 0$, with residue 1 and regular part equal to the constant function 1. The behavior of an analytic function in the neighborhood of an essential singularity is quite complicated. The function $e^{1/z}$ may be used to illustrate some of these peculiar features. The function does not have a well defined limit as $z \to 0$: the behavior depends on how one approaches the essential singularity. For instance when approached along the real axis, $e^{1/x} \to \infty$ as $x \to 0^+$ while $e^{1/x} \to 0$ as $x \to 0^-$. On the other hand, when approached along the imaginary axis, $e^{1/iy}$ oscillates rapidly without approaching a limit as $y \to 0$. When approached from the right/left half planes with the real axis excluded, f oscillates increasingly rapidly with growing/decaying amplitude as $z \to 0$.

• This unbounded rapid oscillation is typical of the behavior of an analytic function in the neighborhood of an essential singularity. According to the theorem of **Sokhotski, Casorati and Weierstrass**, an analytic function must oscillate so rapidly and unboundedly in any punctured neighborhood of an essential singularity that it comes arbitrarily close to every complex value.

• A stronger result on the behavior of an analytic function in the vicinity of an essential singularity is **Picard's great theorem**. It tells us that in every punctured neighborhood of an essential singularity, an analytic function assumes every complex value (with one possible exception) infinitely often. For instance, the function $e^{1/z}$ does not vanish anywhere on the punctured complex plane[2]. However, Picard's theorem asserts that it takes every other complex value infinitely often as one approaches the essential singularity at $z = 0$.

• **Singularity at infinity.** An analytic function can have an isolated singularity at the point $z = \infty$ of the extended complex plane. To analyze the behavior around $z = \infty$ we change variables to $w = 1/z$ and consider the function $g(w) = f(1/w)$ in a punctured disc around $w = 0$. We say that f is regular or has a removable singularity, a pole of finite order or an essential singularity at $z = \infty$ if $g(w)$ displays such a behavior at $w = 0$. For example, the identity function $f(z) = z$ corresponds to $g(w) = 1/w$ which has a simple pole at $w = 0$. So we say that $f(z) = z$ has a simple pole at $z = \infty$. More generally, a polynomial of degree N, $f(z) = a_N z^N + a_{N-1} z^{N-1} + \cdots + a_0$ (with $a_N \neq 0$) has a pole of order N at $z = \infty$.

• **Meromorphic functions on $\mathbb{C}$.** An analytic function whose only singularities on the finite complex plane are isolated removable singularities or isolated poles of finite order is called a meromorphic function. Entire functions such as polynomials and the exponential function are automatically meromorphic. The simplest meromorphic functions that are not entire are ratios of polynomials; they have a finite number of poles in the complex plane. The trigonometric functions $\sec z, \operatorname{cosec} z, \tan z, \cot z, \operatorname{sech} z$ as well as their hyperbolic counterparts are meromorphic; they have infinitely many poles which accumulate at $z = \infty$. A meromorphic function cannot have a point of accumulation of poles in the finite complex plane. Such a point would be a nonisolated singularity. For example, $\sin(1/z)$ has zeros at $z = 1/n\pi$ for integer n, which accumulate at $z = 0$. Thus, its reciprocal $\operatorname{cosec}(1/z)$ has an accumulation of simple poles at $z = 0$ and is therefore not a meromorphic function. In general, a meromorphic function f in a domain Ω is expressible as a ratio g/h of analytic functions in Ω with poles of f occurring precisely at the zeros of h.

• A more restricted notion is that of a meromorphic function on the extended complex plane or Riemann sphere. These are analytic functions that have no singularities other than removable singularities and isolated poles of finite order in $\hat{\mathbb{C}}$. Polynomials and rational functions are meromorphic on $\hat{\mathbb{C}}$ but e^z and related trigonometric and hyperbolic functions that have an essential singularity at $z = \infty$ are not.

• **Nonisolated singularities.** Analytic functions can have nonisolated singularities, some of which we will discuss later. Possibilities include (a) accumulation points of isolated singularities (e.g., poles) and (b) as branch cuts. The function $\operatorname{cosec}(1/z)$ has an accumulation point of simple poles at $z = 0$. The square root $\sqrt{z}$ and logarithm $\log z$ are examples of analytic functions that display nonisolated singularities (branch cuts).

[2] This is because $e^w = e^{\Re w} e^{i \Im w}$ and $e^{\Re w} > 0$ for all w.

1.11 Holomorphic functions: Cauchy-Riemann equations

• **Cauchy-Riemann equations.** A complex-valued function f of a complex variable $z = x + iy$ may be written in terms of its real and imaginary parts: $f = u + iv$. We will say that f is holomorphic if u and v are continuously differentiable functions of x and y and satisfy the Cauchy-Riemann (CR)

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{39}$$

The CR equations are a pair of coupled linear partial differential equations.

• **Examples.** The identity function $f(z) = z$ is holomorphic: $u = x$ and $v = y$ implies $u_x = v_y = 1$ while $u_y = -v_x = 0$. On the other hand, the complex conjugate function is not holomorphic: $f(z) = \bar{z}$ implies $u = x$ and $v = -y$ so $u_x \neq v_y$.

• We will now interpret the CR equations in terms of (anti)holomorphic derivatives and in terms of differentiation in the complex plane.

• Let us define the **holomorphic and antiholomorphic derivatives**

$$\partial f = \frac{1}{2}(\partial_x f - i\partial_y f) \quad \text{and} \quad \bar{\partial} f = \frac{1}{2}(\partial_x f + i\partial_y f). \tag{40}$$

These definitions are motivated by the chain rule if we view $z = x + iy$ and $\bar{z} = x - iy$ as functions of x and y and conversely $x = (z + \bar{z})/2$ and $y = (z - \bar{z})/2i$ as functions of z and $\bar{z}$ so that

$$\begin{aligned} \partial &= \partial_z = \frac{\partial}{\partial z} = \frac{\partial x}{\partial z}\partial_x + \frac{\partial y}{\partial z}\partial_y = \frac{1}{2}(\partial_x - i\partial_y) \quad \text{and} \\ \bar{\partial} &= \partial_{\bar{z}} = \frac{\partial}{\partial \bar{z}} = \frac{\partial x}{\partial \bar{z}}\partial_x + \frac{\partial y}{\partial \bar{z}}\partial_y = \frac{1}{2}(\partial_x + i\partial_y) \end{aligned} \tag{41}$$

With these definitions, we see that

$$\bar{\partial} f = \frac{1}{2}(u_x - v_y + i(u_y + v_x)) = 0 \tag{42}$$

is equivalent to the Cauchy-Riemann equations. Thus, the CR equations say that f can depend on x and y through the combination $z = x + iy$ but not through $\bar{z} = x - iy$. A holomorphic function is a function of x and y that lies in the kernel or nullspace of $\bar{\partial}$.

• **Examples.** Thus, any polynomial in z is automatically holomorphic. However, the absolute square function $f(z) = \bar{z}z = x^2 + y^2$ is not holomorphic since it depends nontrivially on $\bar{z}$.

• Similarly, we will define an **antiholomorphic function** as one for which $\partial f = 0$, i.e., one whose real and imaginary parts satisfy the 'anti-CR' equations $u_x = -v_y$ and $u_y = v_x$. Alternatively, an antiholomorphic function is one whose complex conjugate is a holomorphic function. The only functions that are both holomorphic and antiholomorphic are constants. In other words, the intersection of the nullspaces (kernels) of ∂ and $\bar{\partial}$ consists of constant functions.

• It is noteworthy that most functions of x and y are **neither holomorphic nor antiholomorphic**. For example, $f(x,y) = x^2 + y^2 = z\bar{z}$ is not annihilated by ∂ or $\bar{\partial}$. Evidently, the condition of holomorphy is very stringent. Analytic function theory owes its strength *and* limitations to this stringent condition.

• **Complex derivative.** Another viewpoint is based on the complex derivative. A complex valued function of x, y can be viewed as a vector field on the plane: the real and imaginary parts are its x and y components $f = u(x,y) + iv(x,y) \rightsquigarrow u(x,y)\hat{x} + v(x,y)\hat{y}$. A continuously differentiable vector field is one one for which the partial derivatives u_x, u_y, v_x, v_y exist and are continuous functions of x and y. In particular, there need be no relation between u_x and v_y. The complex derivative is a very special type of derivative, quite different from the partial derivatives. We say that $f(x,y)$ has a complex derivative if the following limit exists[3]:

$$f'(z) = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h}. \qquad (43)$$

Here, we require that the limit of the difference quotient must be the same regardless of how h approaches zero in the complex plane. If we choose real values for h, then in terms of u and v, we get

$$f'(z) = \lim_{h\to 0} \frac{u(x+h,y) + iv(x+h,y) - u(x,y) - iv(x,y)}{h} = u_x + iv_x. \qquad (44)$$

On the other hand, if we choose imaginary values $h = i\epsilon$, then we get

$$f'(z) = \lim_{\epsilon\to 0} \frac{u(x,y+\epsilon) + iv(x,y+\epsilon) - u(x,y) - iv(x,y)}{i\epsilon} = -iu_y + v_y. \qquad (45)$$

Equating these two expressions for f' we arrive at the CR equations. It may be shown that other ways in which h may approach zero also lead to the same CR equations. Thus, the CR equations are equivalent to the existence of the complex derivative.

• **Holomorphic function and differential of type (1,0).** We may view z and $\bar{z}$ as an alternative to the coordinates x and y on $\mathbb{C}$. With a slight abuse of notation, if we view $f(z,\bar{z}) = f(x,y)$ as a map from $\mathbb{C} \to \mathbb{C}$ then the differential of f is

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z}. \qquad (46)$$

If f is holomorphic, the second term is absent and $df = \partial f dz$. In this case we say that df is a differential form of type $(1,0)$. Similarly, the differential $\bar{\partial} g(\bar{z}) d\bar{z}$ of an antiholomorphic function $g(\bar{z})$ is a differential form of type $(0,1)$. A holomorphic function is one whose differential (or exterior derivative) is a 1-form of type $(1,0)$.

• **Formula for derivative of a holomorphic function.** Using the chain rule, the derivative of $f(z)$ may be written as

$$\frac{df(z)}{dz} = \frac{\partial x}{\partial z}\frac{\partial f}{\partial x} + \frac{\partial y}{\partial z}\frac{\partial f}{\partial y} = \frac{1}{2}(f_x - if_y) = \frac{1}{2}(u_x + v_y) + \frac{i}{2}(v_x - u_y). \qquad (47)$$

[3] Till now, the symbol $f'(z)$ meant the term-by-term differentiation of a convergent power series. Now we are, in effect, giving an alternate way of finding this derivative.

If f is holomorphic, then the derivative may be written in several equivalent ways

$$f'(z) = u_x + iv_x = -iu_y + v_y = u_x - iu_y = iv_x + v_y. \qquad (48)$$

• **Holomorphic and complex analytic functions.** The condition of holomorphy $\bar{\partial} f = 0$ implies that f is a function of z and not $\bar{z}$. A convergent power series in z is automatically a holomorphic function. Conversely, it can be shown that a holomorphic function may be expanded in a convergent Taylor series around any point in its domain of holomorphy. We will do this using the Cauchy contour integral formula to solve the Cauchy-Riemann equations in §1.12. Thus, the concepts of complex analyticity and holomorphy will be seen to coincide. For this reason, we will often use the two terms interchangeably.

• **Holomorphy and harmonic functions.** A consequence of the Cauchy-Riemann equations is that the real and imaginary parts $u(x, y)$ and $v(x, y)$ of a complex analytic function $f = u + iv$ are harmonic functions. In fact, assuming u and v are twice continuously differentiable, the CR equations $u_x = v_y$ and $u_y = -v_x$ imply that $u_{xx} = v_{xy} = v_{yx} = -u_{yy}$ so that $u_{xx} + u_{yy} = 0$. Similarly, $v_{xx} + v_{yy} = 0$. The 2nd order linear partial differential operator $\Delta = \partial_x^2 + \partial_y^2$ is called the Laplace operator. A function annihilated by the Laplacian is called harmonic.

• **Example.** For example, consider the function $f(z) = e^z = e^{x+iy} = e^x \cos y + ie^x \sin y$. The real and imaginary parts are $u = \Re f = e^x \cos y$ and $v = \Im f = e^x \sin y$. We notice that $u_{xx} = u$ while $u_{yy} = -u$, so that u is harmonic. Similarly, $v_{xx} = v$ and $v_{yy} = -v$ so that $\Delta v = 0$.

• **Conjugate harmonic functions.** We say that the real and imaginary parts of a holomorphic function are a pair of conjugate harmonic functions.

• **Derivative of a holomorphic function is again holomorphic.** The Cauchy-Riemann equations ensure that the derivative $f'(z) = \partial f$ of a holomorphic function is itself holomorphic, provided u and v are C^2 functions of x and y. In fact, $\bar{\partial} f'(z) = \bar{\partial}\partial f = \partial\bar{\partial} f = 0$. We have used the fact that for C^2 functions, partial derivatives commute. Thus, differentiation preserves holomorphy.

1.12 Cauchy's integral theorem

• It is remarkable that the Cauchy-Riemann partial differential equations (subject to suitable boundary conditions) may be explicitly solved via contour integration. This is called the Cauchy integral formula. Let us see how this is done and how we can use it.

• A smooth contour or curve on the complex plane is a map $\gamma : [a, b] \to \mathbb{C}$ with continuous derivatives of all orders. Here $\gamma(t) = x(t) + iy(t)$ can be thought of as the trajectory of a particle on the complex plane, parameterized by time. We will often also consider piecewise smooth curves where the interval is broken up into a union of nonoverlapping subintervals where the map is smooth. The image of such a curve in $\mathbb{C}$ can have isolated 'sharp corners'. A contour is simple if it does not intersect itself: $t \neq t'$ must imply that $\gamma(t) \neq \gamma(t')$. A contour is closed if $\gamma(a) = \gamma(b)$.

• Given a complex-valued function $f((x,y))$, its integral along the smooth curve $\gamma_1 : [a,b] \to \mathbb{C}$ is denote $\int_{\gamma_1} f dz$ and defined as

$$\int_{\gamma_1} f dz \equiv \int_a^b f(\gamma_1(t)) \, \frac{d\gamma_1(t)}{dt} \, dt \qquad (49)$$

For a piecewise smooth curve, we add up the contributions from each smooth segment.
• We would like to exploit Stokes' theorem to know how this contour integral behaves under a deformation of the contour holding the end points $\gamma_1(a)$ and $\gamma_1(b)$ fixed, when f is a holomorphic function. To do so, suppose $\gamma_2 : [b,c] \to \mathbb{C}$ is another piecewise smooth contour with the same end points ($\gamma_2(b) = \gamma_1(a)$ and $\gamma_2(c) = \gamma_1(b)$), and let $\gamma = \gamma_1 \cup \gamma_2^{-1}$ be the closed contour where γ_1 is traversed first followed by γ_2 traversed in the opposite direction. For simplicity, we will assume that γ is a simple closed contour that divides the complex plane into an interior Ω and an exterior (this is guaranteed by the Jordan curve theorem). In particular $\partial\Omega = \gamma$. Thus, we consider the contour integral

$$\oint_\gamma f dz. \qquad (50)$$

We will view this contour integral as the line integral of a (complex-valued) vector field $\boldsymbol{A} = A_x \hat{x} + A_y \hat{y}$ (or corresponding one-form $A = A_x dx + A_y dy$) along the curve $\boldsymbol{\gamma}(t) = (x(t), y(t))$:

$$\int_\gamma A \equiv \int_a^c (A_x \dot{x} + A_y \dot{y}) dt = \int_a^c \boldsymbol{A} \cdot \frac{d\boldsymbol{\gamma}}{dt} \, dt. \qquad (51)$$

Comparing with the complex contour integral (with $\gamma(t) = x(t) + iy(t)$)

$$\int_\gamma f dz = \int_a^c (f\dot{x} + if\dot{y}) dt, \qquad (52)$$

we find that $A_x = f$ and $A_y = if$. In other words, we have found a (complex-valued) vector field whose line integral is the same as the contour integral of f.
• According to Stokes' theorem, $\int_\Sigma (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{S} = \int_{\partial\Sigma} \boldsymbol{A} \cdot d\boldsymbol{l}$ where Σ is a surface (with infinitesimal area vector $d\boldsymbol{S}$) and $\partial\Sigma$ its boundary (with line element $d\boldsymbol{l}$). Taking $\Sigma = \Omega$ to be the above region on the plane, this may be written as

$$\oint \boldsymbol{A} \cdot d\gamma = \int_\Omega (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot \hat{z} \, dxdy \quad \text{or} \quad \oint_{\gamma = \partial\Omega} A = \int_\Omega dA. \qquad (53)$$

Here $(\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot \hat{z} = \partial_x A_y - \partial_y A_x$. On the other hand, dA is the differential or exterior derivative of the 1-form $A = A_x dx + A_y dy$:

$$dA = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) dx \wedge dy. \qquad (54)$$

In our case, putting $A_x = f$ and $A_y = if$, we find

$$(\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot \hat{z} = i\partial_x f - \partial_y f = i(\partial_x + i\partial_y) f = -2\bar{\partial} f. \qquad (55)$$

Similarly,

$$dA = (i\partial_x f - \partial_y f)\, dx \wedge dy = -2\bar{\partial} f \; dx \wedge dy. \tag{56}$$

We see that $\boldsymbol{\nabla} \times \boldsymbol{A} = 0$ (or $dA = 0$) if and only if f is holomorphic (i.e., $\bar{\partial} f = 0$). In other words, the holomorphy of f is the curl-free nature of an associated planar vector field.

• So if f is holomorphic,

$$\oint_{\gamma = \partial\Omega} f dz = \oint_{\gamma} \boldsymbol{A} \cdot d\boldsymbol{\gamma} = -2 \int_{\Omega} \bar{\partial} f \; dxdy = 0 \tag{57}$$

Thus, if f is holomorphic in the region Ω between the two curves, then we may deform γ_1 to γ_2 without affecting the value of $\int_{\gamma_1} f dz$.

• **Cauchy's integral theorem**[4] states that the contour integral $\oint f dz$ along a closed curve γ vanishes if f is holomorphic on and inside the region bounded by γ.

• For instance, $\oint_\gamma z^n dz = 0$ for $n = 0, 1, 2, \ldots$ and any closed curve γ in the complex plane.

• The case $n = -1$ is an interesting exception. The function $f = 1/z$ is holomorphic everywhere except at the origin so we cannot apply Cauchy's theorem to a contour that encircles the origin. So we need to evaluate the integral by some other method. Suppose γ is a contour that goes round the origin once counterclockwise. To facilitate the integration, we may deform γ to the unit circle $\gamma(t) = e^{it}$ for $0 \leq t \leq 2\pi$ without affecting the value of the integral (this freedom to deform follows from Cauchy's theorem, say by including a narrow 'bridge' between the two contours to create a new closed contour that is simple and closed). Noting that $\dot{\gamma} = ie^{it}dt$,

$$\oint_{S^1} \frac{dz}{z} = \int_0^{2\pi} \frac{e^{it}}{e^{it}} idt = 2\pi i. \tag{58}$$

The nice thing about the unit circle is that the integrand became a constant function of t.

• On the other hand, $\oint_\gamma z^n dz = 0$ for $n = -2, -3, \ldots$. In these cases, we can use a circular contour $\gamma(t) = e^{it}$ for which

$$\oint_{S^1} z^n dz = \int_0^{2\pi} ie^{it} e^{int} dt = \left. \frac{e^{i(n+1)t}}{(n+1)} \right|_0^{2\pi} = \frac{e^{2\pi i(n+1)} - 1}{n+1} = 0. \tag{59}$$

• We may summarize these results using the Kronecker delta:

$$\frac{1}{2\pi i} \oint_\gamma \frac{dz}{z^n} = \delta_{n,1}. \tag{60}$$

[4] **Morera's theorem** furnishes, in a sense, a converse to Cauchy's theorem. Suppose $f(z)$ is a continuous function in some connected open domain D (it need not be simply connected) and suppose $\oint_\gamma f = 0$ for every piecewise continuous closed contour γ in D. Then Morera's theorem tells us that f must be holomorphic in D.

Notice that in this example, among poles, it is only the simple pole that contributes nontrivially to the contour integral. Recalling that the coefficient of $1/z$ is called the residue in a Laurent series representation of a function around $z = 0$, we may also write

$$\oint_\gamma \frac{dz}{z^n} = 2\pi i \operatorname{Res}_{z=0}\left(\frac{1}{z^n}\right). \tag{61}$$

Here γ is a contour that encloses the origin and winds around it once.

• **Winding number.** If the contour went round the origin n times (for some integer n), e.g., $\gamma_n(t) = e^{int}$ for $0 \le t \le 2\pi$, then

$$\oint_{\gamma_n} \frac{dz}{z} = 2\pi i n \tag{62}$$

We say that the winding number of the curve γ_n around the origin (or any point in the interior of the unit disk) is n.

• On the other hand, if a curve γ does not enclose the point z_0, then

$$\oint_\gamma \frac{dz}{z - z_0} = 0 \tag{63}$$

since $1/(z - z_0)$ is holomorphic on the contour and in the region within it.

• **Solution of CR equations via Cauchy integral formula.** The Cauchy integral theorem allows us to solve the Cauchy-Riemann equations for a function holomorphic in a region D, given the values of the function on the boundary $\partial D = \gamma$. The solution is expressed as a contour integral. To see this, suppose f is holomorphic in the region D enclosed by the closed curve $\gamma = \partial D$ and let z_0 lie in the interior of D. Then $\frac{f(z)-f(z_0)}{z-z_0}$ is also holomorphic in D. Despite appearances, z_0 is not a singular point since, by Taylor's theorem[5] for the differentiable function $f(z)$, $f(z) - f(z_0)$ vanishes at least as fast at $z - z_0$. Thus

$$\oint_\gamma \frac{f(z) - f(z_0)}{z - z_0} dz = 0 \quad \text{or} \quad f(z_0) \oint_\gamma \frac{dz}{z - z_0} = \oint_\gamma \frac{f(z)}{z - z_0} dz. \tag{64}$$

Hence we arrive at the promised integral expression for $f(z_0)$ when z_0 lies inside D:

$$f(z_0) = \frac{1}{2\pi i} \oint_\gamma \frac{f(w)}{w - z_0} dw. \tag{65}$$

This formula is sometimes called the **Cauchy transform**. It gives an integral representation of a holomorphic function in the interior of a region in terms of the value of the function on the boundary. It is an example of a linear integral transform since the RHS depends linearly on f.

[5] In more detail, $f(z) = f(z_0) + f'(z_0)(z - z_0) + r(z)(z - z_0)$ where $r(z) \to 0$ as $z \to z_0$.

• **Cauchy differentiation formula.** In fact, we can get integral expressions for all the derivatives of f at z_0 by differentiating under the integral sign (this is allowed since the integrands at each stage are C^1 functions of both w and z):

$$f'(z_0) = \frac{1}{2\pi i}\oint_\gamma \frac{f(w)dw}{(w-z_0)^2}, \quad \cdots, \quad f^{(n)}(z_0) = \frac{n!}{2\pi i}\oint_\gamma \frac{f(w)dw}{(w-z_0)^{n+1}}. \qquad (66)$$

• **Analyticity from holomorphy.** On the other hand, we can use (65) to obtain a convergent power series for f within the region D where it is holomorphic. We write (65) as

$$f(z) = \frac{1}{2\pi i}\oint_\gamma \frac{f(w)dw}{w-z_0-(z-z_0)}. \qquad (67)$$

Expanding $(w-z_0-(z-z_0))^{-1}$ in a power series,

$$\frac{1}{w-z_0-(z-z_0)} = \frac{1}{w-z_0}\left[1+\frac{z-z_0}{w-z_0}+\left(\frac{z-z_0}{w-z_0}\right)^2+\cdots\right] \qquad (68)$$

This series is absolutely convergent as long as $|z-z_0| < |w-z_0|$, i.e., provided z lies within a disc centered at z_0 that lies within D. Inserting this series in (67) and exchanging the order of integration and summation (which is justified since we have uniform convergence[6] in w) we get

$$f(z) = \frac{1}{2\pi i}\oint_\gamma \frac{f(w)dw}{w-z_0} + \frac{z-z_0}{2\pi i}\oint_\gamma \frac{f(w)dw}{(w-z_0)^2} + \frac{(z-z_0)^2}{2\pi i}\oint_\gamma \frac{f(w)dw}{(w-z_0)^3} + \cdots. \qquad (69)$$

Thus we have a convergent power series for $f(z)$ around any $z_0 \in D$. It follows that a holomorphic function is also complex analytic in the domain of holomorphy.

• Comparing with the Taylor series, we recover the Cauchy differentiation formulas (66) for the derivatives of f at z_0.

• Summing up, we showed that an analytic function is automatically holomorphic, as it is annihilated by $\bar{\partial}$. Then we showed that holomorphy implies the Cauchy integral formula. Finally, we showed that the Cauchy integral formula implies a convergent power series expansion (analyticity). Thus, we have explained that the three concepts: (a) analyticity, (b) holomorphy and (c) admitting a Cauchy contour integral formula, all describe the same class of functions. This is the class of analytic or holomorphic functions that we will work with[7].

1.13 Cauchy residue theorem and contour integrals

• **Cauchy's Residue Theorem.** Suppose γ is a simple closed contour that lies in a region where the function $f(z)$ is holomorphic except for isolated poles and essential

[6] Suppose $g_1, g_2, \ldots$ is a sequence of functions in a region Ω such that their sum $\sum_1^\infty g_n(z)$ converges uniformly to a function $g(z)$ in Ω. Then the contour integral of their sum is the same as the sum of their contour integrals along any piecewise smooth curve γ lying in Ω. Thus, uniform convergence allows us to perform term by term integration.

[7] Note that this also means that antiholomorphic functions are the same as antianalytic functions.

singularities that do not lie on γ. If γ encloses the singularities at $z_1, z_2, \ldots, z_n$ and goes round them once counterclockwise, then

$$\oint f(z)dz = 2\pi i \sum_{k=1}^{n} \text{Res}_{z=z_k} f(z). \tag{70}$$

The same formula applies to a meromorphic function where the number of poles enclosed may be infinite and the sum on the RHS may be an infinite sum. The formula is established by deforming the contour so that it it looks like a bunch of back and forth tracks on bridges between small circles around each of the isolated singularities. The contributions from the back and forth tracks cancel leaving small circles, each of which contributes to the contour integral as in the residue formula (61).

• The residue theorem is often applied to evaluate integrals. Three interesting classes of functions may be identified: rational functions integrated over the real line, periodic trigonometric functions over one period and more generally contour integrals of meromorphic functions (such as built from trigonometric and rational functions).

• **Real integral of a rational function.** For example, suppose we wish to evaluate the integral over the real line, of a rational function $f(x)$ (such as the Lorentzian $1/(x^2 + a^2)$) with no real poles. This may be done by 'completing' the real interval $(-R, R)$ with a semicircular contour $Re^{i\theta}$ in the upper ($0 \leq \theta \leq \pi$) or lower ($0 \geq \theta \geq -\pi$) half planes to get a closed (counterclockwise/clockwise) contour that encloses all the poles z_k in the upper/lower half plane. Assuming the function decays at least as fast as $1/|z|$ as $|z| \to \infty$, the semicircular contour does not contribute to the integral in the limit $R \to \infty$. We thus arrive at

$$\int_{-\infty}^{\infty} f(x)dx = 2\pi i \sum_{\Im z_k > 0} \text{Res}_{z=z_k} f(z) = -2\pi i \sum_{\Im z_k < 0} \text{Res}_{z=z_k} f(z). \tag{71}$$

• **Examples.**

1. $\int_{-\infty}^{\infty} \frac{dx}{x^2+1} = \pi$, since $(z^2+1)^{-1} = 1/(z+i)(z-i)$ has simple poles at $z = \pm i$ with residues $\pm(1/2i)$.

2. $\int_{\infty}^{\infty} \frac{dx}{(x^2+a^2)(x^2+b^2)} = \frac{\pi}{ab(a+b)}$ for $a, b > 0$. The function has simple poles at $\pm ia$ and $\pm ib$ with residues $\pm i/[2a(a^2 - b^2)]$ and $\pm i/[2b(b^2 - a^2)]$. Adding $2\pi i \times$ the residues, say in the upper half plane, we get the given result.

• **Integral of meromorphic function** (such as those built from exponentials and polynomials), say over the real line. One or other formula in (71) usually applies in this case, provided we close the contour in the half plane where the integrand decays as $|z| \to \infty$. If it decays in neither half plane, it may be possible to split the integrand into summands to which (71) applies.

• **Examples.**

1. $\int_{\infty}^{\infty} \frac{e^{ix}}{1+x^2} dx = \pi/e$. Here, $e^{iz}/(1 + z^2)$ decays exponentially ($e^{-\Im z}$) as we let $R \to \infty$ at any point on the contour $z = Re^{i\theta}$ for $0 < \theta < \pi$ in the upper half

plane. It has a simple pole at $z = i$ with residue $1/2ie$. Thus, upon closing the contour in the upper half plane, the integral evaluates to $2\pi i/2ie = \pi/e$.

2. $\int_{\infty}^{\infty} \frac{e^{-ix}}{1+x^2} dx = \pi/e$. Here, we need to close the contour in the lower half plane resulting in a clockwise contour. The residue at $z = -i$ is $-1/2ie$. The integral evaluates to π/e. This is not a surprise since the integral is the complex conjugate of the previous one.

3. $\int_{\infty}^{\infty} \frac{\cos x}{1+x^2} dx = \frac{\pi}{e}$. Here, $\cos z/(1+z^2)$ has simple poles at $z = \pm i$ but $\cos z$ grows exponentially in both the upper and half planes. So a common contour does not work. Instead, we split the integrand as the average of $e^{\pm ix}/(1+x^2)$ (or the real part of either) and find that the integral is the average of the previous two results.

4. $\int_\gamma \frac{e^z}{z^n} = 2\pi i \operatorname{Res}_{z=0} \frac{e^z}{z^n} = \frac{2\pi i}{(n-1)!}$ where γ is any closed contour that goes round the origin once counterclockwise. This is a very useful formula, we will return to it when we study the Gamma function.

• The **integral of a periodic trigonometric function** $g(\theta)$ over one period, say $0 \leq \theta \leq 2\pi$ in the real angular variable θ can often be converted via the substitution $z = e^{i\theta}$ to the integral of a meromorphic function $f(z)$ around the unit circle $|z| = 1$ counterclockwise.

• For example, consider the integral

$$I = \int_0^{2\pi} \frac{d\theta}{a - b\cos\theta} \quad \text{for} \quad a > b \geq 0. \tag{72}$$

Substituting $\cos\theta = \frac{1}{2}(z + 1/z)$ and using $d\theta = \frac{dz}{iz}$, we get

$$I = \oint_{|z|=1} \frac{dz}{iz(a - (b/2)(z+1/z))} = 2i \oint_{|z|=1} \frac{dz}{bz^2 - 2az + b} \tag{73}$$

Now, $bz^2 - 2az + b = b(z - z_-)(z - z_+)$ where

$$z_\pm = c \pm \sqrt{c^2 - 1} \quad \text{with} \quad 0 < z_- < 1 < z_+ \quad \text{for} \quad c = \frac{a}{b} > 1. \tag{74}$$

Thus, the pole at $z = z_-$ is the only one that contributes:

$$I = (2\pi i)(2i) \operatorname{Res}_{z=z_-} \frac{1}{b(z-z_-)(z-z_+)} = \frac{4\pi}{b(z_+ - z_-)} = \frac{2\pi}{\sqrt{a^2 - b^2}}. \tag{75}$$

1.13.1 Residue at infinity

• **Residue at** $z = \infty$**.** We defined the residue of an analytic function f at an isolated singularity $z_0 \in \mathbb{C}$ (pole or essential singularity) to be the coefficient of the $1/(z - z_0)$ term in the Laurent series expansion around z_0. The residue at a regular point is postulated to vanish. By Cauchy's theorem (70), the residue at z_0 is expressible as a contour integral around a closed curve γ that goes round the singularity (and no other singularity) once counterclockwise

$$\text{Res}_{z=z_0} f(z) = \frac{1}{2\pi i} \oint_\gamma f(z)\, dz. \tag{76}$$

• Sometimes, contour integrals may be evaluated in a simpler way by viewing the contour as going round the point at infinity. In this context, it is noteworthy that the contour $z = e^{it}$ for $0 \le t \le 2\pi$ is counterclockwise with respect to an interior point (such as $z = 0$) but is clockwise when viewed from infinity.

• Suppose an analytic function f is regular at $z = \infty$ or has an isolated pole or essential singularity at $z = \infty$. The residue at infinity is obtained by making a change of variable to $w = 1/z, dw = -(1/z^2)dz$ in (76) and considering a contour Γ that encircles the point $w = 0$ counter clockwise but does not enclose any point $w_k = 1/z_k$ where z_k is a finite singular point of f:

$$\text{Res}_{z=\infty} f(z) = -\frac{1}{2\pi i} \oint_\Gamma \frac{f(1/w)}{w^2} dw. \tag{77}$$

• We observe that the residue at $z = \infty$ is the coefficient of $1/w$ in the Laurent series for $-\frac{f(1/w)}{w^2}$ around $w = 0$. Note the factor of $-1/w^2$ coming from the integration element.

• We see that the function $f(z) = 1/z$, has the nonzero residue -1 at $z = \infty$, although the function is regular there.

• On the other hand, the residue of $f(z) = z$ vanishes at $z = \infty$ although the function has a simple pole there.

• Notice that the definition of the residue at infinity does not depend on the behavior of the function outside a neighborhood of infinity. In particular, the function may have nonisolated singularities (like branch cuts) in the finite complex plane without affecting the residue at infinity. However, in some cases, the residue at infinity may be related to the singularities of the function in the finite complex plane.

• If f has only isolated poles or essential singularities in the extended complex plane, we may express the residue at $z = \infty$ in terms of an integral along a counterclockwise contour $\tilde{\Gamma}$ that encloses all its singularities z_k in the finite complex z plane:

$$\text{Res}_{z=\infty} f(z) = -\frac{1}{2\pi i} \oint_{\tilde{\Gamma}} f(z)\, dz = -\sum_k \text{Res}_{z=z_k} f(z) \tag{78}$$

The minus sign ensures that this formula agrees with (77) where the contour was counterclockwise when viewed from the point at infinity. Alternatively, upon including the minus sign, we can view this contour as going counterclockwise around the point $z = \infty$ and not enclosing any other singularities of f.

• Under these circumstances, the residue at infinity of f is simply the negative of the sum of residues at isolated singularities in the finite complex plane. In particular, if f

is entire, then its residue at $z = \infty$ vanishes even though f must be singular at infinity if it is not a constant. This is reasonable since a contour that encircles the point at infinity can be shrunk to a point in the finite complex plane without encountering any singularities.

1.13.2 Summation of series using residue calculus

• We wish to evaluate the sum $S = \sum_{n=1}^{\infty} s_n$. Suppose we can realize the summands s_n as the residues of some function $f(z)$ that has poles at $n = 1, 2, 3, \ldots$. Then the series may be written as $(1/2\pi i)$ times the integral of $f(z)$ around contours that encircle the poles. In favorable cases, by deforming the contour, say to go around a finite number of other poles of f, we may be able to evaluate the integral more easily.
• Recall (from Assignment 3) that the poles of $\pi \cot \pi z$ are simple and occur at every integer $z = n$ with residue equal to one. To get a desired residue at these integer points, we may multiply $\pi \cot \pi z$ by a suitable function that is regular and nonvanishing there. The contour integration method is particularly effective if s_n is an even function of n, so that we may write $S = \frac{1}{2} \sum_{n \neq 0} s_n$. Otherwise, we might end up expressing one infinite sum in terms of another infinite sum. The power of this method, when it works, is that it helps us convert an infinite sum into a finite sum.
• Let us consider the well known series $S = \sum_{n=1}^{\infty} \frac{1}{n^2}$. Its convergence is ensured by the integral test. We now notice that

$$\frac{1}{n^2} = \text{Res}_{z=n} \frac{\pi \cot \pi z}{z^2} \quad \text{for integer} \;\; n \neq 0. \qquad (79)$$

Putting $f(z) = \frac{\pi \cot \pi z}{z^2}$, it follows that

$$S = \frac{1}{2} \sum_{n \neq 0} \frac{1}{n^2} = \frac{1}{2} \sum_{n \neq 0} \text{Res}_{z=n} f(z) = \frac{1}{2} \frac{1}{2\pi i} \sum_{n \neq 0} \oint_{\gamma_n} \frac{\pi \cot \pi z}{z^2} dz. \qquad (80)$$

Here, γ_n is a small circular contour winding once counterclockwise around the pole at $z = n$ and enclosing no other singularities of f. By deforming these contours, we may merge them into a pair of counterclockwise hairpin contours, one on the right half plane enclosing the poles at $z = 1, 2, 3, \ldots$ and another on the left half plane enclosing the poles at $z = -1, -2, -3, \ldots$. These hairpin contours are shaped so that their legs asymptotically approach straight lines at constant angles $\pm\theta_0, \mp\theta_0 + \pi$ where $0 < \theta_0 < \pi/2$ is any convenient positive angle. Next, we close these contours in the upper and lower half planes using large circular arcs to get a new hourglass/damaru (with bulging top and bottom) shaped contour γ that encloses only one pole of f: the pole of order three at $z = 0$. The circular contours do not contribute to the integral since $\cot \pi z$ is bounded (approaches one in magnitude) as $\Im z \to \pm\infty$ (show this!) so that for $z = Re^{i\theta}$, $f(z)dz \sim \pi R d\theta / R^2$ in magnitude as $R \to \infty$. The integrals over $-\theta_0 + \pi < \theta < \theta_0$ and $-\theta_0 > \theta > \theta_0 - \pi$ both vanish as $R \to 0$. By Cauchy's residue theorem,

$$S = -\frac{1}{2} \frac{1}{2\pi i} \oint_{\gamma} \frac{\pi \cot \pi z}{z^2} dz = -\frac{1}{2} \text{Res}_{z=0} \frac{\pi \cot \pi z}{z^2}. \qquad (81)$$

Now, since $\pi\cot\pi z$ is an odd function with a simple pole at the origin with unit residue, we have the Laurent expansion

$$\pi\cot\pi z = \frac{1}{z} + a_1 z + a_3 z^3 + \cdots \quad \text{or} \quad \frac{\pi\cot\pi z}{z^2} = \frac{1}{z^3} + \frac{a_1}{z} + a_3 z + \cdots . \tag{82}$$

Thus, the residue of f at $z=0$ is a_1 and $S = -\frac{1}{2}a_1$. It remains to find the linear Taylor coefficient a_1 of $\pi\cot\pi z$. By repeated differentiation, we find $\tan' = \sec^2, \tan'' = 2\tan\sec^2$ and $\tan''' = 2(\sec^2 + 3\tan^2\sec^2)$ so that $\tan'''(0) = 2$ and

$$\tan z = z + \frac{2z^3}{3!} + \cdots . \tag{83}$$

It follows that

$$\cot z = \frac{1}{z}\left(1 + \frac{z^2}{3} + \cdots\right)^{-1} = \frac{1}{z} - \frac{z}{3} + \mathcal{O}(z^3). \tag{84}$$

Consequently,

$$\pi\cot\pi z = \frac{1}{z} - \frac{\pi^2}{3}z + \mathcal{O}(z^3), \tag{85}$$

so that $a_1 = -\pi^2/3$. We conclude that $S = \sum_{n=1}^{\infty}\frac{1}{n^2} = -\frac{1}{2}a_1 = \pi^2/6$. This is a famous result (Basel problem) originally due to Euler (1734), who obtained it in a different way (using an infinite product representation of the entire sine function).

• A similar technique can be applied to sum series such as $\sum_{n=1}^{\infty}\frac{1}{n^2+a^2}$ using the function $f(z) = (\pi\cot\pi z)/(z^2+a^2)$.

• The technique also extends to alternating sign series. In this case, we replace $\pi\cot\pi z$ by $\pi\,\mathrm{cosec}\,\pi z$, whose only poles are simple poles at the integers $z = n$ with residues $(-1)^n$.

1.14 Cauchy principal value and the Hilbert transform

• So far, in contour integrals $\int_\gamma f dz$, our integrands were regular on the contour of integration. If f has a simple pole along the contour, then the integral diverges in the Riemann sense. However, it is still possible to assign a finite Cauchy principal value to such singular integrals via a symmetric limiting procedure. For concreteness, suppose the integral is over a real interval $[a,b]$ and f has a simple pole at $x = c$, then we define

$$\mathcal{P}\int_a^b f(x)dx = \lim_{\epsilon\to 0^+}\left[\int_a^{c-\epsilon} f(x)dx + \int_{c+\epsilon}^b f(x)dx\right]. \tag{86}$$

Loosely speaking, the positive and negative infinities from contributions on either side of the pole cancel out to give a finite answer. This works for simple poles but generally not for higher order poles.

• Cauchy principal value integrals may be evaluated using residue calculus by a judicious choice of contour before taking the appropriate limit.

• Let us consider an example

$$I = \mathcal{P}\int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz. \tag{87}$$

We notice that $|e^{iz}| = |e^{i\Re z} e^{-\Im z}| = e^{-\Im z}$ decays exponentially as $\Im z \to \infty$. Thus, we will close the contour on the upper half plane: $Re^{i\theta}$ for $0 \leq \theta \leq \pi$. In the limit $R \to \infty$, this semicircular arc does not contribute. However, the integrand e^{iz}/z has a simple pole at $z = 0$, which lies on the contour along the real axis. To implement the Cauchy principal value prescription we will skirt this pole by taking a detour in the upper half plane made of a semicircular arc of radius ϵ: $z = \epsilon e^{i\theta}$ with θ running from π to 0. The resulting closed contour consisting of the semicircular arcs of radius ϵ clockwise and R counterclockwise (when viewed from $z = 0$) and real intervals $[-R, -\epsilon]$ and $[\epsilon, R]$ does not enclose any poles of e^{iz}/z. Thus, by the Cauchy integral theorem,

$$\mathcal{P}\int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz + \lim_{\epsilon \to 0} \int_{\pi}^{0} i\epsilon e^{i\theta} d\theta \frac{\exp\left(i\epsilon e^{i\theta}\right)}{\epsilon e^{i\theta}} = 0. \tag{88}$$

In the limit $\epsilon \to 0$, the second integral is simply $-i\pi$. Hence, we get

$$I = \int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz = i\pi. \tag{89}$$

We may use this result to evaluate the Dirichlet integral

$$\int_{0}^{\infty} \frac{\sin x}{x} dx = \frac{1}{2} \Im \mathcal{P} \int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz = \frac{\pi}{2}. \tag{90}$$

• Another interesting consequence is the **Fourier transform of the 'Cauchy kernel'** $\epsilon(x) = 1/\pi x$:

$$\tilde{\epsilon}(k) = \mathcal{P}\int_{-\infty}^{\infty} \frac{e^{-ikx}}{\pi x} dx = -i \operatorname{sgn} k, \tag{91}$$

where $\operatorname{sgn} k$ is ± 1 for $k > 0$ and $k < 0$ while $\operatorname{sgn} 0 = 0$. To see how we arrive at this, we first observe that $\tilde{\epsilon}(0) = 0$ as the integrand is odd, resulting in the principal value integral vanishing. Next, notice the reality property $\tilde{\epsilon}(-k) = \tilde{\epsilon}(k)^*$. It therefore suffices to consider $k > 0$ and put $z = kx$. Then $dz/z = (kdx)/kx = dx/x$ so that

$$\tilde{\epsilon}(k) = \mathcal{P}\int_{-\infty}^{\infty} \frac{e^{-iz}}{\pi z} dz = \frac{I^*}{\pi} = -i \quad \text{for} \quad k > 0. \tag{92}$$

The reality property then implies that $\tilde{\epsilon}(k) = i$ for $k < 0$. Combining, we get $\tilde{\epsilon}(k) = -i \operatorname{sgn} k$. We will make use of this result shortly in discussing the Hilbert transform.

• Indenting and completing the contour using semicircular arcs is a technical tool to enable us to use Cauchy's theorem to evaluate the principal value integral. These are not part of the definition of the Cauchy principal value given in (86), which only involves the symmetric limit of the integrals along the two horizontal intervals. In particular, we could evaluate the Cauchy principal value by closing the contour using a small arc below the pole at $z = 0$.

- **Hilbert transform.** The Hilbert transform is a linear integral transform of a real-valued function $u(x)$ of a real variable x. It is defined as the Cauchy principal value of the convolution of $u(x)$ with the Cauchy kernel $1/\pi x$:

$$(Hu)(x) = \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{\infty}\frac{u(y)}{x-y}dy. \tag{93}$$

Although we do not discuss this aspect here, the Hilbert transform is important in complex analysis because it relates the real and imaginary parts of a complex analytic function. More precisely, if $u(x)$ is the boundary value along the real axis of the real part of a function $f(z)$ analytic in the upper half plane, then $(Hu)(x)$ is the corresponding boundary value of the imaginary part $\Im f(z = x + i\cdot 0)$.

- We should think of H as a linear operator that takes as input a vector u and transforms it into the vector Hu. The vectors u and Hu live in suitable function spaces like $L^2(\mathbb{R})$. A linear transformation A ($u \mapsto Au$) on a finite dimensional vector space with an inner product can be expressed in a basis e_i in terms of its matrix elements $A_{ij} = (e_i, Ae_j)$ with $(Au)_i = \sum_j A_{ij}u_j$. Being an integral transform, we may think of the Hilbert transform in terms of its integral kernel $\epsilon(x, y)$, which are the matrix elements of H in the position basis:

$$(Hu)(x) = \mathcal{P}\int_{-\infty}^{\infty}\epsilon(x,y)u(y)dy \quad \text{where} \quad \epsilon(x,y) = \frac{1}{\pi}\frac{1}{x-y}. \tag{94}$$

In Dirac notation, $\epsilon(x, y) = \langle x|H|y\rangle$ where $|x\rangle$ and $|y\rangle$ are position eigenstates with eigenvalues x and y. Evidently, the operator is not diagonal in position space: its matrix elements are nonzero even if $x \neq y$. However, since the kernel is translation invariant (depends on x and y only through their difference, so that $x \mapsto x + a$ and $y \mapsto y + a$ leaves ϵ unchanged), we might expect it to be simpler (diagonal) in Fourier or wave number space.

- Thus, we consider the Fourier transform of the kernel[8]:

$$\begin{aligned}\tilde{\epsilon}(k,l) &= \mathcal{P}\iint dxdye^{i(kx-ly)}\epsilon(x,y) = \mathcal{P}\int dxdye^{i(kx-ly)}\frac{1}{\pi}\frac{1}{x-y} \\ &= \mathcal{P}\iint dxdy\, e^{ik(x-y)}e^{i(k-l)y}\frac{1}{\pi(x-y)}.\end{aligned} \tag{96}$$

[8] In Dirac notation, $\tilde{\epsilon}(k, l) = \langle k|H|l\rangle$ where $|k\rangle$ and $|l\rangle$ are states of definite wave number. The relative sign $i(kx - ly)$ in the Fourier transform arises since we must take a hermitian adjoint to go from a ket vector to the dual bra vector:

$$\tilde{\epsilon}(k,l) = \langle k|H|l\rangle = \int\langle k|x\rangle\langle x|H|y\rangle\langle y|l\rangle dxdy = \int e^{ikx}\epsilon(x,y)e^{-ily}dxdy. \tag{95}$$

Here, we have used the completeness relation or resolution of the identity in the position basis $\int_{-\infty}^{\infty}|x\rangle\langle x|dx = I$ which says that the sum of the projections to the position eigenstates is the identity operator.

Now we change variable from $-\infty < x < \infty$ to $z = x - y$, which has the same range,

$$\tilde{\epsilon}(k,l) \quad = \quad \int_{-\infty}^{\infty} e^{i(k-l)y} dy \, \mathcal{P} \int_{-\infty}^{\infty} dz \frac{e^{ikz}}{\pi z} = 2\pi\delta(k-l) \, i \, \mathrm{sgn}\, k. \qquad (97)$$

We used the Fourier representation of the Dirac delta function and the Fourier transform of the Cauchy kernel (91). Note that $2\pi\delta(k-l)$ are the matrix elements $\langle k|I|l\rangle$ of the identity operator in the momentum basis. As expected, the kernel of the Hilbert transform is diagonal in wave number space. Since the diagonal entries are purely imaginary, this is the kernel of an antihermitian or skew-adjoint operator. It is a multiplication operator. The multiplier is $i \, \mathrm{sgn}\, k$. Thus, the amplitude of a positive wave number ($k > 0$) mode $\tilde{u}(k)$ is multiplied by i while negative wave number modes ($k < 0$) are multiplied by $-i$.

• Since the diagonal entries of the Hilbert transform kernel in Fourier space have a common value for $k < 0$ and reverse sign for $k > 0$, $\tilde{\epsilon}(k,l)$ can be used to model the Dirac vacuum (in null or light cone coordinates) where all negative momentum states are filled and positive momentum states are empty (more on this shortly).

• In fact, the Hilbert transform kernel squares to the negative identity. This is again possible to see, at least formally, in Fourier space[9]. Indeed,

$$\begin{aligned} \tilde{\epsilon}^2(k,m) &= \int \tilde{\epsilon}(k,l)\tilde{\epsilon}(l,m)\frac{dl}{2\pi} = \int 2\pi i \, \mathrm{sgn}\,(k)\delta(k-l) 2\pi i \, \mathrm{sgn}\,(l)\delta(l-m)\frac{dl}{2\pi} \\ &= -2\pi\delta(k-m)\, \mathrm{sgn}\, k \, \mathrm{sgn}\, m = -2\pi\delta(k-m) = -\langle k|I|l\rangle. \qquad (98) \end{aligned}$$

In position space, this means

$$(H(Hu))(x) = -u(x). \qquad (99)$$

• Suppose $v(x) = (Hu)(x)$, then $(Hv)(x) = -u(x)$. We say that u and v form a Hilbert transform pair.

• The quadratic condition $H^2 = -I$ can be used to encode the Pauli principle for a system of many fermions (like electrons or quarks). It can be rewritten as the condition that the 'density matrix' $\rho = \frac{1}{2}(I + iH)$ is a hermitian (symmetric) operator that is a projection: $\rho^2 = \rho$. Since the eigenvalues of ρ may be interpreted as occupation numbers of states, there can either be zero or one fermion in each state of definite momentum, as required by the Pauli principle. In fact, we see that ρ is diagonal in the Fourier basis and its eigenvalues are one for $k < 0$ and zero for $k > 0$. We say that the sea of negative wavenumber states are filled and the positive ones empty. This is called the Dirac vacuum.

[9] For $k = m \neq 0$, $\mathrm{sgn}\, k \, \mathrm{sgn}\, m = 1$. Some further justification is needed when both k and m are zero. In position space we would need to show that $\mathcal{P} \int \epsilon(x,y)\epsilon(y,z)dy = \delta(x-z)$. Use contour integration to show that the integral vanishes for $x \neq z$.

1.15 Sochotski-Plemelj ($i\epsilon$) formula and discontinuity in the Cauchy transform

• Let us consider the function g of the complex variable z defined by an integral over the real interval $[a, b]$:

$$g(z) = \int_a^b \frac{f(x)}{x-z} dx, \tag{100}$$

for some continuous function f of the real variable x. Here, g is called the Cauchy transform of f. Notice that this is *not* a principal value integral. For reasons similar to those that appeared in the discussion surrounding (69), the integral converges and defines an analytic function of the complex variable z away from the real interval $[a, b]$. We call the real segment $[a, b]$ a branch cut of g (this is simply a name for now, more on branch cuts later). It is natural to ask how g behaves as one approaches a point $x_0 \in [a, b]$ from above or below. In fact, the discontinuity of g across the branch cut is an interesting quantity that arises in many physics and mathematics problems. For instance, the discontinuity of the resolvent of the Hamiltonian across the branch cut along the continuous spectrum is related to the density of energy states. We will now derive a formula for the behavior of g as one approaches the cut from the upper/lower half planes. Such a formula was obtained and used by Sokhotsky and Plemelj in a related setup where the real interval is replaced by a closed contour in the complex plane and the upper and lower half planes are replaced by the regions outside and inside the contour.

• **The $i\epsilon$ prescription for the Cauchy transform.** Suppose $f(x)$ is an analytic function[10] on the real interval $[a, b]$ that does not vanish at any $x_0 \in [a, b]$. Then for any such x_0, $f(x)/(x - x_0)$ has a simple pole at x_0. As it stands, the Cauchy transform $\int_a^b \frac{f(x)}{x-x_0}$ is not defined for $x_0 \in [a, b]$. One way of making sense of this integral was via the Cauchy principal value $\mathcal{P} \int_a^b \frac{f(x)}{x-x_0} dx$ for $a < x_0 < b$, where a symmetric limit from either side of the simple pole was taken. There is another way of 'regularizing' the corresponding singular integral[11]. We modify the real contour of integration near x_0 via a semicircular arc of small radius $\epsilon > 0$ in the lower or upper half plane, thereby skirting the pole at x_0. Since f is analytic at every point of the interval, it must also be analytic in a small neighborhood of the interval in the complex plane. The resulting integrals are denoted

$$g(x_0 \pm i\epsilon) = \int_a^b \frac{f(x)}{x - (x_0 \pm i\epsilon)} dx. \tag{101}$$

The notation $x_0 \pm i\epsilon$ which literally means that the pole at x_0 has been shifted to the upper or lower half planes is often used to convey that the contour of integration has been modified to go around the pole at x_0 via a semicircular arc of radius ϵ centered

[10] Analyticity of f may be replaced by a weaker condition such as C^1 or C^0 for most of the results in this section.

[11] Regularization is a process by which a divergent quantity is made finite by changing its definition through a regulator, here called ϵ. One then studies what happens when the regulator is removed (here by taking $\epsilon \to 0$) and asks if the limit is finite and if so whether it depends on how the regulator is removed.

at x_0 in the lower or upper half planes respectively. Whether one indents the contour or shifts the pole, the limiting answers are the same. Henceforth, we will imagine indenting the contours. The horizontal portions of both these integrals tend to the Cauchy principal value as $\epsilon \to 0$. However, the limiting values of the semicircular integrals are not the same. Indeed, putting $z = x_0 + \epsilon e^{i\theta}$ (where $\theta \in [\pi, 2\pi]$ for the lower semicircle and $\theta \in [\pi, 0]$ for the upper semicircle), we find using $dz/(z - x_0) = i\epsilon e^{i\theta} d\theta/(\epsilon e^{i\theta}) = i d\theta$ that

$$\lim_{\epsilon \to 0} g(x_0 \pm i\epsilon) = \lim_{\epsilon \to 0} \int_a^b \frac{f(x)}{x - x_0 \mp i\epsilon} dx = \mathcal{P} \int_a^b \frac{f(x)}{x - x_0} dx \pm i\pi f(x_0). \quad (102)$$

The first term on the RHS involves the Cauchy principal value integral ($-\pi$ times the Hilbert transform of f); it does not depend on how the contour was modified in the vicinity of the pole. The second term on the RHS retains a memory of whether the contour went below/above the pole. It is a real version of the Sokhotsky-Plemelj formula where the integration is over a simple closed contour in the complex plane and one considers the limiting values of g as one approaches the contour from inside or outside. This result may also be written in a short-hand notation:

$$\lim_{\epsilon \to 0} \frac{1}{x - (x_0 \pm i\epsilon)} = \mathcal{P}\left(\frac{1}{x - x_0}\right) \pm i\pi\delta(x - x_0), \quad (103)$$

Here, $\delta(x - x_0)$ is the Dirac delta distribution supported at x_0. This relation among distributions is to be understood as applicable upon multiplying by a function $f(x)$ and integrating over a real interval containing x_0. We may say that the $i\epsilon$ prescription relates the integral transforms obtained from convolution by the Cauchy principal value and Dirac delta kernels.

• As noted, $g(x_0 \pm i\epsilon)$ do not have a common limit as $\epsilon \to 0$. In fact, the difference is given by

$$\lim_{\epsilon \to 0} [g(x_0 + i\epsilon) - g(x_0 - i\epsilon)] = 2\pi i f(x_0), \quad (104)$$

since the Cauchy principal value cancels out. This difference is simply the discontinuity in the Cauchy transform $g(z)$ across its branch cut $[a, b]$.

1.16 Multivalued functions, branch cuts and Riemann surfaces

• $\arg z$ **as a multivalued function.** We met our first example of a multivalued function quite early: the argument of a complex number $\arg z$. At the point $z = x + iy$, with $|z| \neq 0$, $\arg z$ is any angle θ such that $x = |z| \cos\theta$ and $y = |z| \sin\theta$. Evidently, if $\arg z = \theta$ is one such angle, then so is $\theta + 2n\pi$ for any integer n. We say that $\arg z$ is multivalued and that it has infinitely many branches labelled by the integers n. The origin is clearly singular, we could assign any value for θ at the origin and the above equations would be satisfied. Moreover, if we follow a simple closed curve counterclockwise around the origin, $\arg z$ does not return to its initially assigned value, but a value that is 2π larger. We call the origin a branch point. More generally, a branch point of a function is one with the property that the function does not return to its initially assigned value upon going round it via an arbitrarily small simple closed contour. In particular, the function cannot be defined as a continuous function in any neighborhood of the branch point. Points on the complex plane other than the origin are not branch points of $\arg z$. The $\arg z$ function can be defined as a continuous function in a small disk around any $z_0 \neq 0$. On the other hand, the point at infinity $z = \infty$ is another branch point of $\arg z$ since $\arg(1/w)$ displays a similar feature around $w = 0$, which is seen by putting $w = |w|e^{i\phi}$.

We could eliminate the multivalued nature of $\arg z$ by choosing the 'branch' where $-\pi < \arg z < \pi$ and call it the principal branch Arg z. Notice that the principal branch Arg z is discontinuous across the negative real axis: $\lim_{\epsilon \to 0^\pm}$ Arg $(x + i\epsilon) = \pm\pi$ for $x < 0$. We call the negative real axis a branch cut. We observe that in this example, the branch cut joins the two branch points.

• **The logarithm and its Riemann surface.** We will be interested in multivalued holomorphic functions. It is easy to see that $\arg z$ is not a holomorphic function. It is real-valued and nonconstant ($u_x, u_y \neq 0, v_x = v_y \equiv 0$) and therefore cannot satisfy the Cauchy-Riemann equations. However, one verifies that $\arg z = \arctan(y/x)$ is harmonic. Thus, it may be realized[12] as, say, the imaginary part $v(x, y)$ of a holomorphic function $f = u + iv$. Putting $v(x, y) = \arctan(y/x)$ we find $v_y = \frac{x}{x^2+y^2}$ and $v_x = -\frac{y}{x^2+y^2}$. Thus

$$u_x = v_y = \frac{x}{x^2 + y^2} \quad \text{and} \quad u_y = -v_x = \frac{y}{x^2 + y^2}. \tag{105}$$

Now, we see that the function $u(x, y) = \frac{1}{2}\log(x^2 + y^2)$ has precisely these partial derivatives. So we have found the harmonic conjugate of $v = \arg z$, it is $u = \ln|z|$, which is uniquely defined at all points in the complex plane punctured at the origin. Thus, we are led to consider the holomorphic function $\log z \equiv \ln|z| + i \arg z$, which inherits the multivalued features of $\arg z$. In particular, $z = 0$ is a branch point: $\log z$ increases by $2\pi i$ when we go once counterclockwise round the origin. The same applies to $z = \infty$ if we consider the function $g(w) = \log(1/w) = -\log w$ and follow a contour counterclockwise around the origin of the w plane: $g(w)$ decreases by $2\pi i$. On the other hand, upon following a small closed curve that does not enclose $z = 0$ or $w = 0$, $\log z$ returns to its initial value. So $\log z$ has only two branch points. If we 'cut out' a curve (called the branch cut) joining the branch points $z = 0, \infty$, then we may define a singlevalued holomorphic function on the complement. For instance, following the same convention as for $\arg z$, we take $\log z$ to have a branch cut along the negative real axis. We can then define a sequence of functions labelled by the integers, each of which is singlevalued in the complex plane with the negative real axis cut out:

$$\log_n z = \ln|z| + i(\theta + 2n\pi) \quad \text{for} \quad -\pi < \theta < \pi \quad \text{and} \quad |z| > 0. \tag{106}$$

[12] There is another way of constructing a holomorphic function from $v = \arg z$. Since it is harmonic, $\bar{\partial}\partial v = 0$, so ∂v must be holomorphic. In our case, $\partial v = \frac{1}{2}(\partial_x - i\partial_y)\arctan(y/x) = -\frac{1}{2}(y/(x^2 + y^2) + ix/(x^2 + y^2)) = -i/2z$. However, the resulting holomorphic function $-i/2z$ is singlevalued in this case.

Here $\log_n(z)$ is a holomorphic function on the cut complex plane with a discontinuous increase of $2\pi i$ as one crosses the cut from below to above. The part of the plane just above the cut ($\theta = \pi - \delta$) is called the upper lip while the part just below the cut ($\theta = -\pi + \delta$) is the lower lip. Interestingly, there is no such discontinuity between the values of $\log_n$ and $\log_{n+1}$ as one crosses the cut:

$$\lim_{\delta \to 0^+} \log_n(|z|, \pi - \delta) = \lim_{\delta \to 0^+} \log_{n+1}(|z|, -\pi + \delta) = \ln|z| + 2n\pi + i\pi. \quad (107)$$

It is as though these functions are part of a 'larger' continuous (in fact analytic) function. Riemann's idea was to assemble all these functions on the cut complex plane into one function on a larger surface, now called a **Riemann surface**. Thus, the multivalued logarithm becomes singlevalued when considered as a function on its Riemann surface Σ. The Riemann surface of the logarithm is obtained by stacking the cut complex planes ('sheets' labeled by n) one on top of the other with the upper lip of the $n^{\rm th}$ sheet glued to the lower lip of the $n+1^{\rm st}$ sheet. The Riemann surface is shaped like a helix: a spiral staircase, with the central pillar shrunk vertically to a point, the branch point at the origin. The Riemann surface of the logarithm is an infinite sheeted cover[13] of the punctured complex plane. Sketch it. The function $\log_n$ is viewed as a singlevalued function on the $n^{\rm th}$ sheet. It is called the $n^{\rm th}$ branch of the logarithm. Moreover, $\log_0$ is designated the principal branch. The branch points at $z = 0$ and $z = \infty$ are singular points of the logarithm, viewed as an analytic function on its Riemann surface. However, the branch cut singularity has been dispensed with. Upon making a counterclockwise circuit around the origin, we ascend from one Riemann sheet to the next. A branch point is said to be of order $k = 0, 1, 2, 3, \cdots$ if upon making a minimum of $k+1$ complete cycles around it counterclockwise, the function returns to its original value. The **logarithmic branch points** at $z = 0$ and $z = \infty$ are of infinite order. We call them **transcendental branch points**.

• There are multivalued analytic functions that have branch points of finite order, where the corresponding Riemann surface has finitely many sheets. Examples are provided by algebraic functions such as $f(z) = z^{1/2}$ or $z^{1/3}$ which have so-called algebraic branch points.

• **The square-root function $f(z) = \sqrt{z}$ and its Riemann surface.** We are familiar with the possibility of two distinct square-roots (that differ by a sign) for a complex number other than zero. Putting $z = re^{i\theta}$ where $r > 0$ and say, $0 \leq \theta < 2\pi$, we see that $f_1(z) = \sqrt{r}e^{i\theta/2}$ and $f_2(z) = -\sqrt{r}e^{i\theta/2} = \sqrt{r}e^{i(\theta+2\pi)/2}$ are both possible square-roots. Thus, $f(z) = \sqrt{z}$ is a multivalued (actually double-valued) function. Moreover, if we follow a curve around the origin once counterclockwise, we see that both f_1 and f_2 return to the negative of their initial values: $f_j(r, \theta + 2\pi) = -f_j(r, \theta)$ for $j = 1, 2$. Thus, $z = 0$ is a branch point. In a similar fashion, $z = \infty$ is also a branch point. Putting $w = 1/z = re^{i\theta}$, $g(w) = f(1/w) = \frac{1}{\sqrt{r}}e^{-i\theta/2}$ has the same

[13] This means there is a covering map or projection π from the Riemann surface Σ to the complex plane punctured at the origin $\mathbb{C}^*$ (called the base). The projection $\pi : \Sigma \to \mathbb{C}^*$ has the property that every point $z \in \mathbb{C}^*$ in the base, has an open disk D_z around it whose inverse image (preimage) under π is a disjoint union of open disk-like neighbourhoods D_z^n (for $n \in \mathbb{Z}$) in Σ, one on each sheet, with $\pi : D_z^n \to D_z$ being an invertible continuous map when restricted to D_z^n for each $n \in \mathbb{Z}$.

sign reversal feature when a curve is followed once around $w = 0$. There are no other branch points for the square-root function. Given that we picked θ to lie between zero and 2π, we will choose a branch cut along the positive real axis that joins the branch points at $z = 0$ and $z = \infty$. We notice that $f_1(z)$ is discontinuous across the branch cut: $f_1(r, 0^+) = \sqrt{r}$ while $f_1(r, 2\pi^-) = -\sqrt{r}$ so f_1 jumps up by $2\sqrt{r}$ when going from the lower lip of the cut to the upper lip. On the other hand, f_2 takes the values $-\sqrt{r}$ and $\sqrt{r}$ just above and below the cut. So f_2 drops down by $2\sqrt{r}$ when going from the lower lip of the cut to its upper lip. Thus f_1 just below the cut agrees with f_2 just above the cut and f_2 just below the cut agrees with f_1 just above the cut. Given these properties, we may define the Riemann surface for $\sqrt{z}$ as consisting of two sheets of the cut complex plane with the lower lip of sheet one glued to the upper lip of sheet two across the cut and vice-verse. Try to draw this surface. This 2-sheeted Riemann surface (away from the branch points) is a double cover of the complex plane. On this Riemann surface, we may define a single-valued continuous (indeed analytic) function $\sqrt{z}$, equal to f_1 on the first sheet and f_2 on the second sheet. We see that winding twice around the origin, the function returns to its original value although this did not happen after one cycle. Thus, the square-root branch points at $z = 0$ and $z = \infty$ are of order one.

• $n^{\rm th}$ **root branch points.** In a similar vein one may consider the multivalued function $f(z) = z^{1/n}$ for $n = 2, 3, 4, \ldots$. In each case, there are $n^{\rm th}$ root branch points of order $n - 1$ at $z = 0$ and $z = \infty$. All these are algebraic branch points. They may be joined by a branch cut running along the positive real axis. The corresponding Riemann surfaces have n sheets with the lower lip of the $n^{\rm th}$ sheet being glued to the upper lip of the first one.

• The function $f(z) = \sqrt{z^2 - 1}$ or more generally $f(z) = \sqrt{(z - a)(z - b)}$ for $a \neq b$ has square-root branch points at $z = a, b$ and its branch cut can be chosen to lie along the line segment joining a and b (what is another nice choice?). The corresponding Riemann surface has two sheets. Although $f(z)$ fails to return to its initial value upon following a closed contour around any one of the branch points, it does return to its initial value when the cycle encloses both branch points.

• The function $f(z) = z^\alpha$ where α is irrational (and possibly complex) has branch points at $z = 0, \infty$. However, these are not algebraic branch points, they are sometimes called winding points and are examples of transcendental branch points. Putting $z = re^{i\theta}$, we see that when $\theta \mapsto \theta + 2n\pi$, $z^\alpha \mapsto e^{2n\pi i\alpha} z^\alpha$. For irrational α the prefactor cannot equal one. So these winding points are branch points of infinite order. The corresponding Riemann surfaces have infinitely many sheets labelled by an integer n, just as for the logarithm. In fact, the two functions are related: $z^\alpha = e^{\alpha \log z}$.

1.17 Analytic continuation

• **Analytic functions are rigid.** The condition of analyticity is a very strong one. In particular, it allows us to find the behavior of an analytic function elsewhere from the knowledge of its values on a rather restricted set of points. Analytic functions are rigid in this sense. Cauchy's integral formula (65) is one manifestation of this: if f is analytic on a simple closed curve γ and in the domain D enclosed by γ, then its values in D are determined by its values on γ.

• **Identity theorem: an analytic function is determined by its values on a set with an accumulation point.** If we are willing to forego an explicit formula like Cauchy's, then we can make do with less than the knowledge of f on the boundary of D. Indeed, suppose f is analytic in an open connected domain D. Then its values in D are uniquely determined by its values on a set of points Σ in D which have a point of accumulation z_∞ in D. To see the uniqueness, let us suppose there is another function $g(z)$ that is also analytic in D and agrees with f on Σ. Then the difference $f - g$ is analytic in D and has an accumulation point of zeros at z_∞. From the discussion in §1.10, we deduce that $f - g$ must be identically zero in D, whence $f = g$.

• In particular, if two functions $f_1(z)$ and $f_2(z)$ that are analytic in D agree on a segment of a curve[14] or in an open neighborhood[15] of a point of D, then they must coincide on D. A limitation of these uniqueness theorems is that they do not tell us how to find the values of $f = f_1 = f_2$ on the rest of D. A strategy to do this would be to use the values of f, say, along the curve, to compute all its derivatives at some point z_0 on the curve and then use the method of analytic continuation described below.

• As alluded to, knowledge of all derivatives of f at one point z_0 in the given connected domain D of analyticity allows us to determine it throughout D and even extend it beyond in favorable cases. This leads to the idea of analytic continuation. Given that f is analytic at $z_0 \in D$, we may expand it in a convergent Taylor series $f(z) = \sum_{k=0}^\infty a_k(z-z_0)^k$ where $a_k = \frac{1}{k!} f^{(k)}(z_0)$. This series has a nonzero radius[16] of convergence r_0. Now, suppose z' is another point in D. If $|z' - z_0| < r_0$ we may determine $f(z')$ simply by evaluating this series. Even if z' lies outside the disk of convergence $D_0 = \{z : |z - z_0| < r_0\}$, connectedness guarantees that there is a continuous curve $\gamma : [0,1] \to D$ that joins them: $\gamma(0) = z_0$ and $\gamma(1) = z'$. Next, given any other point $z_1 \in D_0$ that lies on γ, we may differentiate the convergent Taylor series term by term to evaluate $f(z_1)$ and all derivatives of f at z_1. These are used to obtain the Taylor series of f around z_1.

$$f(z) = \sum_{k=0}^\infty \frac{1}{k!} f^{(k)}(z_1)(z-z_1)^k. \tag{108}$$

This Taylor series around z_1 will have a positive radius of convergence r_1. Now, it is always possible to choose $z_1 \in \gamma \cap D_0$ in such a way that a portion of the disc $D_1 = \{z : |z - z_1| < r_1\}$ of convergence around z_1 lies outside D_0 and covers a portion of γ outside D_0. This allows us to evaluate f outside D_0 but inside D_1. We have thus extended the domain of knowledge of f to $D_0 \cup D_1$ and thereby to a larger segment of the curve γ. Repeating this procedure, we cover the curve γ by a

[14] For example, the real segment $[-1,1]$ has the sequence $z_n = 1/n$ that accumulates at the point $z = 0$ within the segment.

[15] For example, the neighborhood $|z| < 2$ has the sequence $z_n = 1/n$ that accumulates at the point $z = 0$.

[16] Notice that the radius of convergence r_0 is at least as large as the distance from z_0 to the nearest point on the boundary ∂D. This is because (by assumption) f does not have any singularities within D, so the nearest singularity can only occur on ∂D. Illustrate with a figure.

finite sequence of, say, $n+1$ overlapping disks in each of which f has a computable convergent Taylor expansion. In this manner, we arrive at a convergent Taylor series for f around a point z_n whose disk of convergence includes the point z'. We say that we have analytically continued f from z_0 to the point z'. By the identity theorem, the analytic continuation of f from D_0 to the point z' is unique. We are guaranteed to get the same answer for $f(z')$ regardless of which curve γ we pick and which sequence of points $z_1, z_2, \cdots, z_n$ on γ that we pick in the above procedure.

In fact, as long as we do not hit a singularity along such a curve γ (where the radii of convergence of successive disks would shrink to zero), we can use this method to analytically continue f even outside the originally known domain D of analyticity. It is also noteworthy that this method can in principle also be used to find the singularities of f: the radii of convergence of the successive disks would shrink to zero if the chosen curve γ passed through a singularity. Analytic continuation naturally stops there. On the other hand, as long as f is singlevalued, by analytically continuing it, we may find its full domain of analyticity. The boundary of this domain (if any), beyond which the function cannot be analytically continued is called its natural boundary. We will see examples of such a natural boundary below. In fact, even if f is multivalued, we can use analytic continuation to find its Riemann surface and all branches of the function starting from the knowledge of its Taylor series at just one point!

• **Example.** As an application of this method of analytic continuation, consider the analytic function $f(z) = 1 + z + z^2 + \cdots$ defined by this geometric series[17] on the unit disk $|z| < 1$. Suppose we wish to analytically continue this function to $z' = -1$. We choose the curve γ to be a line segment from $z = 0$ to $z' = -1$. We pick the point $z = -1/2$ and use the formula for the sum of the geometric series to evaluate

$$f'(z) = 1/(1-z)^2, \quad f'' = \frac{2}{(1-z)^3}, \cdots, f^{(n)}(z) = n! \frac{1}{(1-z)^{n+1}}. \tag{109}$$

Thus, we get the Taylor series for f around $z_1 = -1/2$, which we denote f_1:

$$f_1(z) = \sum_0^\infty (2/3)^{n+1} (z + \frac{1}{2})^n. \tag{110}$$

This series converges for

$$(2/3)|z + \frac{1}{2}| < 1 \quad \text{or} \quad |z + \frac{1}{2}| < 3/2 \tag{111}$$

So the radius of convergence is $3/2$ and the disk of convergence includes the point $z' = -1$. Thus, although the original geometric series would not work, we evaluate this new series at $z' = -1$ to get $f(-1)$ and differentiate it to find the derivatives at -1. We find

$$f^{(n)}(-1) = \frac{n!}{2^{n+1}} \tag{112}$$

[17] In this case, we know that this series represents the function $1/(1-z)$ within the unit disk. However, the method we describe works even without such a formula.

Thus we arrive at a convergent Taylor series for f around -1

$$f_2(z) = \sum_0^\infty \frac{(z+1)^n}{2^{n+1}}, \tag{113}$$

with radius of convergence equal to 2. We have analytically continued f well outside the unit disk.

• **Natural boundary.** The power series

$$f(z) = \sum_{k=0}^\infty z^{2^k} \quad \text{and} \quad g(z) = \sum_{k=0}^\infty z^{k!} \tag{114}$$

are absolutely convergent in the open disk $|z| < 1$. Thus, they define analytic functions f and g on the open unit disk. However, it can be shown that f has singularities at every $2^{k^{\text{th}}}$ root of unity, which are densely distributed on the unit circle. Similarly, g has singularities at every point on the unit circle. Consequently, neither function can be analytically continued outside the unit disk. We say that the unit circle is a natural boundary for the analytic functions f and g. These are examples of lacunary series. Both power series have large lacunae (gaps) between powers of z with nonzero coefficients.

• **Analytic continuations of each other.** Sometimes, we say that two functions f_1 and f_2 are analytic continuations of each other. To understand what this means, suppose f_1 and f_2 are analytic in open domains D_1 and D_2 that intersect nontrivially and suppose they agree on the overlap $D_1 \cap D_2$. Then by uniqueness, the analytic continuation of f_1 to D_2 must coincide with f_2 there and the analytic continuation of f_2 to D_1 must coincide with f_1 there. In other words, there is a common analytic function on the union $D_1 \cup D_2$ which restricts to f_1 and f_2 on the subdomains. We say that f_1 and f_2 are analytic continuations of each other. It is noteworthy that the intersection $D_1 \cap D_2$ need not be connected and $D_1 \cup D_2$ need not be simply connected.

• **Permanence of functional equations under analytic continuation.** In practice, there may be more convenient ways to achieve analytic continuation than through power series. For instance, functional equations can be useful. Suppose f is analytic in D and satisfies a sufficiently nice[18] functional equation $R(f(z), z) = 0$ such as the periodicity condition $f(z+1) - f(z) = 0$ whenever z and $z+1$ are in D. Then the analytic function $g(z) = f(z+1) - f(z)$ is identically zero when $z, z+1 \in D$. Suppose f is analytically continued to a larger domain D'. Then $g(z) = f(z+1) - f(z)$ can also be analytically continued to places where $z, z+1 \in D'$ and must be the zero function by the uniqueness of analytic continuation (identity theorem). Thus, the relation $f(z+1) - f(z) = 0$ must also hold when $z, z+1 \in D'$. This is the permanence of functional equations. In the present example, we may use periodicity to extend the function and thereby arrive at its analytic continuation (at least to some parts of the complex plane) without having to use power series. For example, we could use its 2π

[18] By nice, we mean that the operations involved in the relation $R(f(z), z)$ (such as sums, products, etc.) do not destroy analyticity.

periodicity to analytically continue the function $\sin z$ if it was originally defined, say, in the strip $0 < \Re z < 2\pi$. In the case of the Gamma function, a functional equation enables us to analytically continue it to the left half plane (see §1.17.1).

1.17.1 Example: Analytic continuation of the Gamma function

For $n = 1, 2, 3, \ldots$ the factorial is defined as $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$, with the recursive property $n! = n(n-1)!$. Euler's Gamma function generalizes the factorial to complex arguments. It is defined by Euler's integral of the second kind:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt. \tag{115}$$

This integral converges absolutely for $\Re z > 0$. For $z = 0$, the integral diverges due to the $1/t$ behavior of the integrand as $t \to 0^+$. The integral also converges uniformly on bounded subsets on the right half of the complex z plane. Thus, by a general theorem on such integral representations, it defines an analytic function for $\Re z > 0$. Alternatively, one may differentiate under the integral sign and see that $\Gamma'(z)$ exists for $\Re z > 0$ so that Γ is holomorphic on the open right half plane. Although, the integral (115) does not converge for $\Re z \leq 0$, we would like to extend (analytically continue), to the extent possible, the Gamma function to an analytic function on the left half plane. To do so, it would help to relate $\Gamma(z)$ to $\Gamma(z+1)$ (the recursion formula $n! = n(n-1)!$ suggests that such a relation may exist).

- **Functional equation.** It is clear that $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$. Integrating by parts, for $\Re z > 0$ we get

$$\Gamma(z+1) = \int_0^\infty e^{-t} t^z dz = [-t^z e^{-t}]_0^\infty + \int_0^\infty e^{-t} z t^{z-1} dt = z\Gamma(z). \tag{116}$$

Thus, for a natural number, $\Gamma(n+1) = n!$. If Γ may be analytically continued outside the right half plane, then by the permanence of functional equations, the relation

$$\Gamma(z+1) = z\Gamma(z) \tag{117}$$

must hold also for $\Re z \leq 0$. To begin with, we may use it to define Γ in the strip $-1 < \Re z \leq 0$ in terms of its values in the strip $0 < \Re z \leq 1$:

$$\Gamma(z) = \frac{1}{z}\Gamma(z+1) = \frac{1}{z}\Gamma(1) + \Gamma'(1) + \mathcal{O}(z). \tag{118}$$

In particular, we see that Γ must be regular in this strip aside from a simple pole at $z = 0$ with residue $\Gamma(1) = 1$.

- The constant $\gamma = \Gamma'(1) \approx 0.57721566\ldots$ is called Euler's constant.
- Next, we define Γ in the strip $-2 < z \leq -1$:

$$\Gamma(z+1) = z\Gamma(z) = z(z-1)\Gamma(z-1) \quad \Rightarrow \quad \Gamma(z-1) = \frac{1}{(z-1)z}\Gamma(z+1) \tag{119}$$

As before, we find that Γ is regular in this strip except for a simple pole at $z=-1$. In fact,

$$\Gamma(z-1)=\frac{1}{z(z-1)}\Gamma(1)+\frac{1}{z(z-1)}z\Gamma'(1)+\mathcal{O}(z)=-\frac{\Gamma(1)}{z}-\Gamma'(1)+\mathcal{O}(z). \quad (120)$$

So the residue of the simple pole at $z=-1$ is $-\Gamma(1)=-1$.

• More generally,

$$\Gamma(z-n)=\frac{\Gamma(z+1)}{(z-n)(z-n+1)\cdots z}=\frac{\Gamma(1)}{(-1)^n n(n-1)\cdots 1}\frac{1}{z}+\mathcal{O}(z^0). \quad (121)$$

Thus Γ has a simple pole at every nonpositive integer $-n=0,-1,-2,-3,\ldots$ with residue $(-1)^n/n!$ and is regular everywhere else. In particular, it is a meromorphic function.

1.18 Entire functions

• An entire function $f(z)$ is one which is complex analytic at every point z of the finite complex plane. Entire functions are also called integral functions. The partition function of a classical statistical mechanical system with a finite number of degrees of freedom must be an entire function of the inverse temperature. When one takes the thermodynamic limit (number of particles to infinity and volume to infinity, holding the density finite) it can develop singularities at some temperatures which may be interpreted in terms of phase transitions.

• **Polynomials** are the simplest entire functions. An entire function (such as e^z) that is not a polynomial is called a **transcendental entire function**. Since they are generalizations of polynomials, it is worth noting some relevant properties of polynomials.

• A polynomial of degree N has a **pole** of order N at $z=\infty$. To see this, we put $w=1/z$ and examine the behavior around $w=0$. Transcendental entire functions (such as e^z) have **essential singularities** at $z=\infty$.

• Nonconstant polynomials are unbounded. The same applies to entire functions. This is called the **Cauchy-Liouville Theorem**, which states that a bounded entire function must be a constant.

• By the **fundamental theorem of algebra**, a polynomial of degree N with complex coefficients

$$p(z)=c_N z^N+c_{N-1}z^{N-1}+\cdots+c_1 z+c_0 \quad \text{with} \quad c_N\neq 0, \quad (122)$$

has N complex roots $z_1,z_2,\cdots,z_N$ (repeated according to multiplicity). A polynomial may be factorized into linear factors, one for each root:

$$p(z)=c_N\prod_{k=1}^{N}(z-z_k). \quad (123)$$

Thus, the polynomial is determined (up to the constant c_N) by the locations of its zeros. One may ask if a similar result holds for entire functions: to what extent are they determined by their zeros?

• The **exponential function** e^z is entire. It has no zeros on the finite complex plane. Thus, the exponential of an entire function is also entire and again has no zeros[19]. Examples include $e^{z^3+3z+2i}$ and e^{e^z}. It follows that an entire function without any zeros is determined at best up to multiplication by the exponential of an entire function.

• This is more or less the general pattern. An entire function $f(z)$ is determined by its zeros $z_1, z_2, \cdots$ up to multiplication by $e^{h(z)}$ for some entire function h. Assuming f is not identically zero, the zeros, if infinite in number, cannot have any finite accumulation point since f would have to be singular there. Thus, the zeros can only accumulate at infinity. So when there are infinitely many zeros, we will assume that the sequence $z_k \to \infty$. As with a polynomial, an entire function is expressible as a (possibly infinite) product. Hadamard developed a canonical form for this product for entire functions of finite genus (to be defined below). This was generalized by Weierstrass. The linear factor $(z - z_k)$ for each nonzero root is written as $(1 - z/z_k)$ times a nonvanishing entire function to ensure that the product converges. Thus, we denote the nonzero roots by $z_k \neq 0$ and allow for f to have a zero of order $m \geq 0$ at the origin. Hadamard's infinite product representation of an entire function takes the form

$$f(z) = z^m e^{Q(z)} \prod_{k \geq 1} \left(1 - \frac{z}{z_k}\right) \exp\left(\frac{z}{z_k} + \frac{1}{2}\frac{z^2}{z_k^2} + \cdots + \frac{1}{p}\frac{z^p}{z_k^p}\right). \tag{124}$$

Here $Q(z)$ is a polynomial of degree q. On the other hand, p is the smallest nonnegative integer such that the reciprocal root power sum

$$\sum_{k=1}^{\infty} \frac{1}{|z_k|^{p+1}} \tag{125}$$

converges.

• The Hadamard product for e^z is e^z, it has no zeros so the product is empty so that $p = 0$. In this case, $Q(z) = z$ so $q = 1$.

• On the other hand, $\sin \pi z$ has zeros at every integer $z_k = k$ and although the harmonic sum $\sum_{k=1}^{\infty} \frac{1}{k^1}$ diverges, $\sum_{k=1}^{\infty} \frac{1}{k^{1+1}} = \pi^2/6$. So $p = 1$. It has the Hadamard canonical product

$$\sin \pi z = \pi z \prod_{k \in \mathbb{Z}, k \neq 0} (1 - z/k) e^{z/k}. \tag{126}$$

• Similarly, $\cos \pi z$ has zeros at odd multiples of $1/2$, and we find that the reciprocal root square sum again converges so that $p = 1$. It admits the canonical product

$$\cos \pi z = \prod_{k \in \mathbb{Z}, k \text{ odd}} (1 - 2z/k) e^{2z/k}. \tag{127}$$

• We will say more about infinite products and their convergence later if time permits.

[19] In fact, any nonvanishing entire function may be expressed as the exponential of an entire function, which is called its holomorphic logarithm.

• The **genus of an entire function** that admits a Hadamard product representation is defined as the larger of the integers p and q: $g = \max\{p, q\}$. Evidently, it is the largest power that appears in the exponential factor in the product. So the genus tells us something about how the function grows for large $|z|$. A polynomial has genus zero as there is no exponential factor at all. The exponential function e^z has genus one. The transcendental entire functions e^{z^n} for $n = 1, 2, 3 \cdots$ have genus n. Through p, the genus also encodes the rate at which the roots approach infinity. For example, $\sin z$ has roots at integer multiples of π and $\cos z$ has roots at odd multiples of $\pi/2$. One sees that the smallest nonnegative integer for which the series (125) converges is $p = 1$ in both cases. Moreover, the polynomial Q vanishes in both cases. So the entire functions $\sin z$ and $\cos z$ also have genus one. If the roots approach infinity more slowly, then p would have to be larger. On the other hand, if there are only a finite number of roots, then $p = 0$. A transcendental entire function with finitely many zeros is a polynomial times the exponential of an entire function.

2 Manifolds

2.1 Some references on manifolds

1. Loring Tu, *An Introduction to Manifolds*.
2. Bernard Schutz, *Geometrical Methods of Mathematical Physics*.
3. Govind Krishnaswami, *Classical Mechanics: From Particles to Continua and Regularity to Chaos*, Appendix B.

2.2 The concept of a manifold

By a manifold, we have in mind a space like a circle (denoted S^1), the plane, the surface of a sphere (S^2) or the 3d Euclidean space $\mathbb{R}^3$ in which a particle can move. A manifold is a space where every point[20] has an open neighborhood[21] that looks like[22] Euclidean space $\mathbb{R}^n$ for some fixed positive integer n, which is called the dimension of the manifold. By considering sufficiently many such overlapping open neighborhoods of points, we obtain an open covering of the space. Thus, roughly, a manifold is a space that can be covered by charts or 'coordinate' patches, as in Fig. 2a and Fig. 3a. The charts together are said to furnish an atlas for the manifold. The terminology is borrowed from cartography, where an atlas consists of several overlapping charts which can, for instance, together describe a continent.

The idea is to use existing notions on smooth functions, vector fields, differentials and Cartesian tensors on $\mathbb{R}^n$ to develop corresponding notions for the manifold via a combination of patchwork and consistency conditions between overlapping charts. It took a long time for a satisfactory definition of a manifold to be arrived at (in the work of Hermann Weyl (1912) and Hassler Whitney (1930s)), with examples playing a key role. Here, we will introduce a number of concepts and technical terms from the theory of manifolds. The reader who is meeting these for the first time should not despair, as they are invariably accompanied by illustrative examples.

2.2.1 Analogy with cell phone networks and cartography

The idea of covering a space with overlapping patches of a simple sort is practically realized in cell phone networks, which we caricature now. For instance, a city

[20] In Section 2.13 we will extend the concept of a manifold to one with a boundary. The points on the boundary will not have open neighborhoods homeomorphic to $\mathbb{R}^n$ and need to be treated differently.

[21] The open neighborhoods we have in mind are simple ones: they must come in one piece and be contractible (shrunk continuously to a point). Examples: In 1d they are continuously deformed (stretched/bent) versions of open intervals (a, b) on the real line. In 2d and 3d they are continuous deformations of the open disk $x^2 + y^2 < 1$ and open ball $x^2 + y^2 + z^2 < 1$. The open interval, disk and ball is each continuously deformable into $\mathbb{R}, \mathbb{R}^2$ and $\mathbb{R}^3$. Similarly, we have open balls in higher dimensions. Our neighborhoods look like them. By contrast, a pair of disjoint intervals on the real line does not come in one piece and may be shrunk to a pair of points but not to a single point. Similarly, an annulus $1 < x^2 + y^2 < 2$ can be shrunk to a circle but not to a point.

[22] By 'looks like', we mean *continuously deformable into*. A rubber balloon undergoes continuous deformation as it is inflated. More precisely, by 'looks like', we mean *homeomorphic to*. A homeomorphism is a continuous map with a continuous inverse. An untied balloon is homeomorphic to a disc-shaped rubber sheet since the latter can be stretched into a balloon without tearing the rubber sheet.

is covered by cells (say disks), each serviced by a cell phone tower. Each point in the city lies in at least one such cell and communication to/from the cell phone is transmitted via some protocols associated to the corresponding tower (manner of storage, encryption, etc.). If a phone lies in the intersection of two cells, then two towers can simultaneously communicate with it. The data received by the two towers can be related to each other via a suitable transformation between the protocols followed by each tower. This is crucial when a person is traveling in the back seat of a car and speaking on a cell phone. When moving from one cell to another, the two towers must agree on what the person is saying when the phone is in the intersection, before the 'future' tower takes over from the 'past' tower. Evidently, the city is our manifold and the cells are our coordinate patches. The transformations between data received by two towers from the overlap of two cells play the role of transition functions that we will soon encounter. A similar analogy, which explains much of the terminology, may be made with cartography, where the charts or maps prepared by two explorers have to be related (e.g., the scales of magnification may be different) in regions they both explore. □

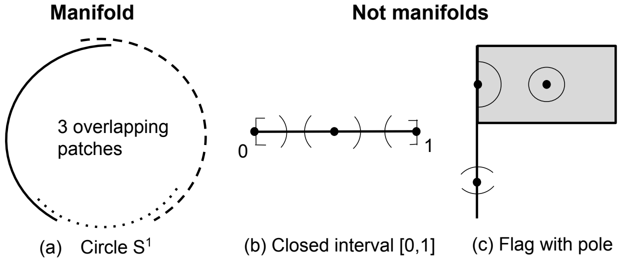


Figure 2: (a) The circle manifold S^1 covered here by three overlapping open neighborhoods. (b) The closed interval $[0, 1]$ is *not* a manifold as 0 and 1 do not have open neighborhoods that look like $\mathbb{R}$: it is a manifold with boundary. (c) A flag with pole is not a manifold: open neighborhoods do not all have the same dimension (1 on the pole and 2 on the flag) and the neighborhoods of all points do not look alike.

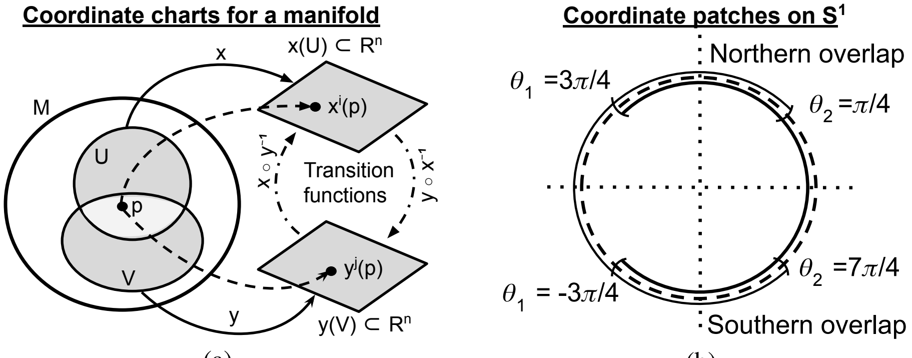
2.2.2 Coordinate charts and transition functions

Returning to our definition of a manifold, why do we insist on open neighborhoods of the same dimension to cover a manifold? Examples of spaces we do not want to regard as manifolds will reveal why. For instance, the closed interval $C = [0, 1] \subset \mathbb{R}^1$ is *not* a manifold[23]. The points 0 and 1 do not have any open neighborhoods[24] lying in C while all other points $0 < x < 1$ have open neighborhoods (see Fig. 2) of the form $(x - \epsilon, x + \epsilon) \subset C$ for some sufficiently small ϵ [we could take ϵ to be the smaller of $x/2$ and $(1 - x)/2$]. Intuitively, C looks different in the vicinity of 0 and 1 from how it looks elsewhere. We do not want to allow such 'inhomogeneities' in a manifold. This is why we insist on open neighborhoods. Similarly, the space that is shaped like

[23] However, $[0, 1]$ may be viewed as the manifold $(0, 1)$ with boundary included, see Section 2.13.

[24] 'Closed-open' neighborhoods of 0 such as $[0, 1/2)$ are not homeomorphic to $\mathbb{R}$.

the multiplication sign × is not a manifold: it looks different at the center and the extremes compared to how it looks elsewhere. A cloth flag attached to a pole is also not a manifold: points on the lower part of the pole look different from points on the cloth: the former typically have 1d open neighborhoods while the latter typically have 2d open neighborhoods (see Fig. 2). This is why we insist that all neighborhoods have the same dimension.



(a) (b)

Figure 3: (a) Coordinate charts and transition functions for a manifold M (which can be thought of as the surface of a sphere S^2). (b) Overlapping coordinate patches for the circle S^1 (dashed curve) as a manifold. A minimum of two (open) coordinate patches is needed to cover the circle: here they are the Eastern and Western patches indicated by thick and thin solid curves. The angle θ is measured counterclockwise from the the horizontal axis.

On the other hand, the unit circle S^1 defined as the set of points (x, y) on the plane with $x^2+y^2 = 1$ is a one-dimensional manifold. As shown in Fig. 3b, the circle can be covered by two patches: the Eastern and Western neighborhoods $-3\pi/4 < \theta_1 < 3\pi/4$ and $\pi/4 < \theta_2 < 7\pi/4$ defined in terms of a polar angle measured counterclockwise with respect to the positive x-axis. θ_1 and θ_2 are called local[25] coordinates in their respective patches. Thus, the circle is one-dimensional. Each of these angular patches is continuously deformable (by stretching) into the real line $\mathbb{R}$ since every open interval can be continuously mapped to the whole real line (e.g., $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$). These patches intersect in a pair of 'upper' and 'lower' intervals: running from North-East to North-West and from South-West to South-East. When a point lies in such an intersection, either of the coordinates can be expressed in terms of the other via a 'transition function' or coordinate transformation. For the circle, we have in the upper intersection $\theta_1 = \theta_2$ and in the lower intersection $\theta_1 = \theta_2 - 2\pi$. The manifold is differentiable if these transition functions between coordinate systems in each such intersection is a differentiable map from $\mathbb{R}^n \to \mathbb{R}^n$ (or between open subsets of $\mathbb{R}^n$ which are homeomorphic to $\mathbb{R}^n$). In the circle example, the transition functions are linear maps of one real variable, so the circle is a differentiable manifold of dimension one. If the transition functions are infinitely differentiable (as is the case here), we say

[25] *Local* means coordinates are defined on a patch rather than *globally* on the whole manifold.

the manifold is smooth[26] or C^∞. Note that the circle cannot be covered by a single open coordinate chart: the largest ones such as $-\pi < \theta_1 < \pi$ unfortunately exclude one point while $-\pi < \theta_1 \leq \pi$ or $0 \leq \theta_1 \leq 2\pi$ fail to be open subsets of $\mathbb{R}$.

Sometimes we are lucky, and a single coordinate patch is sufficient to cover the whole manifold or the portion we are interested in. This is the case with the plane or a disk ($x^2 + y^2 < 1$) or 3d Euclidean space $\mathbb{R}^3$, which can be covered by a single patch with, say, Cartesian coordinates. In particular, $\mathbb{R}^n$ for each[27] $n = 1, 2, 3, \ldots$ is automatically a smooth manifold, as are all the open subsets of $\mathbb{R}^n$ that may be continuously shrunk to a single point. Unfortunately, the circle, which is a 1d manifold, is not an open subset of $\mathbb{R}^1$ and S^2 is not an open subset of $\mathbb{R}^2$, so we cannot cover them with a single chart[28] and need to work harder to find an atlas for these manifolds. The circle or the 2-sphere S^2 require a minimum of two coordinate patches. For S^2, the patches (each continuously deformable into a disk) consisting of all latitudes strictly above the Tropic of Capricorn and all latitudes strictly below the Tropic of Cancer furnish one possible atlas. These patches intersect over the tropics.

Given a manifold M, suppose a point $p \in M$ lies in the intersection of two coordinate patches so that p may be assigned the coordinates $x = (x^1, \cdots, x^n)$ or $y = (y^1, \cdots y^n)$. Then the 'transition function' from x to y is given by the equations for the coordinate transformation $y^i = y^i(x)$ and conversely $x^j = x^j(y)$ for $1 \leq i, j \leq n$. For the manifold to be smooth, both the transformation $x \mapsto y$ and its inverse $y \mapsto x$ must be smooth[29] maps between open subsets of $\mathbb{R}^n$ (see Fig. 3a).

• **Example: Inertial and noninertial coordinates for Newtonian spacetime.** Newtonian spacetime may be viewed as a manifold. For simplicity, suppose space is one dimensional, then the spacetime manifold is simply $\mathbb{R}^2$. A coordinate patch on spacetime is sometimes called a coordinate frame. Newton's first law says that spacetime admits a special frame called an inertial frame where isolated bodies move at constant velocity. We usually use the globally defined Cartesian coordinates (t, x) for such an inertial frame, which covers all of $\mathbb{R}^2$ with a single patch. Other coordinate frames are also of interest. Galileo found that inertial frames are not unique, for instance a uniformly moving frame (relative to one guaranteed by Newtons 1st law) is also inertial. The coordinate transformation to such a boosted frame is given by $t' - t, x' = x - ct$ where c is the relative velocity of the frame. Since these coordinate transition formulae are linear (and hence invertible and infinitely differentiable both ways), Newtonian space-time is a smooth manifold. Noninertial coordinate frames are

[26] If the transition functions are continuous, we call it a C^0 manifold or a topological manifold. If they are once differentiable with a continuous derivative, we call it a C^1 manifold. More generally, we have the notion of a C^k manifold if the transition functions are continuously differentiable k times for some $k = 0, 1, 2, \ldots$.

[27] For some purposes, it is convenient to regard $\mathbb{R}^0$ as a zero-dimensional manifold with only one point. A zero-dimensional manifold is either a point or a discrete set of points. For example, the zero-dimensional sphere S^0 is the pair of points $\{-1, 1\}$ satisfying $x^2 = 1$ in $\mathbb{R}^1$.

[28] If we could cover S^1 with a single chart, the chart (and hence S^1) would be an open subset of $\mathbb{R}^1$. Note, however, that merely being an open subset of $\mathbb{R}^n$ does not mean we can cover a manifold with a single chart, since our charts are assumed to be homeomorphic to open balls. For instance, the annulus $1 < x^2 + y^2 < 2$ is an open subset of $\mathbb{R}^2$, but we need a minimum of two charts to cover it.

[29] $y^i(x)$ is smooth if partial derivatives of all orders ($\partial y^i / \partial x^j, \partial^2 y^i / \partial x^j \partial x^k, \cdots$) exist.

also of interest. An example is one that is accelerating uniformly relative to an inertial frame. The coordinate transformation to such a (noninertial Cartesian) frame is given by $t' = t, x' = x - \frac{1}{2}at^2$ (assuming the origins of the two frames are the same).

2.2.3 Refining an atlas

Given a smooth manifold (which must necessarily come with an atlas of smoothly interrelated coordinate charts), we are free to add a chart to the atlas, provided we are consistent. For instance, if the new chart with coordinate y overlaps an existing chart with coordinate x, the transformation $y = y(x)$ and its inverse must be smooth. For the Euclidean plane, the Cartesian coordinates ($x^1 \in \mathbb{R}, x^2 \in \mathbb{R}$) furnish a one-chart atlas. Suppose we wish to add a chart consisting of plane polar coordinates ($y^1 = r$, $y^2 = \phi$). We can do this provided we choose the polar coordinate chart to be an open set on which the transformation to/from Cartesian coordinates is smooth. This is the case, for instance, if we choose the polar coordinate chart to be defined on $\mathbb{R}^2$ with the origin and negative horizontal axis excluded so that $y^1 = r \in (0, \infty)$ and $y^2 = \phi \in (-\pi, \pi)$. The Cartesian product $(0, \infty) \times (-\pi, \pi)$ is clearly an open subset of $\mathbb{R}^2$ continuously deformable into $\mathbb{R}^2$. Note that if we retained the negative real axis, the patch consisting of the punctured plane (plane with the origin removed) would not be continuously deformable into $\mathbb{R}^2$. In this region of overlap we have the familiar coordinate transformation $y^1 = \sqrt{(x^1)^2 + (x^2)^2}$ and $y^2 = \arctan(x^2/x^1)$ and the inverse transformation $x^1 = y^1 \cos y^2$ and $x^2 = y^1 \sin y^2$ which are both seen to be smooth. The smoothness of the transition function would fail at the origin and on the negative horizontal axis.

2.3 Maps between manifolds: homeomorphisms, diffeomorphisms

Having defined manifolds, we can now consider maps between manifolds. We will use such maps to say when two manifolds are to be considered the same. Two manifolds are topologically equivalent (or homeomorphic) if they are related by an invertible continuous map. The surface of a cube can be continuously deformed into that of a sphere, so they are homeomorphic. If two differentiable manifolds can be related via an invertible differentiable map, then they are called diffeomorphic.

To make precise the notion of a continuous, differentiable or smooth map between manifolds, we make use of the corresponding concept for maps between Euclidean spaces or open subsets thereof. So to begin with, a map $f : \mathbb{R}^p \to \mathbb{R}^q$ given by $y^i = f^i(x)$ is continuous if y^i are continuous functions of x^j. It is differentiable if all the first partials $\frac{\partial y^i}{\partial x^j}$ exist (it is continuously differentiable if these partial derivatives exist and are continuous). Finally, it is smooth if partial derivatives of all orders exist. Moving on from Euclidean spaces to manifolds, a map $\phi : M^p \to N^q$ is continuous/differentiable/smooth if in each coordinate patch, the corresponding maps between $\mathbb{R}^p$ and $\mathbb{R}^q$ are continuous/differentiable/smooth. For consistency, if coordinate patches overlap, then the individual maps should agree on the overlap.

Two manifolds M, N are said to be homeomorphic if there is a continuous bijective (one-to-one and onto) map $f : M \to N$ with continuous inverse. They are diffeomorphic if continuity is replaced with smoothness. Homeomorphic or diffeomorphic manifolds must have the same dimension and cannot be distinguished insofar as their topological/smooth structure is concerned. The circles $x^2 + y^2 = 1, x^2 + y^2 = 2$ and the ellipse $x^2 + 2y^2 = 1$ are all diffeomorphic (see Prob. **??**) as are the sphere $x^2+y^2+z^2 = 1$ and the ellipsoid $x^2+y^2+2z^2 = 1$ or the open interval $(0, 1)$ and the real line $\mathbb{R}$ (see Prob. **??**). On the other hand, the surface of a cube is not smooth (due to the sharp edges and corners), so it is not diffeomorphic to the sphere, though the two are homeomorphic. Similarly, the surface of a sphere and that of a torus (inflated tyre tube or vadai) are not homeomorphic: one can show that there is no continuous bijection between them since the latter has a 'hole/handle' which the former lacks.

The concept of a manifold that we have defined does not possess any notion of distances between points or lengths of tangent vectors or angles between tangent vectors. To define these 'geometric' concepts we need additional structure on the manifold, such as a metric (see Section 2.9). At present, our manifolds are either topological manifolds (if the transition functions are continuous) or differentiable/smooth manifolds (if the transition functions are differentiable/smooth). Thus, our manifolds currently lack any geometric rigidity of shape or size. In particular, the surface of a triaxial ellipsoid ($x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ with no two among a^2, b^2, c^2 equal) and that of a round sphere ($x^2 + y^2 + z^2 = 1$) are identical as topological or smooth manifolds since they can be continuously or smoothly deformed into each other.

2.4 Submanifolds: immersions and embeddings

It is tempting to think of a submanifold as a subset of a manifold M that acquires a manifold structure when charts of M are suitably restricted. However, this is a little too restrictive for some purposes. While the unit circle $x^2+y^2 = 1$ is a submanifold of the Euclidean x-y plane $\mathbb{R}^2$ in this sense, we would also like to admit the cubic curve and 3-petaled rose shown in Fig. 4 (c) and (d) (but not the cardioid (a) and cycloid (b)) as suitable submanifolds of $\mathbb{R}^2$. On the face of it, these curves are not manifolds due to the self-intersections. However, there is a simple way to view them as images of bona fide manifolds sitting inside ('included in') the plane. Without attempting to be very precise, we outline a framework for the idea of a submanifold. Given an n-dimensional manifold M, an s-dimensional manifold S ($s \leq n$) and an 'inclusion' map $i : S \hookrightarrow M$, we can specify what we mean by immersed and embedded submanifolds. For example, S could be $\mathbb{R}$, thought of as an infinitely long (or open stretch of) rope and M could be the plane $\mathbb{R}^2$. The inclusion map is some way of laying the rope on the plane. The question is one of whether S sits inside M in a sufficiently nice way. For example, we readily admit the x-axis contained in $\mathbb{R}^2$ and the interval $(0, 1) \subset \mathbb{R}$ as submanifolds. Questions arise, for instance, when the image of S in M involves sharp corners (as in the curve that looks like the character V or the cardioid and cycloid of Fig. 4 (a) and (b)) or self-intersections (as in the curve that looks like α or the 3-petaled rose of Fig. 4(d)). Very roughly, if the tangent to S behaves nicely (sharp corners are absent), we will say that S is an immersed submanifold while we will call it an embedding if it has neither sharp corners nor self-intersections. The manner in which S sits inside M can be encoded in properties of the inclusion map $i : S \hookrightarrow M$

which takes any point $x \in S$ to the corresponding point $i(x) \in M$. If the derivative of the inclusion map, which is the $n \times s$ Jacobian matrix of first partials, has the maximum possible rank[30] s everywhere, then S is said to be an immersed submanifold (this eliminates sharp corners but allows for self-intersections, as in the symbol α or the planar cubic curve $y^2 = x^3 + x^2$ of Fig. 4(c)). Thus, in an immersion, the inclusion map need not be 1-1, though its derivative must be 1-1. An immersion where the inclusion map is also 1-1 (this eliminates self-intersections) is called an embedding. The n-sphere $S^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} \text{ such that } (x^1)^2 + (x^2)^2 + \cdots + (x^{n+1})^2 = 1\}$ is an embedded submanifold of $\mathbb{R}^{n+1}$ for $n = 0, 1, 2, \ldots$. An important theorem of Whitney states that essentially any smooth n-dimensional manifold M (defined as above using an atlas) can be realized as a smoothly embedded submanifold[31] of $\mathbb{R}^{2n}$.

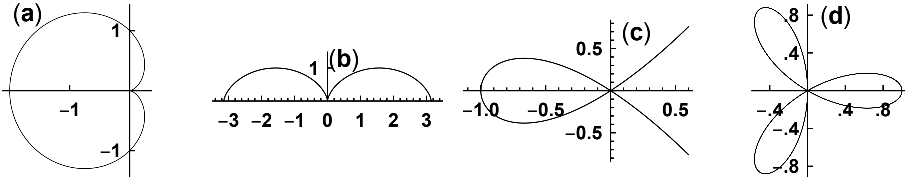


Figure 4: Plane curves (a) cardioid $x = \cos t(1 - \cos t), y = \sin t(1 - \cos t)$ (b) cycloid $x = (t - \sin t)/2, y = (1 - \cos t)/2$ (c) cubic $y^2 = x^3 + x^2$ and (d) 3-petaled rose $r = \cos 3\theta$ given in parametric, implicit and polar forms. The cardioid and cycloid fail to be immersed submanifolds of the plane since the rank of their Jacobians drop from 1 to 0 at the sharp corners. E.g., for the cycloid, the transpose of the Jacobian is $J^t = (\dot{x}, \dot{y}) = ((1 - \cos t)/2, (\sin t)/2) = (0, 0)$ at $t = 2n\pi$ where n is an integer. The cubic curve and the rose are immersions but not embeddings: they have no sharp corners but display self-intersections. If we think of the image curve as the path traced by an ant walking on the plane, the Jacobian is the velocity vector. If the ant does not momentarily come to rest, its path may be modeled as an immersion (self-intersections occur when an ant returns to an earlier location while at a sharp corner, the ant must momentarily come to rest and abruptly change direction). If the curve is the world line of a massive particle in space-time parametrized by proper time, then it must be an embedding since the 4-velocity (**??**) cannot vanish and the world line cannot have self-intersections.

• The circle $(x^1)^2 + (x^2)^2 = 1$ is a smooth embedding of S^1 in $\mathbb{R}^2$. Note that here we take $S = S^1$ (and not $S = \mathbb{R}$) and furnish it with an atlas with at least two charts. If we lay the real line in the shape of a circle on the plane, then we would not get an embedding (although it would still be an immersion) since the inclusion map would be many-to-one.

• The 3-petaled rose is an immersion of S^1 in the plane. It is not an embedding. It may also be viewed as an immersion of $\mathbb{R}$ in the plane.

[30] The row (column) rank of a matrix is the number of linearly independent rows (columns). They can be shown to be equal and this common number is called the rank of the matrix.

[31] In favorable cases, one may be able to embed the n-dimensional manifold M in a Euclidean space of dimension less than $2n$, as is the case with $S^n \hookrightarrow \mathbb{R}^{n+1}$.

2.5 Connected and simply connected manifolds

A manifold is connected if it comes in one piece. For example, the disjoint union of two open real intervals $(0,1) \cup (2,3)$ is not connected. To define the concept of connectedness, we imagine a point-like ant walking on the manifold M. If it can reach any point from any other point via a continuous path $\gamma(t)$ that lies in M, then M is connected. More precisely, M is path connected if any two points $p, q \in M$ can be joined by a continuous path $\gamma : [0,1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$. For example, $\mathbb{R}^n$ and S^n for $n = 1, 2, 3, \ldots$ are connected. If a manifold is not path connected, then it is called disconnected. The real line punctured[32] at the origin is a disconnected manifold. For a disconnected manifold, the connected component of $p \in M$ is the submanifold consisting of points $q \in M$ that can be reached from p via continuous paths lying in M. The line punctured at the origin has two connected components $(-\infty, 0)$ and $(0, \infty)$. The connected component of the point 2 is the right half-line. Similarly, $S^0 = \{-1, 1\}$ is disconnected, it has two connected components: $\{-1\}$ and $\{1\}$.

The orthogonal group $O(3)$ (3×3 real matrices with $A^t A = A A^t = I$) is disconnected, it has two connected components. The connected component in which the identity matrix lies is the subgroup $SO(3)$ where the determinant is always equal to one. The other component (where the determinant is minus one) consists of orthogonal transformations which involve reflections in an essential way.

A connected manifold M is called simply connected if any nontrivial closed curve in M can be continuously deformed (shrunk) to a point while remaining in M. The two-sphere S^2 is simply connected, a rubber band on a globe can always be shrunk to a point while remaining on the globe. On the other hand, S^1, the torus, the surface of an infinite circular cylinder and the punctured plane are all connected but not simply connected (see Fig 5).

2.6 Smooth functions or scalar fields

In essence, a smooth real function (or a 'scalar field') f on a smooth manifold M is a way of assigning a smoothly varying real number to each point on the manifold. They are important since the observables or dynamical variables of a mechanical system are smooth functions on the phase space manifold. More precisely, scalar fields are smooth real-valued functions of the coordinates in any given patch with the consistency condition that, when $p \in M$ lies in the intersection of coordinate patches, the value of the function at p must be the same irrespective of which coordinate system is used to describe p. In other words, if the function is described by the formulae $f(x)$ and $g(y)$ in two coordinate patches, then we must have $g(y(x)) = f(x)$ at each point p of the overlap. Sometimes, we turn this around and say that given a scalar field $f(x)$ in one coordinate system, under a change from $x \mapsto y$ given by the transformation $y = y(x)$ (and its inverse $x = x(y)$), the formula for the function becomes $F(y) = f(x(y))$. Henceforth, we will mostly take this second viewpoint and speak in terms of how objects transform under a change of coordinates

[32] To puncture the line $\mathbb{R}$ is to remove (or excise) one point from it.

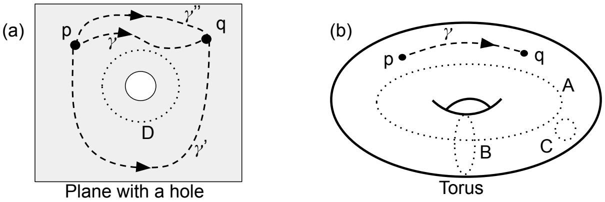


Figure 5: (a) The plane with a hole (disk excised) is homeomorphic to the once punctured plane. Any two points p, q can be joined by a continuous path γ: it is path connected. However, it is not simply connected: the closed curve D cannot be continuously shrunk to a point. It is *multiply connected*: there are many paths from p to q that are not continuously deformable into each other: direct (γ), around the hole (γ'), winding twice around the hole, etc. Though γ is *not* homotopic to γ' (they cannot be continuously deformed into each other), γ and γ'' are homotopic to each other (a rubber band stretched from p to q along γ can be deformed to γ''). A homotopy between $\gamma, \gamma'' : [a,b] \to M$ is a continuous map $\Gamma : [a,b] \times [0,1] \to M$ with $\Gamma(t;0) = \gamma(t)$ and $\Gamma(t;1) = \gamma''(t)$. (b) The torus too is path connected but not simply connected: the closed curves A, B winding around the torus cannot be shrunk to a point, though C can. C is a *contractible* closed curve or one that is *homotopic* to a point.

in a region of overlap between two coordinate patches. The space of smooth functions on M is denoted $C^\infty(M)$ or $\mathcal{F}(M)$. If M is the phase space of a classical system, then the set of observables is given by $\mathcal{F}(M)$. It is a commutative algebra: it is closed under real linear combinations $af + bg \in \mathcal{F}$ and pointwise products $(fg)(x) = f(x)g(x) = g(x)f(x) = (gf)(x)$ for all $f, g \in \mathcal{F}(M)$, with products distributing over sums. The property $fg = gf$ encodes the fact that classical observables commute under multiplication, a feature that is not always true in the quantum theory, where observables are Hermitian operators.

2.7 Vector fields

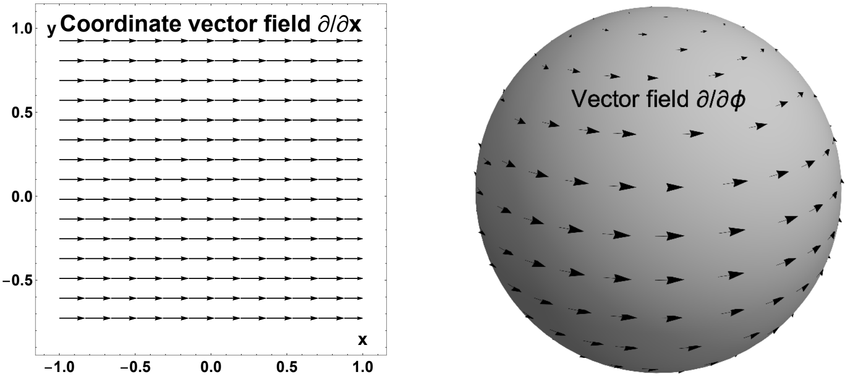
Given local coordinates x^i in a chart on an n-dimensional manifold M, we have the notion of coordinate vector fields. These are defined as the first order partial differential operators $\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \cdots, \frac{\partial}{\partial x^n}$, which are often abbreviated ∂_{x^i} or ∂_i for $i = 1, \cdots, n$. We may think of these differential operators as acting on smooth functions to produce other smooth functions (their partial derivatives).

Geometrically, we may think of the coordinate vector fields at a point p as tangent vectors to M at p. For instance, ∂_1 is the tangent vector to the coordinate curve parametrized by x^1 passing through p holding $x^2, \cdots, x^n$ fixed. As a consequence, ∂_x is a tangent vector field on the $x-y$ plane $\mathbb{R}^2$ that points rightward at every point, as shown in Fig. 6a. More generally, we may view any vector at p as the velocity vector of some smooth (or at least differentiable) curve passing through p. For instance if $(x(t), y(t))$ is a curve on the plane $\mathbb{R}^2$, then its velocity vector at the point on the curve corresponding to parameter t is $(\dot{x}(t), \dot{y}(t))$. In the language introduced above, this velocity vector is written as $\dot{x}(t)\frac{\partial}{\partial x} + \dot{y}(t)\frac{\partial}{\partial y}$. Tangent vectors at p may thus be regarded as equivalence classes of curves passing through p. For this purpose, two curves are considered equivalent if they possess the same velocity vector at p.

Coordinate vector fields furnish a basis for more general vector fields on M. A general vector field is given by a linear combination

$$v = \sum_{i=1}^{n} v^i(x) \frac{\partial}{\partial x^i} \equiv v^i(x)\partial_i. \tag{128}$$

The set of n functions $v^i(x)$ are called the components of v in the coordinate basis. A vector field restricted to a point $p \in M$ is called a tangent vector at p. The set of tangent vectors at p is the tangent space $T_p(M)$, a real vector space of dimension n. The coordinate tangent vectors $\partial_1, \cdots, \partial_n$ at p furnish a basis[33] for $T_p(M)$. E.g., the tangent space to the 2-sphere at a point on the equator may be visualized as a vertical tangent plane spanned by the coordinate tangent vectors $\frac{\partial}{\partial\phi}$ and $\frac{\partial}{\partial\theta}$ (see Fig. 6b).



(a) (b)

Figure 6: (a) Coordinate vector field ∂_x on the x-y plane. (b) Azimuthal coordinate vector field $\partial_\phi = -y\partial_x + x\partial_y$ on the unit sphere with the z-axis pointing vertically upwards. At the North and South poles $x = y = 0$ and $z = \pm 1$. At the poles, ∂_ϕ vanishes, they are zeros of ∂_ϕ. In fact, there is no nonvanishing smooth vector field on a sphere: loosely speaking, it is not possible to comb hair on the sphere. Here, a smooth distribution of hair combed tangent to a sphere may be regarded as a vector field on the sphere. The vector field has a zero at a bald spot where there is no hair.

Transformation of vector fields on the overlap between two patches. The set of n functions $v^i(x)$ are called the components of v in the coordinate basis. Though each is a function within a coordinate patch, they do not transform as scalar functions under a change of coordinates. The components of a vector field have a special transformation law that follows from the chain rule in multivariable calculus. Suppose the

[33] A *noncoordinate basis* for vector fields is a collection of n vector fields $e_1, \cdots, e_n$ that are linearly independent at each point but whose pairwise commutators $[e_i, e_j]$ are not all zero. The latter condition ensures that they are not expressible as ∂_{x^i} in any local coordinate system x^i. For example, $e_1 = y\partial_x$ and $e_2 = x\partial_y$ furnish a noncoordinate basis for vector fields, say, in the first quadrant of the Euclidean plane ($x > 0, y > 0$). We find that $[e_1, e_2] = y\partial_y - x\partial_x = (y/x)e_2 - (x/y)e_1 \neq 0$.

same vector field is expressed in another coordinate system y^i:

$$v = \tilde{v}^j(y) \frac{\partial}{\partial y^j}. \tag{129}$$

Since $y = y(x)$, we may relate the two sets of coordinate vector fields via a Jacobian:

$$\frac{\partial}{\partial x^i} = \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j} = J^j_i \frac{\partial}{\partial y^j} \quad \text{so that} \quad v = v^i \frac{\partial}{\partial x^i} = v^i \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j}. \tag{130}$$

Comparing with (129) we find how the components of a vector field transform:

$$\tilde{v}^j(y) = v^i(x(y)) \frac{\partial y^j}{\partial x^i} \quad \text{or} \quad \tilde{v}^j = J^j_i v^i. \tag{131}$$

Thus, the components of a vector field transform via the Jacobian matrix[34]: the new j^{th} component is a linear combination of all the old components (quite unlike how n scalar fields would transform). Such a transformation is called *contravariant*. The prefix *contra*[35] arises from the manner in which the coordinate vector fields transform, i.e., via the inverse of the Jacobian matrix[36]:

$$\frac{\partial}{\partial y^j} = \frac{\partial x^i}{\partial y^j} \frac{\partial}{\partial x^i} = (J^{-1})^i_j \frac{\partial}{\partial x^i}. \tag{132}$$

Thus, tangent vector fields are also called contravariant vector fields. A vector field on a smooth manifold is called smooth if the components $v^i(x)$ are smooth functions of the coordinates in each patch and components in overlapping patches are related by the above transformation formula. The matrix elements J^j_i entering the transformation formula for v^i between overlapping coordinate patches are automatically smooth since the manifold is smooth.

• Example: On $\mathbb{R}^2$ we have a single patch Cartesian coordinate system (x, y). We also have plane polar coordinates (r, ϕ) consisting of radial and azimuthal coordinates on a suitable open set (say on the complement of the negative horizontal axis $x \le 0, y = 0$). Then the azimuthal coordinate vector field ∂_ϕ can be expressed in the Cartesian basis. Using $x = r\cos\phi, y = r\sin\phi$, we find

$$\frac{\partial}{\partial \phi} = \frac{\partial x}{\partial \phi} \frac{\partial}{\partial x} + \frac{\partial y}{\partial \phi} \frac{\partial}{\partial y} = -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y} \tag{133}$$

on the overlap of the two coordinate patches. The RHS makes sense on all of $\mathbb{R}^2$ and allows us to extend the azimuthal coordinate vector field to all of $\mathbb{R}^2$. We see that this vector field has a zero at the origin.

[34] We may view this as the product of a matrix with a column vector by regarding J^j_i as the entry in the j^{th} row and i^{th} column of a square matrix J and v^i as the element in the i^{th} row of a column vector.

[35] In Section 2.8 we will meet covector fields. Coordinate covector fields (137) transform via J rather than J^{-1}, so covector fields are called *covariant*.

[36] The Jacobians for $x \mapsto y$ and $y \mapsto x$ are inverse matrices. This is seen by using the chain rule to differentiate $x^i(y(x)) = x^i$ with respect to x^k to get $\frac{\partial x^i}{\partial y^j} \frac{\partial y^j}{\partial x^k} = \delta^i_k$.

Action of vector fields on smooth functions. A vector field v can act on a function on M and give its derivative 'along' v. In a coordinate chart,

$$v(f) = v^i \frac{\partial f}{\partial x^i}. \tag{134}$$

$v(f)$ is a function on M and generalizes the concept of the directional derivative $\boldsymbol{v} \cdot \boldsymbol{\nabla} f$ from vector calculus in $\mathbb{R}^3$. $v(f)$ is also called the Lie derivative of f along v and denoted $\mathcal{L}_v f$. To find $\mathcal{L}_v f$ at a point $x_0 \in M$, we consider the integral curve $x(t)$ (136) of v through x_0 with $x(0) = x_0$. Then $\mathcal{L}_v f = \lim_{s \to 0}[\{f(x(s)) - f(x(0))\}/s]$. In other words, we ask how the function varies between nearby points on the integral curve of v through x_0.

Evidently, a vector field acts linearly on the space of functions: $v(af + bg) = av(f) + bv(g)$ for any pair of scalar fields f, g and real numbers a, b. Since a vector field is a first order differential operator, v acts as a derivation on the space of functions: verify that the Leibniz rule $v(fg) = fv(g) + v(f)g$ is satisfied. Thus, we may view a vector field simply as a linear map from $\mathcal{F}(M) \to \mathcal{F}(M)$ that satisfies the Leibniz rule. The set of vector fields on M is denoted Vect(M), it is an infinite-dimensional real vector space since the coefficients v^i can be arbitrary smooth functions in any given patch. For instance, if $M = \mathbb{R}$ with coordinate x, then the tangent space at any point is one dimensional (and isomorphic to the vector space $\mathbb{R}$). However, the vector fields ∂_x and $x\partial_x$ are linearly independent over the real numbers: there is no nontrivial real linear combination that vanishes identically. In fact, polynomial vector fields (of degree d) on the real line can be written as $\sum_{l=0}^{d} c_l x^l \partial_x$ where $c_0, c_1, \ldots, c_d$ are suitable real coefficients and x^l denotes the l^{th} power of x. Evidently, the space of polynomial vector fields on the real line is infinite dimensional: the whole number d can be arbitrarily large.

Commutator of vector fields. Given a pair of differentiable vector fields on a manifold M, we may define their commutator, which is another vector field. In local coordinates, suppose $u = u^i \partial_i$ and $v = v^i \partial_i$. Then their commutator $[u, v]$ is

$$[u, v] = (u^j \partial_j v^i - v^j \partial_j u^i)\partial_i. \tag{135}$$

Given a function $f : M \to \mathbb{R}$, both $u(v(f))$ and $v(u(f))$ are functions on M. The commutator $[u, v]f$ measures the extent to which the two differ. Notably, $u(v(f))$ and $v(u(f))$ involve both first and second order derivatives, so the composition of vector fields is not a vector field. Pleasantly, we verify that these second order derivatives cancel out in the commutator.

It is easily checked that coordinate vector fields commute since mixed partials of a smooth function are equal: $[\partial_{x^i}, \partial_{x^j}]f = 0$ for any smooth function f and any $1 \leq i, j \leq n$. For example $[\partial_x, \partial_y] = 0$ on $\mathbb{R}^2$.

By making a change of coordinates, one may check that this first order differential operator in (135) transforms as a contravariant vector field.

The commutator is also called the Lie bracket of vector fields since it is linear[37] $[au + bv, w] = a[u, w] + b[v, w]$, antisymmetric $[u, v] + [v, u] = 0$ and satisfies the

[37] Note that $[fu, v] \neq f[u, v]$ in general for a nonconstant smooth function f, see Prob. **??**.

Jacobi identity $[u,[v,w]] + [w,[u,v]] + [v,[w,u]] = 0$ (Prob. **??**) for any three vector fields u, v and w and real numbers a, b. Consequently, the linear space of vector fields $\mathrm{Vect}(M)$ equipped with the commutator Lie bracket is called a real Lie algebra[38].

The commutator $[u, v]$ is also called the Lie derivative of v along u and is written $\mathcal{L}_u v = [u, v]$. From (135), we see that the Lie derivative[39] of v along u includes two contributions: the first is the 'obvious' change in v in the direction of u while the second accounts for the fact that the components of u themselves change with location. It is noteworthy that if we omitted this second term, the first term by itself would not transform as a vector field under a coordinate transformation.

Finally, we note that given a smooth function f, the Lie derivative of a vector field satisfies the Leibniz rule: $\mathcal{L}_u(fv) = (\mathcal{L}_u f)v + f\mathcal{L}_u v$, as we verify in Prob. **??**.

Integral curves of a vector field. Given a vector field v on a manifold, it defines a flow on the manifold. By this, we mean that there is a family of curves on M that are everywhere tangent to v. Precisely, the integral curve through the point $x_0 \in M$ is the solution $x^i(t)$ to the system of first order ODEs

$$\frac{dx^i}{dt} = v^i(x) \quad \text{with} \quad x^i(0) = x_0. \tag{136}$$

Here, x^i are coordinates in a neighborhood of x_0. This system is generally nonlinear since the components $v^i(x)$ can depend on x in a nonlinear fashion. It can be shown that if v is a C^1 vector field (continuously differentiable), then the solution of the above system of ODEs exists and is unique for some time interval. This means that through each point of M there is precisely one integral curve of v. The set of integral curves of a (nonvanishing) vector field on a manifold is sometimes called a congruence (or a congruence of curves).

$*$ **Gradient of a function.** The gradient of a scalar function f is an example of a vector field. However, to define it, we need additional structure on the manifold. For instance, we may combine the notion of the differential of a function and the inverse of a metric tensor to define the gradient $(\text{grad}\, f)^i = (\boldsymbol{\nabla} f)^i = g^{ij}\partial_j f$. These concepts will be introduced in Section 2.8 and Section 2.9. □

• **Tangent bundle.** As a set, the union of all tangent spaces on an n-dimensional manifold M is called the tangent bundle TM. It is a 'fibre bundle' whose 'base space' is M and whose fibres are the tangent spaces. It is a manifold of dimension $2n$. Locally it is a Cartesian product of a coordinate neighbohood and $\mathbb{R}^n$ (the tangent spaces are isomorphic to $\mathbb{R}^n$). For each coordinate neighborhood of M with coordinates x^i we get coordinates on T^*M given by $(x^1, \cdots, x^n, v^1, v^2, \cdots, v^n)$. Here $v = v^j\partial_j$ are tangent vectors at the point with coordinates x^i. If M is the configuration space of a

[38] The Lie algebra of vector fields may be regarded as the Lie algebra of the group of diffeomorphisms of M.

[39] It is tempting to mimic the geometric approach to the Lie derivative of a function to define $\mathcal{L}_u v$ as the $s \to 0$ limit of a difference quotient $\{v(x(s)) - v(x(0))\}/s$, where $x(t)$ is the integral curve (136) of u through x_0 with $x(0) = x_0$. However, there is a difficulty since $v(x(s))$ and $v(x(0))$ live in different tangent spaces and cannot be subtracted. One needs a way to 'push' one of the vectors to the tangent space where the other lives before subtracting. This can be done using the pushforward defined in Section 2.11.

mechanical system, then the initial conditions for Newton's or Lagrange's equations $(x^i, \dot{x}^i)$ define a point in the tangent bundle. The tangent bundle of the circle S^1 is diffeomorphic to a cylinder $S^1 \times \mathbb{R}$.

2.8 Covector fields or 1-forms

Motivating examples. On the Euclidean plane, the differentials dx and dy are examples of covector fields or 1-forms. They are to be thought of as dual to the coordinate vector fields $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ via the 'pairing' $dx(\partial_x) = 1, dx(\partial_y) = 0, dy(\partial_x) = 0$ and $dy(\partial_y) = 1$ which is defined to be linear: for instance, $dx(f\partial_x + g\partial_y) = fdx(\partial_x) + gdx(\partial_y) = f(x,y)$ for any two smooth functions f and g. A general covector field is a linear combination $\phi = a(x,y)dx + b(x,y)dy$ where a and b are smooth functions. A 1-form is also called a Pfaffian differential expression after the German mathematician J F Pfaff who studied equations[40] of the form $a(x,y,z)dx + b(x,y,z)dy + c(x,y,z)dz = 0$. Physically, for a particle moving on a plane, while the velocity $\dot{q}(t) = \dot{q}^1(t)\partial_x + \dot{q}^2(t)\partial_y$ is a tangent vector at each point $(x(t), y(t))$ on a trajectory, the momentum $p(t) = p_1 dx + p_2 dy$ is a covector at each such point on the configuration plane.

Covector fields are also encountered on the thermodynamic state space. The thermodynamic state space of a gas with a fixed number of molecules is a 3d manifold M with coordinates U, V, S which are the internal energy, volume and entropy of the state of the gas (other choices of coordinates are also possible). An infinitesimal process is represented by a tangent vector $v = a\partial_U + b\partial_V + c\partial_S$. We have two distinguished 1-forms on M, the work and heat 1-forms. The work 1-form is $\omega = pdV$. The work done in this infinitesimal process is $\omega(v) = pdV(v) = pb$. According to the 1st law of thermodynamics, the heat 1-form is $\phi = dU + pdV$. The infinitesimal heat added to a gas is given by the action of the heat 1-form ϕ on the tangent vector representing the infinitesimal process. In our notation, the heat added is $a + pb$. Equilibrium states form a 2d hypersurface determined by an equation of state (EOS). Tangent vectors to this EOS surface represent infinitesimal reversible processes. When restricted to this equilibrium surface, the second law postulates that the heat 1-form may be expressed as $\phi = TdS$, where T and S are the absolute temperature and entropy. T is defined only on the equilibrium submanifold.

Dual to tangent space. More generally, a covector field or covariant vector field or 1-form is simply a (smoothly varying) assignment of a covector at each point of a manifold. In more detail, given local coordinates x^i in a patch, we have the coordinate basis 1-forms given by the differentials of the coordinates $dx^1, \cdots, dx^n$. At a point $p \in M$, the basis 1-forms $dx^i(p)$ are said to span the cotangent space to M at p. The cotangent space is denoted $T_p^*(M)$ and is the vector space dual to the tangent space $T_p(M)$. Indeed, $\{dx^i\}$ is the dual basis to $\{\partial_i\}$ defined via the pairing $dx^i(\partial_j) = \delta^i_j$. In general, a covector field on M is a linear combination of the basis 1-forms

[40] The Pfaffian differential equation $\phi = a\,dx + b\,dy + c\,dz = 0$ is said to be integrable if it admits an integrating denominator $T(x,y,z)$ (or integrating factor $1/T$) such that $\phi/T = dS$ is an exact differential. Then $dS = 0$ and the solutions of the Pfaffian differential equation are given by $S(x,y,z) = \sigma$ for any constant σ.

$\phi = \phi_i(x) dx^i$. The n real-valued quantities $\phi_i(x)$ in a coordinate patch are called the components of the covector field. As with the components of a vector field, they are not scalar functions on M but satisfy a special transformation law. Indeed, suppose y^j is another local coordinate system defined on a chart that has an overlap with that of the x^i. On the overlap, the coordinate 1-forms are related by the chain rule

$$dy^j = \frac{\partial y^j}{\partial x^i} dx^i = J^j_i dx^i. \tag{137}$$

We see that coordinate 1-forms transform via the Jacobian matrix (as opposed to its inverse, as was the case for coordinate vector fields in (132)). For this reason, covector fields are called covariant vector fields. Now suppose the same covector field ϕ is expressed in the y basis: $\phi = \tilde{\phi}_j(y) dy^j = \tilde{\phi}_j J^j_i dx^i$. Comparing, we see that the components of a covector field transform via the inverse of the Jacobian[41]:

$$\phi_i = \tilde{\phi}_j \frac{\partial y^j}{\partial x^i} \quad \text{or} \quad \tilde{\phi}_j = (J^{-1})^i_j \phi_i. \tag{138}$$

Compare this with the corresponding formula (131) for components of a vector field. If the components of ϕ in all charts are smooth functions of the local coordinates on a smooth manifold, then ϕ is called a smooth covector field. Since covectors are dual to vectors at each point of M, covector fields are linear functions on the space of vector fields. The value of $\phi = \phi_i dx^i$ on the vector field $v = v^j \partial_j$ is the smooth function or scalar field

$$\phi(v) = \phi_i dx^i (v^j \partial_j) = \phi_i v^j dx^i(\partial_j) = \phi_i v^j \delta^i_j = \phi_i(x) v^i(x). \tag{139}$$

We used linearity of the action of a covector on a vector to pull the components $v^j(x)$ out. $\phi(v)$ is called the contraction of ϕ with v. More generally, for vector fields v, w,

$$\phi(fv + gw) = f\,\phi(v) + g\,\phi(w) \quad \text{for any} \quad f, g \in \mathcal{F}(M). \tag{140}$$

The space of covector fields on M is denoted $\Omega^1(M)$ and is dual to Vect(M) over $\mathcal{F}(M)$. In particular, if ϕ and ψ are 1-forms and f, g scalar fields, then $f\phi + g\psi$ is also a 1-form. Note that $f\phi = \phi f$, the order does not matter.

On the other hand, we can also evaluate a vector field v on a 1-form ϕ to get a scalar field. In fact, since they are dual bases, we also have $\partial_i(dx^j) = \delta^j_i$ so that $v(\phi) = v^j \partial_j(\phi_i dx^i) = v^i \phi_i = \phi(v)$. This allows us to reinterpret vector fields as linear functions on the space of 1-forms. This viewpoint will soon be useful in generalizing vector fields to contravariant tensor fields.

An important class of 1-forms are **differentials of functions** on M: $\phi = df = \frac{\partial f}{\partial x^i} dx^i$. So the partial derivatives of a function should be thought of as components

[41] If we view $(J^{-1})^i_j$ as the entry in the i^{th} row and j^{th} column of a matrix and ϕ_i as the entry in the i^{th} column of a row vector, then this is the product (from the left) of a row vector with the square matrix J^{-1} producing the row vector with j^{th} column entry $\tilde{\phi}_j$.

of a covector rather than a vector[42]. Sometimes, it is convenient to regard functions as covector fields of degree zero and call $\mathcal{F}(M)$ $\Omega^0(M)$. Thus, the differential d is a linear map (over $\mathbb{R}$) from $\Omega^0(M)$ to $\Omega^1(M)$ satisfying the Leibniz rule.

Lack of a canonical isomorphism. The tangent space $T_p(M)$ and the cotangent space $T_p^*(M)$ are both n-dimensional real vector spaces and are therefore isomorphic. However, there is no preferred or canonical isomorphism between them. If a basis, such as a coordinate basis is chosen for vector fields then one gets an isomorphism that maps ∂_i to dx^i and vice versa. However, this depends on the choice of coordinates[43]. Thus, given a smooth manifold, there is no distinguished or natural way to relate vectors to covectors, there are many ways to do this, but none of them is special. The situation changes if the manifold is equipped with a metric tensor. In this case, there is a standard way (called *lowering an index*) of mapping vectors to covectors, which does not depend on the coordinates chosen, as we will see in Section 2.9.

• **Cotangent bundle.** As a set, the union of all cotangent spaces on an n-dimensional manifold M is the cotangent bundle T^*M. This is a ‘fibre bundle’ whose ‘base space’ is M and whose fibres are the cotangent spaces. It is a manifold of dimension $2n$. Locally it is a Cartesian product. For each coordinate neighborhood of M with coordinates x^i we get coordinates on T^*M given by $(x^1, \cdots, x^n, \phi_1, \phi_2, \cdots, \phi_n)$. Here $\phi = \phi_j dx^j$ is any cotangent vector at the point with coordinates x^i. If M is the configuration space of a mechanical system, then the phase space is its cotangent bundle. Hamilton's equations are equations for a vector field on this cotangent bundle.

2.9 Tensors of rank two and 2-forms

Vector fields $v = v^i \partial_i$ that we encountered in Sect. 2.7 are called contravariant tensor fields of rank one (or of type (1,0) as their components (v^i) have one upper index), while 1-forms introduced in Sect. 2.8 are called covariant tensor fields of rank one (or of type (0,1)). More generally, one may define tensors of higher rank that find varied uses in physics and mathematics. For instance, as we will see later in this section, the metric is a symmetric tensor of rank two that is used to define lengths and angles on a manifold while the symplectic form is an antisymmetric second rank tensor on phase space that arises in Hamilton's equations.

At a point $p \in M$ lying in a patch with local coordinates x^i, we may consider the tensor product of the tangent space with itself $T_p(M) \otimes T_p(M)$. This is the space of dimension n^2 with basis consisting of $\partial_i \otimes \partial_j$ for $1 \leq i, j \leq n$. A type (2,0) tensor field or second rank contravariant tensor field is then a linear combination

$$t = t^{ij}(x) \partial_i \otimes \partial_j. \tag{141}$$

[42] There is a related vector field, the gradient of f, introduced in Sect. 2.7. However, it requires an inverse metric tensor for its definition $(\text{grad } f)^i = g^{ij} \partial_j f$. The components of df and grad f are numerically equal if $g^{ij} = \delta^{ij}$, see Section 2.9.

[43] For example, on $\mathbb{R}$ with coordinate x, we have an isomorphism mapping $dx \leftrightarrow \partial_x$. If we change to a new coordinate $y = 2x$, then $dy = 2dx$ and $\partial_y = \frac{1}{2}\partial_x$. The new isomorphism between cotangent and tangent spaces $dy \leftrightarrow \partial_y$ is *different* since it takes dx to $\frac{1}{4}\partial_x$. Thus, there are many isomorphisms between the spaces of covectors and vectors, none of which can be considered coordinate-independent or standard.

Without further ado, we note that upon changing coordinates $x \mapsto y$, the components t^{ij} transform via the Jacobian matrix, just as for contravariant vector fields, except that there are now two Jacobian factors

$$t = \tilde{t}^{kl} \frac{\partial}{\partial y^k} \otimes \frac{\partial}{\partial y^l} \quad \text{where} \quad \tilde{t}^{kl} = J_i^k J_j^l t^{ij} \quad \text{and} \quad J_i^k = \frac{\partial y^k}{\partial x^i}. \tag{142}$$

Just as vector fields act linearly on 1-forms to produce functions, second rank contravariant tensors act bilinearly[44] on a pair of 1-forms to produce functions:

$$t(\phi, \psi) = t^{ij} \partial_i \otimes \partial_j (\phi_k dx^k, \psi_l dx^l) = t^{ij} \phi_k \psi_l \partial_i (dx^k) \partial_j (dx^l) = t^{ij} \phi_i \psi_j. \tag{143}$$

2.9.1 Poisson tensor

A physically important example of a (2,0) tensor is the Poisson tensor on the phase space of a mechanical system: $r = r^{ij} \partial_i \otimes \partial_j$, which has the further property of antisymmetry: $r^{ij} = -r^{ji}$. The Poisson bracket of a pair of smooth functions (observables) is the function $\{f, g\} = r(df, dg) = r^{ij} \partial_i f \partial_j g$. For a particle moving on a line, $M = \mathbb{R}^2$ with canonical coordinates $\xi = (q, p)$ and $r^{ij} = (0, 1| -1, 0)$. Given a Hamiltonian function H on phase space, the Poisson tensor allows us to define the Hamiltonian vector field V_H. It is the vector field which acts on any 1-form ϕ via $V_H(\phi) = r(\phi, dH)$. The Hamiltonian vector field defines time evolution of any observable through $\dot{f} = V_H(df)$. Trajectories on phase space are the integral curves of V_H. They are governed by the ODEs $\dot{\xi}^i = V_H^i = r^{ij} \partial_j H$. For the canonical Poisson tensor on $\mathbb{R}^2$, they reduce to Hamilton's canonical equations $\dot{\xi}^1 = \dot{q} = r^{12} \partial_2 H = \frac{\partial H}{\partial p}$ and $\dot{\xi}^2 = \dot{p} = r^{21} \partial_1 H = -\frac{\partial H}{\partial q}$. □

Similarly, we have covariant tensor fields of rank two or tensors of type (0,2):

$$t = t_{ij} dx^i \otimes dx^j, \tag{144}$$

which are linear combinations of the tensor products of the coordinate basis covector fields. Such a tensor transforms via two factors of the inverse Jacobian:

$$\tilde{t}_{kl} = (J^{-1})_k^i (J^{-1})_l^j \, t_{ij}. \tag{145}$$

In summary, each upper index on a tensor transforms via J and each lower one via J^{-1}. Covariant tensors of rank two can act on a pair or vector fields and produce a scalar function, they are bilinear maps from Vect$(M)\times$ Vect(M) to $\mathcal{F}(M)$.

2.9.2 Metric tensor

An important example of a 2nd rank covariant tensor field is the metric tensor $g = g_{ij} dx^i \otimes dx^j$, which has the further property of being symmetric $g_{ij} = g_{ji}$ and

[44] $t(\phi, \psi)$ is bilinear if it is linear in *both* the entries. For instance, $t(f\phi_1 + g\phi_2, \psi) = ft(\phi_1, \psi) + gt(\phi_2, \psi)$ for any functions f, g and 1-forms ϕ_1, ϕ_2, ψ. Bilinearity is a consequence of the definition of a dual space: vector fields are dual to 1-forms and act linearly on 1-forms (as discussed in Sect. 2.8). So pairs of vectors fields (written as a tensor product $\partial_i \otimes \partial_j$) act linearly on pairs of 1-forms (ϕ, ψ).

nondegenerate [g_{ij} an invertible matrix]. A metric allows us to generalize the concept of the dot product of vectors in Euclidean space to tangent vectors at a point p on a manifold M. The kinetic energy term $T = \frac{1}{2} m_{ij} \dot{q}^i \dot{q}^j$ of a Lagrangian quadratic in velocities defines a metric tensor on the configuration space of a mechanical system. A metric is called Riemannian if g_{ij} is a positive-definite matrix at every point on the manifold. Since the metric defines a real symmetric matrix at each point, its eigenvalues are real. Positive definiteness means the eigenvalues are strictly positive. There cannot be a zero eigenvalue since the metric is assumed nondegenerate (invertible).

• If the metric is not positive definite (or negative definite), then it must eigenvalues of both signs. Such a metric is called pseudo-Riemannian. The pair of integers (p, q) specifying the number of positive and negative eigenvalues is called the signature of the metric. An example of a pseudo-Riemannian metric is the Lorentzian metric tensor of space-time; e.g., Minkowski space in Cartesian coordinates $x^\mu = (ct, x, y, z)$ has the metric given by the constant diagonal matrix $g_{\mu\nu} = \text{diag}(1, -1, -1, -1)$ where $\mu, \nu = 0, 1, 2, 3$). It has signature $(1, 3)$.

• A virtue of an invertible metric tensor is that it defines an isomorphism from vectors to covectors: $v \mapsto v'$ where $v'_i = g_{ij} v^j$. The inverse metric with components g^{ij} maps covectors to vectors $g^{ij} v'_j = v^i$. Thus, on a Riemannian manifold, the tangent and cotangent spaces are canonically isomorphic. We say that the metric and its inverse can be used to lower and raise indices. In particular, we may use the inverse metric g^{ij} to define the *gradient of a function* (see Sect. 2.7) by raising the index of the components of the 1-form df: $(\text{grad}\, f)^i = (\boldsymbol{\nabla} f)^i = g^{ij} \partial_j f$.

• A metric tensor gives a manifold a rigid geometric shape[45]. The square of the length of the vector $v = v^i \partial_i \in T_pM$ is defined as

$$g(v, v) = g_{ij} dx^i \otimes dx^j (v^k \partial_k, v^l \partial_l) = g_{ij} v^i v^j. \tag{146}$$

Given a pair of tangent vectors $u, v \in T_pM$, their inner product is defined as $g(u, v) = g_{ij} u^i v^j$. The cosine of the angle between them is $g(u, v)/\sqrt{g(u, u) g(v, v)}$.

• It is conventional to use the symbol ds^2 for the expression for the metric tensor $g_{ij} dx^i \otimes dx^j$. Often, the tensor product symbol $\otimes$ is suppressed. Sometimes, ds^2 is called the 'square of the line element'. What this means, for instance, is that $g(\dot{x}, \dot{x}) = g_{ij} \dot{x}^i \dot{x}^j$ is the square of the length of the velocity vector $\dot{x} = \dot{x}^i \partial_i$ to a curve $x^i(t)$.

[45] The surface of a round sphere and an ellipsoid are diffeomorphic, they have the same topology. However, they differ geometrically: notions of lengths of tangent vectors and angles between them differ. When the distances between corresponding points on two diffeomorphic manifolds are the same, the notions of lengths and angles are preserved and the two manifolds are said be isometric (have the same geometry). This happens when the diffeomorphism preserves the metric tensor. A page of a book and a cylinder are isometric since the page can be bent into a cylinder without stretching or tearing.

2.9.3 Two-forms

An antisymmetric second rank covariant tensor is called a 2-form. To make the antisymmetry manifest, one defines the wedge product[46]

$$dx^i \wedge dx^j = dx^i \otimes dx^j - dx^j \otimes dx^i \tag{147}$$

and writes a 2-form with antisymmetric components ω_{ij} as (show the $2^{\rm nd}$ equality!)

$$\omega = \omega_{ij} dx^i \otimes dx^j = \frac{1}{2}\omega_{ij} dx^i \wedge dx^j. \tag{148}$$

Note that $dx^1 \wedge dx^1 = 0$, etc. Geometrically, two-forms are related to area elements on tangent two-planes in a manifold. The familiar area 'element' $dx\, dy$ on a plane is more precisely the 2-form $dx \wedge dy$. The antisymmetry of the wedge product allows us to encode the orientation of the area element, which in vector calculus is conveyed by the inward/outward normal $\hat{\boldsymbol{n}}$ in an 'infinitesimal area vector' $dx\, dy\, \hat{\boldsymbol{n}}$ on a surface parametrized by x and y.

• In Euclidean space $\mathbb{R}^3$ with Cartesian coordinates, the components of the wedge product of 1-forms df and dg are related to those of the cross product $\boldsymbol{\nabla} f \times \boldsymbol{\nabla} g$ whose magnitude measures the area of a parallelogram spanned by the vectors $\boldsymbol{\nabla} f$ and $\boldsymbol{\nabla} g$.

• The space of 2-forms is denoted $\Omega^2(M)$. Recall that functions can also be regarded as 0-forms and that we could go from functions to 1-forms by taking the differential: $df = (\partial_i f) dx^i$. The differential of a function is also called its exterior derivative. Interestingly, there is a similar way of going from 1-forms to 2-forms by (exterior) differentiation. Given a 1-form $\phi = \phi_j dx^j$, we *define* its exterior derivative

$$\omega = d\phi = d\phi_j \wedge dx^j, \tag{149}$$

which is a 2-form. To find its components we write

$$d\phi = \frac{\partial \phi_j}{\partial x^i} dx^i \wedge dx^j = \frac{1}{2}(\partial_i \phi_j - \partial_j \phi_i) dx^i \wedge dx^j \quad \text{whence} \quad \omega_{ij} = \partial_i \phi_j - \partial_j \phi_i. \tag{150}$$

We used the antisymmetry of the wedge product in the second step, relabelled indices and used the definition (148) to identify the antisymmetric tensor ω_{ij}.

However, unlike ordinary differentiation that can be done repeatedly to produce higher order derivatives of a function, the square of the exterior derivative vanishes[47]. Indeed, using the definition in (149), the exterior derivative of the 1-form df is

$$d(df) = d(\partial_j f) dx^j = (\partial_i \partial_j f) dx^i \wedge dx^j = 0. \tag{151}$$

[46] The wedge product can be written as a sum over permutations of two objects: $dx^1 \wedge dx^2 = \sum_{\sigma \in S_2} \text{sgn}\,(\sigma) dx^{\sigma(1)} \otimes dx^{\sigma(2)}$. Here S_2 is the permutation or symmetric group consisting of two elements, the identity ($\sigma(1) = 1, \sigma(2) = 2$) and the exchange transposition ($\sigma(1) = 2, \sigma(2) = 1$). sgn (σ) is the sign of the permutation: -1 to the power of the number of pairwise transpositions needed to write σ as a product of exchanges. The identity has sign $+1$ and the exchange has sign -1.

[47] We say that the exterior derivative is nilpotent of degree two: $d^2 = 0$

Here $\partial_i \partial_j f$ is symmetric under $i \leftrightarrow j$ exchange due to the equality of mixed partials, while the wedge product $dx^i \wedge dx^j$ is antisymmetric, so the sum vanishes. Thus, $d^2 f = 0$. This identity is a generalization of the vector identity $\boldsymbol{\nabla} \times (\boldsymbol{\nabla} f) = 0$ valid for real-valued functions on Euclidean space $\mathbb{R}^3$.

• A 1-form ϕ that is the differential of a smooth function ($\phi = df$) is called an exact 1-form. A 2-form ω that is the exterior derivative of a 1-form $\omega = d\alpha$ is said to be an exact 2-form.

• A 1-form ϕ whose exterior derivative vanishes is called a closed one-form. Any exact 1-form is automatically closed. Soon we will define the exterior derivative of a 2-form. As with 1-forms, we will say that a 2-form ω is closed if $d\omega = 0$.

• Just as a 1-form acts linearly on vector fields to produce functions $\phi(v) = \phi_i v^i$, a 2-form acts as a skew-symmetric bilinear map from pairs of vector fields to $\mathcal{F}(M)$:

$$\omega(u, v) = \omega_{ij} dx^i \otimes dx^j (u^k \partial_k, v^l \partial_l) = \omega_{ij} dx^i (u^k \partial_k) dx^j (v^l \partial_l) = \omega_{ij} u^i v^j. \quad (152)$$

Here, the 1st (2nd) factor in a tensor product acts on the 1st (2nd) entry of the ordered pair (u, v). We used linearity of the action of forms on vector fields (139) and the pairing $dx^i(\partial_k) = \delta^i_k$.

• A 2-form can be used to define an area for infinitesimal parallelograms in each tangent space to a manifold. For example, if ∂_i, ∂_j are two coordinate tangent vectors at x, then the area of the parallelogram they span is defined as $\omega(\partial_i, \partial_j) = \omega_{ij}(x)$. However, note that ω could assign a zero 'area' to a parallelogram spanned by linearly independent vectors. For the assigned area of such a parallelogram to be nonzero, we must ask that ω be nondegenerate. More on this when we discuss a symplectic form.

• **Examples of 2-forms.**

• (i) An interesting example of a 2-form is the electromagnetic field strength tensor F, called the Faraday tensor. It is a 2-form on $\mathbb{R}^4$ (the 4-dimensional Minkowski spacetime). It is conventional to denote the Cartesian coordinates on $\mathbb{R}^4$ by x^0, x^1, x^2, x^3 with $x^0 = ct$ where t is time and c is the speed of light. F is the exterior derivative of the 1-form 'gauge potential':

$$A = A_\mu dx^\mu \quad \text{and} \quad F = dA = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu \quad \text{where} \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (153)$$

as in (150). Here $\mu, \nu = 0, 1, 2, 3$ and $A_\mu = (\phi, -\boldsymbol{A})$ is a combination of the scalar and vector potentials of electrodynamics. It may be shown that the electric and magnetic fields appear as the components of F.

• (ii) An example of a 1-form in mechanics is the so-called canonical or Liouville 1-form on the $2n$-dimensional phase space $M = \mathbb{R}^{2n}$ of a system with n-degrees of freedom:

$$\alpha = p_i dq^i = p_1 dq^1 + p_2 dq^2 + \cdots + p_n dq^n. \quad (154)$$

We may view $\mathbb{R}^{2n}$ as the cotangent bundle of the configuration space $\mathbb{R}^n$, which is the base space of the bundle. Notice that $\xi = (q^1, \cdots, q^n, p_1, \cdots p_n)$ together furnish coordinates on $\mathbb{R}^{2n}$. Here, q^i are coordinates on the base space while p_j are coordinates on the fibers (components of a cotangent vector or momentum covector in the coordinate basis $p = p_j dx^j$). For $a = 1, 2, \cdots, 2n$, ξ^a are called Darboux or canonical coordinates on the cotangent bundle. Notice that α has no components along the dp_i. Its exterior derivative is a 2-form (using $\frac{\partial p_i}{\partial q^j} = 0$ and $\frac{\partial p_i}{\partial p_j} = \delta_i^j$):

$$\begin{aligned} \omega &= d\alpha = dp_i \wedge dq^i = -dq^i \wedge dp_i = -(dq^1 \wedge dp_1 + \cdots + dq^n \wedge dp_n) \\ &= \frac{1}{2}(-dq^1 \wedge dp_1 - \cdots - dq^n \wedge dp_n + dp_1 \wedge dq^1 + \cdots + dp_n \wedge dq^n). \end{aligned} \tag{155}$$

From this we may read off the antisymmetric components of $\omega = \frac{1}{2}\sum_{a,b=1}^{2n} \omega_{ab} d\xi^a \wedge d\xi^b$. The only nonzero ones are:

$$\omega_{i,n+i} = -1 \quad \text{and} \quad \omega_{n+i,i} = 1 \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{156}$$

This ω is called the canonical symplectic 2-form. It is invertible at each point. The inverse is the canonical Poisson tensor ($\omega_{ab} r^{bc} = \delta_a^c$). For one degree of freedom ($n = 1$), $\alpha = p dq$ and

$$\omega = \frac{1}{2}(-dq \wedge dp + dp \wedge dq) = \frac{1}{2}(\omega_{11} dq \wedge dq + \omega_{12} dq \wedge dp + \omega_{21} dp \wedge dq + \omega_{22} dp \wedge dp) \tag{157}$$

so that $\omega_{12} = -\omega_{21} = -1$ and $\omega_{11} = \omega_{22} = 0$ and $\omega = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Referring back to our discussion of the Poisson tensor earlier in this section, we observe that given a Hamiltonian function H on phase space, the Hamiltonian vector field is defined via $\omega(\cdot, V_H) = dH(\cdot)$. In components, $\omega_{ab} V_H^b = \partial_a H$ or inverting, $V_H^c = r^{ca} \partial_a H$. The integral curves of this vector field are called the phase space trajectories of the Hamiltonian system.

- On $\mathbb{R}^2$ with Cartesian coordinates $(x^1, x^2) = (x, y)$, we have the standard surface area form $\Omega = \frac{1}{2}\epsilon_{ij} dx^i \wedge dx^j$. Here ϵ_{ij} is the antisymmetric Levi-Civita symbol with $\epsilon_{12} = -\epsilon_{21} = 1$. Writing out the terms, $\Omega = \frac{1}{2}(dx^1 \wedge dx^2 - dx^2 \wedge dx^1) = dx^1 \wedge dx^2$. This is the usual surface integration element. Often, when we write $\int f(x, y) dx dy$ we mean the integral of the 2 form $f(x, y) dx \wedge dy$. We will discuss integration of forms later.

2.9.4 Mixed second rank tensors

Aside from contravariant and covariant tensors, we also have mixed second rank tensors[48] of type (1,1): $t = t^i_j \partial_i \otimes dx^j$. They transform via one Jacobian and one inverse Jacobian factor: $\tilde{t}^k_l = J^k_i (J^{-1})^j_l t^i_j$. A (1,1) tensor restricted to a point $p \in M$ can be viewed as a linear transformation on the tangent space $T_p(M)$. Indeed, contracting it with a tangent vector gives another tangent vector:

$$t(\cdot, v) = t^i_j \partial_i(\cdot) dx^j (v^k \partial_k) = t^i_j v^j \partial_i(\cdot) \quad \text{or} \quad v^i \mapsto v'^i = t^i_j v^j. \tag{158}$$

[48] We have arbitrarily chosen to place the ∂_i ahead of the dx^j in the tensor product. The opposite order can also be followed throughout.

The $\cdot$ is a placeholder for an unspecified 1-form that t could act on via the first slot. Similarly, $t^i_j v^j \partial_i(\cdot)$ is the action of a vector field on an unspecified 1-form. Analogously, t may also be viewed as a linear map from 1-forms to 1-forms, taking $\phi_i \mapsto \phi'_i = t^j_i \phi_j$. The components t^i_j of a (1,1) tensor define a matrix in the coordinate basis, and the above transformation rule written in matrix notation, $\tilde{t} = JtJ^{-1}$ is just a similarity transformation!

2.10 Higher rank tensor fields and forms

More generally, we have tensor fields of type (p,q) for $p,q \geq 0$ which, in local coordinates, are given by the linear combinations

$$t = t^{i_1 \cdots i_p}_{j_1 \cdots j_q} \partial_{i_1} \otimes \cdots \otimes \partial_{i_p} \otimes dx^{j_1} \otimes \cdots \otimes dx^{j_q}. \tag{159}$$

Their components transform via p factors of J for upper indices and q factors of J^{-1} for lower indices. Such a tensor field can act linearly on p one-forms and q vector fields to produce a function: $t(\phi, \psi, \cdots, u, v, \cdots) = t^{i_1 \cdots i_p}_{j_1 \cdots j_q} \phi_{i_1} \psi_{i_2} \cdots u^{j_1} v^{j_2} \cdots$. Thus, algebraically, (p,q) tensor fields are simply multilinear maps from p copies of $\Omega^1(M)$ and q copies of Vect(M) to the space of scalar functions on M. For instance, $t(f\phi_1 + g\phi_2, \ldots) = ft(\phi_1, \ldots) + gt(\phi_2, \ldots)$ for any scalar functions f and g.

Of particular importance are the p-forms, which are covariant antisymmetric tensor fields[49] of rank p (or of type $(0,p)$),

$$\omega = \omega_{i_1 \cdots i_p} dx^{i_1} \otimes \cdots \otimes dx^{i_p}. \tag{160}$$

Antisymmetry means the components are antisymmetric under interchange of *any* pair of indices. As a consequence, a p-form on an n dimensional manifold must be identically zero if $p > n$ (at least one basis 1-form must appear twice in the tensor product, which when contracted with an antisymmetric coefficient, must vanish). Forms can be written as linear combinations of p-fold wedge products of coordinate 1-forms, which are obtained by antisymmetrizing the p-fold tensor product:

$$\omega = \frac{1}{p!} \omega_{i_1 \cdots i_p} \, dx^{i_1} \wedge \cdots \wedge dx^{i_p}. \tag{161}$$

For instance, a three-fold wedge product is a sum over all permutations of three objects (which comprise the symmetric group[50] S_3) weighted by the signs of the permutations

[49] We may take linear combinations of p-forms ω, ψ: $f\omega + g\psi$ for any smooth functions f, g to produce other p-forms. The space of p-forms is denoted $\Omega^p(M)$.

[50] The symmetric group on 3 letters has $3! = 6$ elements. The identity $\sigma(i) = i$ denoted $(1)(2)(3)$ has sign 1. There are three pairwise transpositions $(12)(3), (1)(23)$ and $(2)(31)$ which have sign -1. For example (23) means 2 and 3 are mapped to each other. Thus, $(1)(23)$ means $\sigma(1) = 1, \sigma(2) = 3, \sigma(3) = 2$. There are also two cyclic permutations $(123) = (12)(23)$ and $(132) = (12)(13)$ which have been written as products of pairwise exchanges composed from right to left. Here (132) means $\sigma(1) = 3, \sigma(3) = 2$ and $\sigma(2) = 1$. In the composition $(12)(13)$, 3 is mapped to 1 which is then mapped to 2, so that 3 is on the whole mapped to 2. On the other hand, 2 is directly mapped to 1. The cyclic permutations have sign $+1$ as they are products of an even number of transpositions. See also Footnote 46.

(see Footnote 46):

$$\begin{aligned} dx^1 \wedge dx^2 \wedge dx^3 &= \textstyle\sum_{\sigma \in S_3} \operatorname{sgn}(\sigma)\, dx^{\sigma(1)} \otimes dx^{\sigma(2)} \otimes dx^{\sigma(3)} \\ &= dx^1 \otimes dx^2 \otimes dx^3 - dx^2 \otimes dx^1 \otimes dx^3 - dx^1 \otimes dx^3 \otimes dx^2 \\ &- dx^3 \otimes dx^2 \otimes dx^1 + dx^2 \otimes dx^3 \otimes dx^1 + dx^3 \otimes dx^1 \otimes dx^2. \quad (162) \end{aligned}$$

• Let us count the number of independent p-forms on an n-dimensional manifold. Zero forms are smooth functions. In a coordinate neighborhood, any smooth function is written as $f(x)1$. We view the constant function 1 as the only basis element with the coefficient being an arbitrary function. Any 1-form may be written as $\phi_i(x)dx^i$. So we have n independent basis 1-forms with coefficients being functions. Similarly, any 2-form is expressible as $\omega = \omega_{ij}(x)dx^i \wedge dx^j$ where $\omega_{ij}(x)$ are antisymmetric coefficients. Due to antisymmetry, there are $n(n-1)/2 = \binom{n}{2}$ independent basis 2-forms. Similarly they are $\binom{n}{p}$ independent basis p-forms for $0 \leq p \leq n$. In particular, any n-form may be written as $\Omega = \rho(x)dx^1 \wedge \cdots \wedge dx^n$: any n-form is a multiple (by some function of the coordinates) of $dx^1 \wedge \cdots \wedge dx^n$.

• An example of a 3-form on $\mathbb{R}^3$ is the **Euclidean volume form** whose components in Cartesian coordinates are given in terms of the Levi-Civita symbol:

$$\Omega = \frac{1}{3!}\epsilon_{ijk} dx^i \wedge dx^j \wedge dx^k. \quad (163)$$

Combining the six nonzero terms using the antisymmetry of the wedge product, we verify that Ω is simply the familiar volume element $\Omega = dx^1 \wedge dx^2 \wedge dx^3$. The Levi-Civita symbol generalizes to $\mathbb{R}^n$: $\epsilon_{i_1 \cdots i_n}$ is antisymmetric under every exchange of indices and satisfies $\epsilon_{12\cdots n} = 1$.

• Volume elements transform via a Jacobian determinant. Let us consider the case of $\mathbb{R}^3$, although the formulae generalize to n dimensions. On the one hand,

$$\Omega = \frac{1}{3!}\epsilon_{ijk} dx^i \wedge dx^j \wedge dx^k = dx^1 \wedge dx^2 \wedge dx^3. \quad (164)$$

Changing coordinates to y,

$$\Omega = \frac{1}{3!}(J^{-1})^l_i (J^{-1})^m_j (J^{-1})^n_k \epsilon_{lmn} dy^i \wedge dy^j \wedge dy^k \quad (165)$$

This is the usual coordinate transformation formula: the components of a covariant third rank tensor transform via three factors of the inverse Jacobian. The question we pose is how the simplified expression $dx^1 \wedge dx^2 \wedge dx^3$ transforms. To answer this, we use the formula $\epsilon_{ijk} \det A = A^l_i A^m_j A^n_k \epsilon_{lmn}$, which holds for any 3×3 matrix and apply it to $A = J^{-1}$. We then find that

$$\Omega = \frac{1}{3!}(\det J^{-1})\epsilon_{ijk} dy^i \wedge dy^j \wedge dy^k. \quad (166)$$

As before, all the six nonzero terms in the sum contribute equally, leaving us with

$$\Omega = dx^1 \wedge dx^2 \wedge dx^3 = (\det J^{-1}) dy^1 \wedge dy^2 \wedge dy^3. \quad (167)$$

This is the sense in which we say that the volume element transforms via a Jacobian determinant. Note that this generalizes to n dimensions where we use the formula $\epsilon_{i_1\cdots i_n}\det A = A^{j_1}_{i_1}A^{j_2}_{i_2}\cdots A^{j_n}_{i_n}\epsilon_{j_1\cdots j_n}$ for any $n\times n$ matrix A and choose A as the inverse of the Jacobian matrix.

$*$ **Levi-Civita tensor density.** We define the Levi-Civita symbol ϵ_{ijk} to have the same components in any coordinate system. Suppose $x^i \to \tilde{x}^i$, is a change of coordinates, then $\tilde{\epsilon}_{ijk} = \epsilon_{ijk}$. What is more, for any invertible matrix A, we have the identity

$$\epsilon_{ijk} = \det A^{-1}\, A^l_i A^m_j A^n_k\, \epsilon_{lmn}. \tag{168}$$

Applying this to the inverse Jacobian matrix $A = J^{-1}$ where $(J^{-1})^l_i = \frac{\partial x^l}{\partial \tilde{x}^i}$, we get

$$\tilde{\epsilon}_{ijk} = \epsilon_{ijk} = \det\left(\frac{\partial \tilde{x}^a}{\partial x^b}\right)\frac{\partial x^l}{\partial \tilde{x}^i}\frac{\partial x^m}{\partial \tilde{x}^j}\frac{\partial x^n}{\partial \tilde{x}^k}\epsilon_{lmn}. \tag{169}$$

If the determinant did not appear on the right, ϵ_{ijk} would transform as the components of a $(0,3)$ tensor. To account for the first power of the Jacobian determinant, we say that ϵ_{ijk} transform as the components of a tensor density of weight 1. □

2.11 Pushforward and pullback of tensors

Given a pair of smooth manifolds X and Y with dimensions n and n' and a smooth map $\phi : X \to Y$, we may, in favorable cases, use ϕ to move tensor fields between the manifolds. This finds application, for instance, in deducing the induced metric on a submanifold embedded in Euclidean space. However, moving tensors only works in certain directions. Forms and more generally covariant tensor fields on Y may be 'pulled back' to X, the pullback being denoted ϕ^*. On the other hand, vector fields (and more generally contravariant tensor fields) on X may, in some cases, be 'pushed forward' to Y via ϕ_*. Combining these, if ϕ is a diffeomorphism (*invertible* smooth map with smooth inverse) then the pullback and pushforward via ϕ and ϕ^{-1} may be used to move arbitrary tensor fields in either direction.

Pullback. The simplest tensor field is a scalar function. Given a smooth function $f : Y \to \mathbb{R}$, its pullback is the function $\phi^* f : X \to \mathbb{R}$ defined as $(\phi^* f)(x) = f(\phi(x))$ for any $x \in X$. In other words, we simply compose f with ϕ to go from X to $\mathbb{R}$ in two steps, $\phi^* f : X \xrightarrow{\phi} Y \xrightarrow{f} \mathbb{R}$. For example, suppose $\phi : S^2 \to \mathbb{R}^3$ is the map $\phi(\theta,\varphi) = (x = \sin\theta\cos\varphi, y = \sin\theta\sin\varphi, z = \cos\theta)$ and let $f(x,y,z) = z$ be the height function. Then the pullback $(\phi^* f)(\theta,\varphi) = \cos\theta$ is the function that assigns the cosine of the polar angle to any point on the sphere. On the other hand, it is generally not possible to define the pushforward of a function. For this reason, we will view scalar functions as covariant (rather than contravariant) tensors of rank zero. This is one reason we viewed $C^\infty(M)$ as the space of zero-forms $\Omega^0(M)$.

More generally, the gadget that helps us do this pushing and pulling is the linearization or differential $d\phi$ of the map ϕ. Suppose x^i and y^j are local coordinates on X and Y and $y = \phi(x)$ or $y^j = \phi^j(x)$. Then the linearization at the point x is represented by the $n' \times n$ Jacobian matrix with entries $\frac{\partial \phi^j}{\partial x^i}$. Next, given a 1-form $\omega_j dy^j$ on Y we define its pullback at a point $x \in X$, denoted $(\phi^* \omega)(x)$ via

$$(\phi^* \omega)_i(x) = \frac{\partial \phi^j}{\partial x^i} \omega_j(\phi(x)). \tag{170}$$

Notice that no assumption on the invertibility of ϕ has been made. This suggests why it is not possible, in general, to pushforward a differential form. If ϕ is invertible (say when $X = Y$ and ϕ is a diffeomorphism), we may multiply by the inverse Jacobian and formally recover the coordinate transformation formula of (138). However, there is a conceptual difference: while a map $\phi : X \to X$ actively moves points around, a coordinate transformation only relabels them. The generalization to the pullback of covariant rank-p tensor fields (including p-forms) is:

$$(\phi^* \omega)_{i_1 \cdots i_p}(x) = \frac{\partial \phi^{j_1}}{\partial x^{i_1}} \cdots \frac{\partial \phi^{j_p}}{\partial x^{i_p}} \omega_{j_1 \cdots j_p}(\phi(x)). \tag{171}$$

Evidently, the pullback of a smooth function is the special case when $p = 0$.

Pushforward. Pushing forward vector fields or contravariant tensors is not so straightforward. To begin with, we note that the linearization of ϕ defines a linear transformation $d\phi$ between tangent spaces. If $y = \phi(x)$, then $d\phi(x) : T_x X \to T_y Y$. Once coordinates are chosen, this map is represented by the $n' \times n$ Jacobian matrix. Now if $v = v^i \partial_{x^i} \in T_x X$, then it is natural to define its pushforward to be the image of the vector v under the linear transformation $d\phi$. Thus, we are tempted to define the pushforward $\phi_* v$ as the vector field whose components at $y = \phi(x)$ are given by

$$(\phi_* v)^j(y) = \frac{\partial \phi^j}{\partial x^i} v^i(x) \quad \text{for} \quad j = 1, 2, \ldots, n'. \tag{172}$$

Note that if ϕ is not surjective (say, if $n' > n$) then this does not define a vector field on all of Y. Nevertheless, we can try to define a pushforward vector field on the image $\phi(X) \subset Y$. However, there is a further difficulty with (172): suppose ϕ is many to one with $y = \phi(x_1) = \phi(x_2)$. Then it may happen that the images of v via the Jacobians $\frac{\partial \phi^j}{\partial x^i}$ at x_1 and x_2 are not the same, giving rise to an ambiguity in the definition of the vector field. This difficulty does not arise when ϕ is one-to-one and we can write $x = \phi^{-1}(y)$ on the RHS of (172) to arrive at a pushforward vector field $\phi_* v$ on $\phi(X)$. See Prob. **??** for a simple example. The definition has a straightforward generalization to rank p contravariant tensor fields for any $p = 1, 2, \ldots$:

$$(\phi_* t)^{j_1 \cdots j_p}(y) = \frac{\partial \phi^{j_1}}{\partial x^{i_1}} \frac{\partial \phi^{j_2}}{\partial x^{i_2}} \cdots \frac{\partial \phi^{j_p}}{\partial x^{i_p}} t^{i_1 \cdots i_p}(x). \tag{173}$$

As before, we notice the formal similarity with the coordinate transformation laws [e.g., (142)] for contravariant tensor fields when X and Y are the same manifold. What is more, if ϕ is a diffeomorphism then it is both injective and surjective so that (173) unambiguously defines a pushforward tensor field on all of Y.

Pullback of a metric: induced metric via an example. Consider the unit sphere S^2 ($x^2+y^2+z^2=1$) embedded as a submanifold of $\mathbb{R}^3$. If we use polar coordinates ($\xi^1=\theta, \xi^2=\varphi$) on S^2, the embedding is defined by a smooth map $\phi : S^2 \to \mathbb{R}^3$ given by $\phi(\theta,\varphi) = (x = \sin\theta\cos\varphi, y = \sin\theta\sin\varphi, z = \cos\theta)$. Now, $\mathbb{R}^3$ has the standard flat Euclidean metric whose components in Cartesian coordinates are $g_{ij} = \delta_{ij}$. We may pullback this rank-2 covariant symmetric tensor field to get an 'induced' metric h_{ab} on S^2 with components

$$h_{ab} = (\phi^* g)_{ab} = \frac{\partial \phi^i}{\partial \xi^a}\frac{\partial \phi^j}{\partial \xi^b} g_{ij}. \tag{174}$$

This formula defines the induced metric and is of course not special to the above example. In the case of the embedding $\phi : S^2 \hookrightarrow \mathbb{R}^3$, the induced metric $h_{ab}(\theta,\varphi)$ is the familiar 'round sphere' metric. Evaluate its components.

• By the Nash embedding theorems, essentially any Riemannian manifold M of dimension n with metric tensor g can be isometrically embedded in a Euclidean space of sufficiently large dimension N. This means the metric on M can be realized as the pull back of the Euclidean metric on R^N for a suitable embedding. In particular, the standard metrics on spheres, ellipsoids, hyperboloids, etc., of various dimensions are obtained as pullbacks of Euclidean metrics using familiar embeddings.

• The pushforward of a vector field can be used to define the Lie derivative of a vector field v along a vector field u on a smooth manifold M. $\mathcal{L}_u v$ is the derivative of v along the integral curves of u. Suppose t is the parameter along the integral curve $x(t)$ of u through the point $x(0)$. We cannot take the difference between $v(x(0+\delta t))$ and $v(x(0))$ since the vectors lie in distinct tangent spaces. However, for each small t the flow ϕ_t defined by u defines a smooth 1-1 onto map in a sufficiently small neighborhood of $x(0)$. Thus, we can use this flow to pushforward the vector $v(x(0))$ to the point $x(\delta t)$ and do the subtraction in the tangent space to M at $x(\delta t)$. In this way, we can define the Lie derivative as the limit of the difference quotient

$$\mathcal{L}_u v(x(0)) = \lim_{\delta t \to 0} \frac{v(x(\delta t)) - \phi_{\delta t *} v(x(0))}{\delta t}. \tag{175}$$

With some more effort, one can show that the formula one obtains this way is the same as the commutator of vector fields $[u,v]$.

2.12 Exterior algebra, exterior derivative and Bianchi's identity

Exterior algebra. In Sect. 2.8 and Sect. 2.9, we introduced 1- and 2-forms. One-forms can be used to describe the momentum of a particle on the configuration space of a mechanical system, the Liouville form '$p\,dq$' on phase space, the infinitesimal heat added to a gas in a thermodynamic process or the electromagnetic 'scalar' and 'vector' potentials. Two-forms are used to model infinitesimal area elements, the electromagnetic field strength tensor $F_{\mu\nu}$ and the symplectic form ω_{ij} in mechanics. Furthermore, the wedge product of two 1-forms was seen to produce a 2-form. On the other hand, the differential or exterior derivative of a function df was shown to give a

1-form, while the exterior derivative of a 1-form led to a 2-form. In this section, we extend the wedge product and exterior derivative to forms of any rank (introduced in Sect. 2.10) and also discuss an analog of the Leibniz rule for the exterior derivative of a wedge product. The space of differential forms with these algebraic properties is called the exterior algebra. These developments are then applied to understand some properties of the symplectic form ω of Hamiltonian mechanics.

Recall from Sect. 2.10, that a differential form of order $p = 0, 1, 2, \ldots$ or p-form ω in a patch with coordinates x^i is a linear combination of p-fold wedge products of coordinate 1-forms:

$$\omega = \frac{1}{p!}\omega_{i_1 \cdots i_p} dx^{i_1} \wedge \cdots \wedge dx^{i_p} \tag{176}$$

where $\omega_{i_1 \cdots i_p}$ is totally antisymmetric. Given any smooth p-forms ω, ψ and smooth functions f, g, $f\omega + g\psi$ is also a smooth p-form. Thus, the space of p-forms denoted $\Omega^p(M)$ is said to be a module (see Footnote **??**) over the ring of smooth real-valued functions on M ($\mathcal{F}(M)$ of Sect. 2.6).

Owing to the antisymmetry of $dx^{i_1} \wedge \cdots \wedge dx^{i_p}$ there are no nonzero p forms for $p > n$ on a manifold of dimension n. For instance on $\mathbb{R}$, there is only one coordinate 1-form dx and the only possible coordinate basis 2-form $dx \wedge dx$ vanishes by antisymmetry (there is no concept of area on a line). On $\mathbb{R}^2$, we have two coordinate basis 1-forms dx and dy, one independent basis 2 form $dx \wedge dy = -dy \wedge dx$ and no nonzero 3-forms as $dx \wedge dy \wedge dx$, etc., all vanish. In fact, the number of linearly independent p-forms at a point is the binomial coefficient $\binom{n}{p}$ since each choice of p distinct coordinate 1-forms $dx^{i_1}, \ldots, dx^{i_p}$ furnishes one coordinate basis p-form $dx^{i_1} \wedge \ldots \wedge dx^{i_p}$. In particular, there is only one ($= \binom{n}{0}$) independent 0-form and one ($= \binom{n}{n}$) independent n-form. What we mean is that any 0-form is some smooth function times the constant function 1 and any n-form is some smooth function times the volume form $dx^1 \wedge \cdots \wedge dx^n$.

We may take the direct sum of the spaces of p-forms to obtain the $\sum_{p=0}^{n} \binom{n}{p} = 2^n$ dimensional space of all differential forms on M:

$$\Omega(M) = \oplus_{p=0}^{n} \Omega^p(M). \tag{177}$$

In addition to taking linear combinations of forms, we may take their wedge product. For the coordinate basis forms

$$(dx^{i_1} \wedge \cdots \wedge dx^{i_p}) \wedge (dx^{i_{p+1}} \wedge \cdots \wedge dx^{i_{p+q}}) = dx^{i_1} \wedge \cdots \wedge dx^{i_{p+q}}. \tag{178}$$

For example, $(dx \wedge dy) \wedge (dz \wedge dw) = dx \wedge dy \wedge dz \wedge dw$. By repeated use of the antisymmetry property $dx^i \wedge dx^j = -dx^j \wedge dx^i$, we may show that the wedge product of a p-form ω and a q-form ψ is (anti)commutative:

$$\omega \wedge \psi = (-1)^{pq} \psi \wedge \omega. \tag{179}$$

Let us explain the origin of the sign. Suppose $\omega = dx$ and $\psi = dy \wedge dz$ so that $p = 1$ and $q = 2$. Then dx has to 'pass through' dy and dz producing two minus signs resulting in $dx \wedge (dy \wedge dz) = (-1)^{1 \cdot 2}(dy \wedge dz) \wedge dx$. Similarly, suppose we consider $(dx \wedge dy) \wedge (du \wedge dv \wedge dw)$. Here we move dy first through the 3-form picking up a $(-1)^3$ and then move dx and get another $(-1)^3$. Thus we see the emergence of p factors of $(-1)^q$ leading to the sign $(-1)^{pq}$.

Equipped with this wedge product, $\Omega(M)$ is called the exterior algebra. A special case is the wedge product of a p-form ω and a 0-form f: $\omega \wedge f = (-1)^0 f \wedge \omega = f\omega$.

Exterior derivative. The exterior derivative may be extended to a map from p-forms to $(p+1)$-forms: $d : \Omega^p(M) \to \Omega^{p+1}(M)$ for any $p = 0, 1, 2, \ldots$ satisfying the three axioms:

1. Linearity: $d(a\omega + b\psi) = ad\omega + bd\psi$ for any $a, b \in \mathbb{R}$ and p-forms ω, ψ.
2. Leibniz (antiderivation) rule: $d(\omega \wedge \phi) = d\omega \wedge \phi + (-1)^p \omega \wedge d\phi$ where $\omega \in \Omega^p(M)$ and ϕ is any form.
3. Nilpotent[51] of degree two: $d^2\omega = 0$ for any p-form ω.

The need for the minus sign in this Leibniz rule is already evident if we consider the wedge product of a 1-form $\phi = \phi_j dx^j$ and a zero form f. Now $\phi \wedge f = f \wedge \phi = f\phi$. We will calculate $d(\phi \wedge f)$ from first principles and see the emergence of the minus sign. In fact, using $d\phi = \partial_i \phi_j dx^i \wedge dx^j$, we get

$$\begin{aligned} d(\phi \wedge f) &= d(f\phi) = \partial_i (f\phi_j) dx^i \wedge dx^j = \left((\partial_i f)\phi_j + f\partial_i \phi_j\right) dx^i \wedge dx^j \\ &= df \wedge \phi + f d\phi = -\phi \wedge df + d\phi \wedge f = d\phi \wedge f - \phi \wedge df. \end{aligned} \quad (180)$$

• Explicit formula.

$$d\alpha = \frac{1}{p!}\partial_k \alpha_{i \cdots j} dx^k \wedge dx^i \wedge \cdots \wedge dx^j = \frac{1}{(p+1)!}(p+1)\partial_{[k}\alpha_{i \cdots j]} dx^k \wedge dx^i \wedge \cdots \wedge dx^j. \quad (181)$$

For example $d(\alpha_i dx^i) = \frac{1}{2!} 2\partial_{[j}\alpha_{i]} dx^j \wedge dx^i$ with $(d\alpha)_{ji} = 2\partial_{[j}\alpha_{i]} = \partial_j \alpha_i - \partial_i \alpha_j$.
• Example: Exterior derivative of a 2-form and the Bianchi formula. Suppose $\omega = \frac{1}{2}\omega_{jk} dx^j \wedge dx^k$ is a 2-form with antisymmetric components ω_{jk}. Then what is its exterior derivative? Using linearity and the Leibniz rule,

$$d\omega = \frac{1}{2} d(\omega_{jk} dx^j \wedge dx^k) = \frac{1}{2}(\partial_i \omega_{jk}) dx^i \wedge dx^j \wedge dx^k. \quad (182)$$

Since $dx^i \wedge dx^j \wedge dx^k$ is antisymmetric under exchange of any pair of indices, only the similarly antisymmetric part of $\partial_i \omega_{jk}$ can contribute. Antisymmetrizing as in (162), we write

$$\begin{aligned} d\omega &= \frac{1}{12}\left(\partial_i \omega_{jk} - \partial_j \omega_{ik} - \partial_k \omega_{ji} - \partial_i \omega_{kj} + \partial_k \omega_{ij} + \partial_j \omega_{ki}\right) dx^i \wedge dx^j \wedge dx^k \\ &= (1/3!)\left(\partial_i \omega_{jk} + \partial_k \omega_{ij} + \partial_j \omega_{ki}\right) dx^i \wedge dx^j \wedge dx^k, \end{aligned} \quad (183)$$

[51] $d^2 = 0$ is a generalization of the vector calculus identities $\boldsymbol{\nabla} \times \boldsymbol{\nabla} f = 0$ and $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}) = 0$ for any smooth function f and vector field $\boldsymbol{v}$ in $\mathbb{R}^3$.

In the first equality, every one of the 6 terms contributes equally (check this), this explains the division by 6. We used the antisymmetry of ω in the last step. Thus $(d\omega)_{ijk} = \partial_i \omega_{jk} + \partial_k \omega_{ij} + \partial_j \omega_{ki}$. This is Bianchi's formula for the exterior derivative of a 2-form. It follows that $d\omega = 0$ iff the Bianchi identity $\partial_i \omega_{jk} + \partial_k \omega_{ij} + \partial_j \omega_{ki} = 0$ is satisfied in each coordinate neighborhood.

Closed and exact forms on a manifold M. A p-form ω such that $d\omega = 0$ is said to be closed. On the other hand, if $\omega = d\phi$ for some $(p-1)$-form ϕ, then ω is said to be exact (generalizing the idea of an exact differential). Since $d^2 = 0$, an exact form is automatically closed. The converse need not be true: a closed from need not be exact. The linear space of closed p-forms is called $Z^p(M)$ while the linear space of exact p-forms is called $B^p(M)$. The quotient linear space of closed p-forms modulo exact p-forms is called the p^{th} **de Rham cohomology** (group) of the manifold, denoted $H^p(M)$.

• **Example from Maxwell Theory.** The homogeneous Maxwell equations $\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$ and $\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{\nabla} \times \boldsymbol{E} = 0$ are together the statement that the Faraday 2-form $F = (1/2) F_{\mu\nu} dx^\mu \wedge dx^\nu$ on Minkowski space-time $\mathbb{R}^4$ (with $x^\mu = (ct, x, y, z)$ for $\mu = 0, 1, 2, 3$) is a closed 2-form: $dF = 0$. Here, $F_{0i} = E_i$ and $F_{ij} = -\epsilon_{ijk} B_k$ for $1 \leq i, j, k \leq 3$. It follows from Poincaré's Lemma that $\mathbb{R}^4$ has trivial cohomology groups: any closed form is exact. Thus, F must be exact and expressible as $F = dA$ for some 'gauge potential' 1-form $A = A_\mu dx^\mu$. This is why we may express $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ and $\boldsymbol{E} = -\boldsymbol{\nabla}\phi - c^{-1}\frac{\partial \boldsymbol{A}}{\partial t}$ in terms of the scalar and vector potentials which are the components of $A_\mu = (\phi, -\boldsymbol{A})$.

Symplectic form and Bianchi's identity. Suppose $\alpha = p_i dq^i$ is the canonical Liouville 1-form on the phase space $\mathbb{R}^{2n}$ of a mechanical system with n degrees of freedom. Then we have seen that the canonical symplectic form is given by $\omega = d\alpha = dp_i \wedge dq^i$. It follows that $d\omega = d^2\alpha = 0$. In other words, the canonical symplectic form is closed. More generally (see Prob. **??**) the Jacobi identity implies that the inverse ω of any invertible (but not necessarily canonical) Poisson tensor r satisfies the Bianchi identity $\partial_i \omega_{jk} + \partial_k \omega_{ij} + \partial_j \omega_{ki} = 0$. From the foregoing discussion, this is simply the condition $d\omega = 0$. So the closedness of the symplectic form is a restatement of the Jacobi identity satisfied by the Poisson bracket. We begin to see the economy and clarity that the use of differential forms can bring to tensor calculus. What is more, given a smooth 'Hamiltonian' function H on M, we may use ω to define a vector field v_H called the Hamiltonian vector field via the formula

$$v_H(\cdot) = \omega^{-1}(\cdot, dH). \tag{184}$$

Here ω^{-1} is a contravariant (antisymmetric) second rank tensor that can act on a pair of 1-forms, one of which is chosen to be dH. The resulting object is a vector field as it can act linearly on an (unspecified) 1-form.

More generally, even if we do not have canonical (q-p-type) coordinates on phase space and do not have available the canonical Liouville 1-form $\alpha = p_i dq^i$, we may still wish to define a symplectic form using physical or geometric considerations. From the foregoing, the essential conditions it must satisfy are invertibility and the Bianchi identity. Thus, one defines a symplectic manifold as a sufficiently smooth manifold that is equipped with a closed nondegenerate (i.e., invertible) two-form ω called the symplectic form. Though it is required to be closed, ω need not be exact. We will see an example in the context of the 2-sphere.

2.13 Integration on manifolds and Stokes' theorem

We now move from the exterior differential calculus to the integral calculus on a manifold. This will allow us to generalize the concepts of line, surface and volume integrals to manifolds. To do this, we first need the idea of an oriented manifold.

Orientability of a manifold. We may orient a curve $\boldsymbol{\gamma}$ in 3d space by placing arrows on it so that the curve is traversed in only one direction. A parametrized curve $\boldsymbol{\gamma}(s) : (0,1) \to \mathbb{R}^3$ with $\dot{\boldsymbol{\gamma}} \neq 0$ everywhere has a natural orientation, namely the direction in which $\dot{\boldsymbol{\gamma}}$ points (which is the same as that of increasing s). If $\dot{\boldsymbol{\gamma}}$ vanishes somewhere, we would not know which way the arrow points there. Moreover, we can also have the situation where the parametrized curve $\boldsymbol{\gamma}$ retraces the image so that there would be points on the curve where the arrow points in both directions. This may be avoided by assuming that $\dot{\boldsymbol{\gamma}} \neq 0$. Given a vector field $\boldsymbol{v}$ and such a curve $\boldsymbol{\gamma}$, we may define the line element $d\boldsymbol{\gamma} = \dot{\boldsymbol{\gamma}}(t)dt$ and the line integral of $\boldsymbol{v}$ along $\boldsymbol{\gamma}$: $\int_{\boldsymbol{\gamma}} \boldsymbol{v} \cdot d\boldsymbol{\gamma}$. Note that $\boldsymbol{\gamma}$ need not be an integral curve of $\boldsymbol{v}$. One verifies that this line integral is reparametrization invariant[52]. Indeed, suppose $s = s(t) : [0,1] \to [0,1]$ is a reparametrization (invertible map) and let $\tilde{\boldsymbol{\gamma}}(t) = \boldsymbol{\gamma}(s(t))$. Then

$$\int \boldsymbol{v} \cdot d\tilde{\boldsymbol{\gamma}} = \int_0^1 v_j \frac{d\tilde{\gamma}^j}{dt} dt = \int_0^1 v_j \frac{d\gamma^j}{ds}\frac{ds}{dt} dt = \int_0^1 v_j \frac{d\gamma^j}{ds} ds = \int \boldsymbol{v} \cdot d\boldsymbol{\gamma}. \quad (185)$$

For a 2d surface Σ embedded in $\mathbb{R}^3$, we usually speak of an outward or inward pointing unit normal at each point of Σ. When Σ is defined by the condition $C(x,y,z) = 0$, the normal in the direction of increasing C is given by the unit vector along the gradient $\boldsymbol{\nabla} C$. To be well-defined (unambiguous), when the normal is followed around any closed loop on Σ, it must return to its original direction. When this happens, we say that the surface is oriented. In vector calculus, this normal to the surface is used to define a vectorial area element $(\hat{n} dS)$ that goes into the definition of surface integrals. These concepts can be generalized to manifolds of any dimension and are used to define integration on manifolds. An n-dimensional manifold is orientable if it admits a nonvanishing[53] form of top degree n (called a volume form). The choice of such a form is called an orientation. On $\mathbb{R}^2$ we usually choose the orientation as given by $dx \wedge dy$, the choice $dy \wedge dx$ is equally valid, but would correspond to reversing the orientation. On $\mathbb{R}^n$, the standard volume form is $dx^1 \wedge \cdots \wedge dx^n$.

[52] When we use a line integral to model the work $W = \int \boldsymbol{F} \cdot d\boldsymbol{\gamma}$ done by a force field (viewed as a vector field on the configuration space) as a particle moves along a path, we are asserting that the work done is independent of how fast the particle moves at various places along the path.

[53] A p-form ω is nonvanishing if at each point on the manifold, $\omega(u, v, \ldots) \neq 0$ for any p linearly independent tangent vector fields $u, v, \cdots$.

Admitting a volume form is equivalent to the Jacobian determinants of all the transition functions between coordinate charts being positive, so that all the coordinate charts have a common orientation[54] and the atlas may be called an oriented atlas. The two-sphere is orientable since the standard area form on S^2 is a nonvanishing 2-form (proportional to $\sin\theta d\theta \wedge d\phi$ in polar coordinates, this can be extended to a nonvanishing form at the poles as well by using 2 patches). For a surface in $\mathbb{R}^3$, orientability allows us to unambiguously distinguish two sides of the surface. The Möbius strip is not orientable[55]: one can go from the 'upper' side of the surface to the 'lower' side at the same point by taking a walk on the strip; this is not possible on a cylindrical surface or on a sphere, which are orientable.

Riemannian volume form. On an oriented n-dimensional Riemannian or pseudo-Riemannian manifold M with nondegenerate metric g, one has a natural volume form ω_g. In local coordinates x^i, it is $\omega_g = \sqrt{|\det g|}\, dx^1 \wedge \cdots \wedge dx^n$. Since g_{ij} is invertible, $\det g \neq 0$ so that this is a nonvanishing form. Let us check that this formula holds in any coordinate system. As noted in Sect. 2.10, under a coordinate change $x^i \to y^i$ a volume form $dx^1 \wedge \cdots \wedge dx^n$ transforms to $\det J^{-1} dy^1 \wedge \cdots \wedge dy^n$ where $J^i_j = \frac{\partial y^i}{\partial x^j}$ is the Jacobian matrix. If the transformation is orientation-preserving, then $\det J^{-1} > 0$. Now, the metric components transform to $\tilde{g}_{ij} = g_{kl}(J^{-1})^k_i (J^{-1})^l_j$. Hence, $\det \tilde{g} = \det g \det((J^{-1})^t) \det J^{-1}$. Consequently,

$$\begin{aligned}\sqrt{|\det g|} dx^1 \wedge \cdots \wedge dx^n &= \frac{\sqrt{|\det \tilde{g}|}}{\det J^{-1}} \det J^{-1} dy^1 \wedge \cdots \wedge dy^n \\ &= \sqrt{|\det \tilde{g}|} dy^1 \wedge \cdots \wedge dy^n. \qquad (186)\end{aligned}$$

We see that in any coordinate system, the Riemannian volume form has the same expression.

• Show that the Riemannian volume form on the round unit sphere is $\omega_g = \sin\theta d\theta \wedge d\phi$ in polar coordinates on S^2. Show that the Riemannian volume form in the spherical polar coordinate patch on 3d Euclidean space ($\mathbb{R}^3$) is given by $\omega = r^2 \sin\theta dr \wedge d\theta \wedge d\phi$.

What is more, if one takes any orthonormal basis $\phi^1, \phi^2, \cdots, \phi^n$ for 1-forms on M, then $\omega = \pm\phi^1 \wedge \phi^2 \wedge \cdots \wedge \phi^n$. For example, consider 3d Euclidean space $\mathbb{R}^3$. In Cartesian coordinates, the Euclidean metric has components $g_{ij} = \delta_{ij}$ with unit determinant and the Euclidean volume form is $\omega = dx^1 \wedge dx^2 \wedge dx^3$. In spherical polar coordinates, the nonzero metric components are $g_{rr} = 1, g_{\theta\theta} = r^2, g_{\phi\phi} = r^2 \sin^2\theta$ so that

[54] It is possible to concoct a nonoriented atlas. Consider the Euclidean plane $\mathbb{R}^2$ and define a new manifold via two overlapping patches. The left patch $x < 1$ and the right patch $x > -1$. On the left patch we define local coordinates $\xi^1 = x, \xi^2 = y$ while on the right patch we define local coordinates $\eta^1 = y, \eta^2 = x$. They overlap along the strip $-1 < x < 1$ where the point (x, y) has two addresses or sets of coordinates: (ξ^1, ξ^2) and (η^1, η^2). The transition functions are $\eta^1 = \xi^2$ and $\eta^2 = \xi^1$ resulting in an off-diagonal Jacobian matrix $\frac{\partial \eta^i}{\partial \xi^j} = (0, 1|1, 0)$ with determinant -1. Through this atlas, we have defined a manifold that is not orientable, the two charts have opposite orientations.

[55] A cylinder is constructed by taking a rectangular page from a tall book and pasting the two short edges together: the left of the bottom edge to the left of the top edge. The Möbius strip is obtained by twisting the bottom edge once before pasting it onto the top edge, so that the left of the bottom edge is joined to the right of the top edge.

$\det g = r^4 \sin^2\theta$ and the volume form becomes $\omega = r^2 \sin\theta dr \wedge d\theta \wedge d\phi$. On the other hand, the inverse metric is $g^{ij} = \text{diag}(1, 1/r^2, 1/(r^2\sin^2\theta))$ so that the coordinate 1-forms have squared-lengths $g^{-1}(dr, dr) = 1, g^{-1}(d\theta, d\theta) = 1/r^2, g^{-1}(d\phi, d\phi) = 1/r^2\sin^2\theta$. It follows that $\{dr, rd\theta, r\sin\theta d\phi\}$ is an orthonormal basis for 1-forms. We see that their wedge product is the volume form ω. It is also noteworthy that the length of $d\phi$ diverges at the poles, where $\sin\theta = 0$.

Integration of forms. In vector calculus, we define line integrals, surface integrals and volume integrals. These are examples of the integration of a 1-form along a curve, a 2-form over a surface and a 3-form over a 3d manifold. More generally, a p-form ψ may be integrated over an oriented p-dimensional manifold M to obtain a real number denoted $\int_M \psi$. To evaluate the integral, the manifold is covered by nonoverlapping cells and their boundaries (each lying within a coordinate chart) and the integral is a sum of contributions from each cell[56]. In each cell, the integral is evaluated as in multivariable calculus. In more detail, the p-form ψ can be written as $\psi = f\omega$ where f is a scalar and ω the volume form. Moreover, within a patch with coordinates $x^1, \cdots, x^p$, the volume form may be written as $\omega = \mu(x) dx^1 \wedge \cdots \wedge dx^p$ for some nonvanishing function μ. Then the contribution of the cell C is $\int_C \psi = \int_{x(C)} f(x)\mu(x) dx^1 \cdots dx^p$ where $x(C)$ is the image of the cell in $\mathbb{R}^p$. Examples: (i) We may integrate the 1-form $\psi = f(x)dx$ over the submanifold $I = (1, 2) \cup (3, 6)$ of $\mathbb{R}$:

$$\int_I \psi \equiv \int_1^2 f(x)dx + \int_3^4 f(x)dx + \int_4^6 f(x)dx. \tag{187}$$

Here we have chosen the 'increasing' orientation $\omega = dx$ (as opposed to $-dx$) and broken I into three cells. (ii) The polar coordinate patch $x^i = (\theta, \phi)$ along with its boundary covers the unit sphere S^2. So the integral of the 2-form $\psi = f\omega$ on S^2 may be expressed as

$$\int_{S^2} \psi = \int_{x(S^2)} f(\theta, \phi) \, \sin\theta \, d\theta \wedge d\phi \equiv \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \, f(\theta, \phi) \sin\theta. \tag{188}$$

The orientation has been chosen so that for $f = 1$, the integral of $\omega = \sin\theta d\theta \wedge d\phi$ over S^2 gives the area 4π of the unit sphere.

Manifold with boundary. To discuss Stokes' theorem, we need to generalize the notion of a manifold to include manifolds with boundary. By the definition of Sect. 2.2, the closed unit disk D ($x^2 + y^2 \leq 1$) contained in the plane is not a manifold, since points of D on the rim (with $x^2 + y^2 = 1$) do not have open neighborhoods lying within D. For points on the rim, we will allow neighborhoods of a different sort: roughly those shaped like a half Moon that include nearby points on the rim. There is an obvious sense in which the unit circle S^1 is the boundary of D, which we indicate via $\partial D = S^1$. The set theoretic difference $D \setminus \partial D$ is the open unit disk, it is called the interior of D. More generally, a manifold with boundary is a (topological)

[56] If ψ has bounded components, the boundaries between cells do not contribute to $\int_M \psi$ so it does not matter if these boundaries are omitted or counted a finite number of times.

space M with two types of points: (a) *interior points* which together comprise an n-dimensional manifold (i.e., which have open neighborhoods homeomorphic to $\mathbb{R}^n$ or the n-ball $B_n : x_1^2 + \cdots + x_n^2 < 1$) and (b) *boundary points* which together comprise an $n-1$ dimensional manifold called the boundary (∂M) consisting of points of M which have a neighborhood homeomorphic to a half space ($\boldsymbol{x} \in \mathbb{R}^n$ with $x_1 \geq 0$) or half ball ($\boldsymbol{x} \in B_n$ with $x_1 \geq 0$) with the homeomorphism taking the boundary points to points with $x_1 = 0$.

• A consequence of the definition of a manifold M with boundary is that $\partial\partial M = 0$, i.e., the boundary of the boundary is empty.

• An orientation on M induces an orientation on its boundary ∂M. This is familiar to us from surfaces Σ in 3d Euclidean space, a choice of 'outward' normal on Σ induces an orientation of the curve $\partial\Sigma$ that runs along the boundary of Σ, determined by the right hand thumb rule. Somewhat more generally, the orientation $dx^1 \wedge \cdots \wedge dx^n$ on the interior of the half space induces the orientation $dx^2 \wedge \cdots \wedge dx^n$ on the boundary defined by the condition $x^1 = 0$.

Stokes' theorem. Suppose $\omega = d\phi$ is an exact p-form on a p-dimensional manifold M with boundary denoted ∂M. Then Stokes' theorem

$$\int_M d\phi = \int_{\partial M} \phi \tag{189}$$

expresses the integral of ω over M as that of the $(p-1)$-form ϕ over the $(p-1)$-dimensional boundary ∂M. In particular, the integral of an exact form over a manifold without boundary vanishes. This is a generalization of Gauss' divergence theorem and Kelvin's and Stokes' theorem from vector calculus[57]

$$\int_\Omega \boldsymbol{\nabla} \cdot \boldsymbol{v}\, d^3r = \int_{\partial\Omega} \boldsymbol{v} \cdot d\boldsymbol{S} \quad \text{and} \quad \int_S (\boldsymbol{\nabla} \times \boldsymbol{v}) \cdot d\boldsymbol{S} = \oint_{\partial S} \boldsymbol{v} \cdot d\boldsymbol{l}. \tag{191}$$

Here, Ω is a 3d region in $\mathbb{R}^3$ while S is a surface in $\mathbb{R}^3$.

In fact, (189) is also a generalization of the **fundamental theorem of calculus** for the integral of a 1-form over an interval $M = [a, b] \subset \mathbb{R}$: $\int_M f'(x)dx = \int_{\partial M} f = f(b) - f(a)$. Here f is a zero form and the boundary ∂M is the 0-dimensional disconnected manifold consisting of two points a and b. Integration of f on a zero dimensional manifold is a sum of the values of the function at the points weighted by the values of the volume form that specifies the orientation. What do we mean by an orientation on a 0-dimensional manifold? We mean the specification of a nonvanishing 0-form, i.e., a function. The choice of values of this function consistent with the

[57] It is also a generalization of Green's theorem, which is a planar version of Stokes' theorem. Suppose $\boldsymbol{v} = v_x \hat{x} + v_y \hat{y}$ is a vector field on the x-y plane and let S be a region in the plane bounded by the curve ∂S. Then $\boldsymbol{\nabla} \times \boldsymbol{v}$ has only a z component while $d\boldsymbol{S} = dx\, dy\, \hat{z}$ and $d\boldsymbol{l} = dx\, \hat{x} + dy\, \hat{y}$, so that Stokes' theorem becomes

$$\int_S (\partial_x v_y - \partial_y v_x) dx\, dy = \oint_{\partial S} (v_x dx + v_y dy). \tag{190}$$

fundamental theorem of calculus is $+1$ at b and -1 at a. In other words, the orientation dx of the interval gives ∂M an orientation ($+1$ at b and -1 at a) leading to the relative sign on the RHS.

As an **application of Stokes' theorem**, let us show that the standard **area form on the unit 2-sphere** is not exact. It is given by $\omega = \sin\theta \, d\theta \wedge d\phi$. It is closed $d\omega = 0$ as ω is a top degree form. To show it is not exact, we begin by noting that $\int_{S^2} \omega = 4\pi$ is the surface area of the unit sphere[58]. If $\omega = d\alpha$ for some 1-form α, then by Stokes' theorem, this integral must vanish, since S^2 has no boundary: $\int_{S^2} d\alpha = \int_{\partial S^2} \alpha = 0$. This would lead to a contradiction. Thus, ω cannot be exact. However, locally in a coordinate patch, it can be written as $\omega = d\alpha$ for $\alpha = -\cos\theta \, d\phi$ (local exactness is called the Poincaré lemma). The difficulty is that α cannot be smoothly extended to a 1-form on all of S^2. We have already met a symptom of this, the 1-form $d\phi$ has a length that diverges at the poles ($g^{-1}(d\phi, d\phi) = 1/\sin^2\theta$ where g^{-1} is the inverse of the standard round metric on the 2-sphere.) Thus ω is a closed but not exact 2-form on the sphere. It is therefore a nontrivial element of the **2nd de Rham cohomology group** of the sphere: $H^2(S^2)$.

• Application of Stokes' theorem to calculating **areas of planar regions**. Suppose D is some closed and bounded region in the x-y plane with boundary ∂D being a closed curve (more generally, the boundary could be the disjoint union of several closed curves). We will take the volume form on the plane to be $dx \wedge dy$. Then the area enclosed by D is

$$\text{Ar}(D) = \int_D dx \wedge dy. \tag{192}$$

To apply Stokes' theorem, we notice that on the plane, $dx \wedge dy$ is an exact 2 form $dx \wedge dy = d(xdy)$. Thus,

$$\text{Ar}(D) = \int_D dx \wedge dy = \int_{\partial D} xdy. \tag{193}$$

So we may evaluate areas via a line integral which is simpler than a surface integral.

In fact, this is done in mechanics if $x = p$ and $y = q$ are the momentum and position of a particle with one degree of freedom. The closed curve is a periodic trajectory on phase space. The line integral $\oint pdq$ is 2π times the **action of the trajectory** which is also the area enclosed by the periodic trajectory.

2.14 Laplace-Beltrami operator on a Riemannian manifold from variational principle

• The exterior derivative of a scalar function is a 1-form. To convert this to a vector field (the gradient) we need an inverse metric. In fact, on a Riemannian manifold (M, g), the gradient of a scalar function ϕ is defined as the vector field whose components in a coordinate basis are $(\boldsymbol{\nabla}\phi)^i = g^{ij}\partial_j\phi$. Here, g^{ij} are the entries of the inverse metric.

[58] Although polar coordinates do not cover all of the 2-sphere, they cover it except for a set of measure zero (any one longitude) which does not affect this integral.

• Can we generalize the concepts of the divergence of a vector field and the Laplacian of a scalar function, familiar from Euclidean space to general Riemannian manifolds? We will use the above concept of the gradient to achieve this generalization.

• The Laplace operator acting on a scalar function f on 3d Euclidean space $\mathbb{R}^3$ in Cartesian coordinates is familiar

$$\Delta f = \boldsymbol{\nabla}^2 f = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) f(x,y,z), \tag{194}$$

with a straightforward extension to $\mathbb{R}^n$. It appears in many places including the Laplace, wave, heat, Poisson and Schrödinger equations. In vector calculus, the Laplacian arises as the divergence of the gradient: $\boldsymbol{\nabla}^2 f = \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} f$. Depending on the symmetries of the setup, we often need the Laplacian in spherical, cylindrical or other curvilinear coordinate systems and it is challenging to remember the formula, although it is possible to obtain the formula by changing coordinates from Cartesian to, say, polar. It turns out that there is a generalization of the Laplacian to a Riemannian manifold M with metric tensor g. It is called the Laplace-Beltrami operator and admits a relatively simple expression in terms of the metric in any particular coordinate system. By using this formula we may obtain an explicit expression for the Laplacian in the desired coordinate system.

• We can obtain the above formula for the Laplacian using a variational principle for the Poisson equation. We will use the language of electrostatics, although this is not essential to the argument. Poisson's equation for the electrostatic potential ϕ is $\boldsymbol{\nabla}^2\phi = -\frac{\rho}{\epsilon_0}$ where ϵ_0 is a constant called the permittivity of free space and $\rho(x)$ is the electric charge density. The electric field is given by $\boldsymbol{E} = -\boldsymbol{\nabla}\phi$. In Cartesian coordinates x^i this follows from requiring that the 'energy' functional[59]

$$\begin{aligned} W[\phi] &= \frac{1}{2}\epsilon_0 \int \boldsymbol{E}^2 d^3x - \int \rho(x)\phi(x) d^3x \\ &= \frac{1}{2}\epsilon_0 \int \delta^{ij}(\partial_i\phi)(\partial_j\phi)\; d^3x - \int \rho(x)\phi(x) d^3x \end{aligned} \tag{195}$$

be stationary with respect to infinitesimal variations of ϕ. To see why this is the case, let us evaluate $W[\phi+\delta\phi]$ to linear order in $\delta\phi$:

$$\begin{aligned} W[\phi+\delta\phi] &= \frac{1}{2}\epsilon_0 \int \delta^{ij}\partial_i(\phi+\delta\phi)\partial_j(\phi+\delta\phi) d^3x - \int \rho(x)(\phi+\delta\phi) d^3x \\ &= W[\phi] + \frac{1}{2}\epsilon_0 \int \left[2\delta^{ij}\partial_i\phi\partial_j\delta\phi - \frac{2}{\epsilon_0}\rho\delta\phi\right] d^3x + \cdots \\ \Rightarrow \quad \delta W &= \frac{1}{2}\epsilon_0 \int [-2\partial_j(\delta^{ij}\partial_j\phi)\delta\phi - (2/\epsilon_0)\rho\delta\phi] d^3x + \mathcal{O}(\delta\phi)^2. \end{aligned} \tag{196}$$

We integrated by parts and assumed there are no contributions from boundaries. Thus,

$$\delta W = 0 \quad \Rightarrow \quad \delta^{ij}\partial_i\partial_j\phi = -\rho/\epsilon_0 \quad \text{or} \quad \boldsymbol{\nabla}^2\phi = -\rho/\epsilon_0. \tag{197}$$

[59] The first term may be interpreted as the energy in the electric field and the second term as the energy due to the charges.

To obtain the Laplacian on a (pseudo-)Riemannian manifold, we will generalize the energy functional (195) to a manifold with metric g and then extremize it. Suppose we work on a coordinate patch with coordinates x^i and metric tensor g_{ij}. The volume element d^3x is now generalized to the Riemannian volume form $\sqrt{g}dx^1 \wedge \cdots \wedge dx^n \equiv \sqrt{g}d^nx$ where $g = |\det g_{ij}|$. The square of the gradient $\delta^{ij}\partial_i\phi\partial_j\phi$ is generalized to the square of the length of the 1-form $d\phi$, i.e., $g^{ij}\partial_i\phi\partial_j\phi$. Then W becomes

$$W[\phi] = \frac{1}{2}\epsilon_0 \int g^{kl}(\partial_k\phi)\ (\partial_l\phi)\sqrt{g}\, d^nx - \int \rho(x)\phi(x)\sqrt{g}\, d^nx. \tag{198}$$

Considering a small variation $\phi \mapsto \phi + \delta\phi$, we ask that the first variation of W vanish. Integrating by parts,

$$\delta W = -\epsilon_0 \int \partial_k \left(g^{kl}(\partial_l\phi)\sqrt{g}\right) \delta\phi\, d^nx - \int \rho(x)\delta\phi\sqrt{g}d^nx. \tag{199}$$

Now $\delta W = 0$ for any $\delta\phi$ implies

$$\epsilon_0\partial_k\left(\sqrt{g}g^{kl}\partial_l\phi\right) = -\sqrt{g}\rho \quad \text{or} \quad \frac{1}{\sqrt{g}}\partial_k\left(\sqrt{g}g^{kl}\partial_l\phi\right) = -\frac{\rho}{\epsilon_0}. \tag{200}$$

Comparing with $\Delta\phi = -\frac{\rho}{\epsilon_0}$ we read off a formula for the Laplacian

$$\Delta\phi = \frac{1}{\sqrt{g}}\partial_k\left(\sqrt{g}g^{kl}\partial_l\phi\right). \tag{201}$$

This is called the Laplace operator or **Laplace-Beltrami operator** on scalar fields on a Riemannian manifold with metric tensor g.

- **Divergence of a vector field.** Recall that $g^{kl}\partial_l\phi$ are the components of the gradient of ϕ. Thus, if we view the Laplacian as $\Delta\phi = \text{div grad}\,\phi$, then we may read off a formula for the divergence of a vector field

$$\text{div}\, v = \boldsymbol{\nabla}\cdot v = \frac{1}{\sqrt{g}}\partial_i(\sqrt{g}v^i). \tag{202}$$

2.15 Hodge dual and volume form duals

- Consider a Riemannian manifold M of dimension n with coordinates x^i in a patch with metric g_{ij} and inverse g^{ij}. We have seen that the spaces of p-forms and $(n-p)$-forms at a point of M have the same dimension $\binom{n}{p}$. Similarly, the spaces of totally antisymmetric rank-p and rank-$(n-p)$ contravariant tensors also have the same dimension $\binom{n}{p}$. A rank-p antisymmetric contravariant tensor is called a p-vector. This suggests that there may be isomorphisms between these four spaces. On a Riemannian manifold, there are such canonical linear isomorphisms (called dualities), which are formulated in terms of the volume form dual and Hodge dual.
- The volume form dual maps a p-vector to an $(n-p)$-form. It can also be used to map a p-form to an $(n-p)$-vector. The Hodge dual maps a p-form to an $(n-p)$-form or a p-vector to an $(n-p)$-vector.

• The metric volume form on M is the n-form

$$\omega = \frac{1}{n!}\omega_{ij\dots k}dx^i \wedge dx^j \wedge \cdots \wedge dx^k = \frac{1}{n!}\sqrt{g}\,\epsilon_{ij\dots k}dx^i \wedge dx^j \wedge \cdots \wedge dx^k \quad (203)$$

where $g = \det g_{ij}$. The components of the inverse metric volume form are given by:

$$\omega^{ij\cdots k} = \frac{1}{\sqrt{g}}\,\epsilon^{ij\cdots k} \quad (204)$$

Here, $\epsilon_{12\dots 3} = (\epsilon^{12\dots 3})^{-1} = 1$ and the ϵ symbol is totally antisymmetric. These formulae are valid in any coordinate system. Note that $\epsilon_{ij\dots k}\epsilon^{ij\cdots k} = n!$.

• **Volume form dual.** The volume form dual $*$ of a q-vector T (antisymmetric contravariant tensor of rank q) gives an $(n-q)$-form with components:

$$(*T)_{l\dots m} = \frac{1}{q!}\omega_{ij\dots kl\dots m}T^{ij\dots k}. \quad (205)$$

In particular, the volume form dual of a vector is simply the contraction $*v = \omega(v)$. Similarly, one can get an $(n-p)$-vector from a p-form α using the inverse volume form:

$$(*\alpha)^{l\dots m} = \frac{1}{p!}\omega^{ij..kl\dots m}\alpha_{ij\dots k} \quad (206)$$

Moreover, for a p-form α we have

$$**\alpha = (-1)^{p(n-p)}\alpha. \quad (207)$$

Similarly for a q-vector T we have,

$$**T = (-1)^{q(n-q)}T. \quad (208)$$

In other words, taking the volume form dual twice in succession is the identity operation up to a possible sign.

• **Hodge dual.** The Hodge dual $\star$ of a p-form gives an $(n-p)$-form by first taking the volume form dual of the p-form to get an $(n-p)$-vector and then lowering the $(n-p)$ indices using the metric. Thus, $\star = g*$ is a composition of the metric isomorphism with the volume form dual. Explicitly,

$$(\star\alpha)_{i_1\cdots i_{n-p}} = g_{i_1 j_{p+1}}\cdots g_{i_{n-p}j_n}\frac{1}{p!}\frac{1}{\sqrt{g}}\epsilon^{j_1\cdots j_n}\alpha_{j_1\cdots j_p}. \quad (209)$$

We may also first raise the indices of the p-form α to get a p-vector and then take its volume form dual to get an $(n-p)$-form. The two approaches give the same result, so, $\star = *g = g*$. Moreover, if (M, g) is Riemannian, we have the property that for a p-form α,

$$\star\star\alpha = (-1)^{p(n-p)}\alpha. \quad (210)$$

• **Hodge dual on $\mathbb{R}^3$.** Let us work out the Hodge dual of coordinate 1-forms on 3d Euclidean space with Cartesian coordinates. Since Hodge duality is a linear map, it is determined by its action on basis forms. For the Euclidean metric $g_{ij} = \delta_{ij}$, $\det g = 1$. Moreover, $\omega_{ijk} = \sqrt{g}\epsilon_{ijk} = \epsilon_{ijk}$ and $\omega^{ijk} = \frac{1}{\sqrt{g}}\epsilon^{ijk} = \epsilon^{ijk}$ with $\epsilon^{ijk} = \epsilon_{ijk}$ being numerically the same. Suppose $\alpha = \alpha_i dx^i$, then its volume form dual is the 2-vector $(*\alpha)^{jk} = \omega^{ijk}\alpha_i$ and its Hodge dual is the 2-form

$$(\star\alpha)_{lm} = \omega^{ijk}\alpha_i g_{lj} g_{mk} = \delta_{lj}\delta_{mk}\epsilon^{ijk}\alpha_i. \tag{211}$$

Taking $\alpha = dx^1$ we find $(\star dx^1)_{23} = \epsilon^{123} = 1$. Thus, $\star dx^1 = dx^2 \wedge dx^3$. In a similar way,

$$\star\, dx^2 = dx^3 \wedge dx^1 \quad \text{and} \quad \star\, dx^3 = dx^1 \wedge dx^2. \tag{212}$$

Next we obtain the volume form dual and Hodge dual of the zero form 1. Its volume form dual is a 3-vector:

$$(*1)^{ijk} = \frac{1}{0!}\omega^{ijk} = \epsilon^{ijk}. \tag{213}$$

The Hodge dual is the Riemannian volume 3-form:

$$(\star 1)_{ijk} = g_{il} g_{jm} g_{kn}\epsilon^{lmn} = \epsilon_{ijk} \quad \Rightarrow \quad \star 1 = dx^1 \wedge dx^2 \wedge dx^3. \tag{214}$$

For $\mathbb{R}^3$, $p(n-p)$ is always even, so $\star\star$ is the identity. Thus

$$\star\,(dx^1 \wedge dx^2) = dx^3, \quad \star(dx^2 \wedge dx^3) = dx^1, \quad \star(dx^3 \wedge dx^1) = dx^2. \tag{215}$$

• **Hodge dual on $\mathbb{R}^2$.** The Hodge dual of the 0-form 1 is the volume form:

$$(*1)^{ij} = \epsilon^{ij} \quad \Rightarrow \quad (\star 1)_{ij} = g_{ik} g_{jl}\epsilon^{kl} = \epsilon_{ij} \quad \Rightarrow \quad \star 1 = dx^1 \wedge dx^2. \tag{216}$$

Conversely, $\star(dx^1 \wedge dx^2) = 1$ since $p(n-p) = 0$ for $p = 2$.

On the other hand, suppose $\alpha = \alpha_i dx^i$ is a 1-form. Then

$$(*\alpha)^j = \epsilon^{ij}\alpha_i \quad \Rightarrow \quad (\star\alpha)_k = g_{kj}(*\alpha)^j = g_{kj}\epsilon^{ij}\alpha_i. \tag{217}$$

It follows that

$$(\star\alpha)_1 = g_{11}\epsilon^{21}\alpha_2 = -\alpha_2 \quad \text{and} \quad (\star\alpha)_2 = g_{22}\epsilon^{12}\alpha_1 = \alpha_1 \tag{218}$$

so that

$$\star(\alpha_1 dx^1 + \alpha_2 dx^2) = -\alpha_2 dx^1 + \alpha_1 dx^2 \quad \Rightarrow \quad \star dx^1 = dx^2, \ \star dx^2 = -dx^1. \tag{219}$$

• **Maxwell's equations** may be written as $dF = 0$ and $d \star F = j$ where j is the current density 3-form and F is the Faraday 2-form ($F = dA$). We may convert the current density 3-form to a vector by taking its volume form dual (or convert it to a current 1-form by taking its Hodge dual).

• **Inner product between two p-forms:** We may use the Hodge dual to define an inner product (symmetric bilinear operation) on the space of p-forms. Suppose α and β are both p-forms. To get a number, we would like to define an n-form and integrate it over the manifold. There is an obvious n-form: $\alpha \wedge \star\beta$. Thus, we define the inner product as

$$(\alpha, \beta) = \int \alpha \wedge \star\beta. \tag{220}$$

• For example, in electromagnetism, $F = dA$ and the action is proportional to the inner product of F with itself, $S = \int F \wedge \star F$. When written out in terms of the electric and magnetic fields, the integrand involves $\boldsymbol{E}^2 - \boldsymbol{B}^2$, which is proportional to the Lagrangian density of the Maxwell field.

• An explicit formula will show that this inner product is symmetric under interchange of α and β. Consider two p-forms α and β:

$$\alpha = \frac{1}{p!}\alpha_{ij\cdots k} dx^i \wedge dx^j \wedge \cdots dx^k \quad \text{and} \quad \beta = \frac{1}{p!}\beta_{lm\cdots n} dx^l \wedge dx^m \wedge \cdots dx^n. \tag{221}$$

Then the inner product can be expressed as

$$(\alpha, \beta) = \int \alpha \wedge \star\beta = \int \frac{1}{p!}\alpha_{ij\cdots k}\beta^{ij\cdots k}\sqrt{g} dx^1 \wedge dx^2 \wedge \cdots dx^n. \tag{222}$$

This formula shows that $(\alpha, \beta) = (\beta, \alpha)$. To get the formula on the right, we first need to take the Hodge dual of β to get an $(n-p)$ form. So, first raising all the indices on β we get a p-vector

$$\beta^{pq\cdots r} = g^{pl}g^{qm}\cdots g^{rn}\beta_{lm\cdots n}. \tag{223}$$

Taking its volume form dual gives an $(n-p)$ form:

$$\begin{aligned} (\star\beta)_{st\cdots u} &= \frac{1}{p!}\sqrt{g}\epsilon_{pq\cdots rst\cdots u}\beta^{pq\cdots r} \\ \Rightarrow \quad \star\beta &= \frac{1}{(n-p)!p!}\sqrt{g}\epsilon_{pq\cdots rst\cdots u}\beta^{pq\cdots r} dx^s \wedge dx^t \wedge \cdots dx^u. \end{aligned} \tag{224}$$

Now, we take the wedge product between α and $\star\beta$:

$$\begin{aligned} \alpha \wedge (\star\beta) &= \frac{1}{(p!)^2}\frac{1}{(n-p)!}\tilde{\alpha}_{ij\cdots k}\beta^{pq\cdots r}\epsilon_{pq\cdots rst\cdots u} \\ &\quad \sqrt{g}\, dx^i \wedge dx^j \wedge \cdots dx^k \wedge dx^s \wedge dx^t \wedge \cdots \wedge dx^u \\ &= \frac{1}{(p!)^2}\frac{1}{(n-p)!}\tilde{\alpha}_{ij\cdots k}\beta^{pq\cdots r}\epsilon_{pq\cdots rst\cdots u}\epsilon^{ij\cdots kst\cdots u} \\ &\quad \sqrt{g}\, dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n \end{aligned} \tag{225}$$

Here we use the identity $dx^i \wedge dx^j \wedge \cdots \wedge dx^k = \epsilon^{ij\cdots k} dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n$. Moreover, using the formula for contraction of ϵ with itself (228), we get,

$$\begin{aligned} \alpha \wedge (\star\beta) &= \frac{1}{(p!)^2}\frac{1}{(n-p)!}\alpha_{ij\cdots k}\beta^{pq\cdots r}\ ((n-p)!\ \delta^{ij\cdots k}_{pq\cdots r})\sqrt{g}\, dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n \\ &= \frac{1}{(p!)^2}\alpha_{ij\cdots k}\beta^{pq\cdots r}\ p!\ \delta^i_{[p}\delta^j_q \cdots \delta^k_{r]}\sqrt{g}\, dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n \end{aligned}$$

$$= \quad \frac{1}{p!}\alpha_{ij\cdots k}\beta^{pq\cdots r} \;\; \delta^i_p\delta^j_q\cdots\delta^k_r\sqrt{g}\, dx^1\wedge dx^2\wedge\cdots\wedge dx^n \qquad (226)$$

We dropped the antisymmetrization on the δ's since it is contracted with the antisymmetric contravariant tensor β. Thus, we get

$$\alpha\wedge(\star\beta) = \frac{1}{p!}\alpha_{ij\cdots k}\beta^{ij\cdots k}\sqrt{g}dx^1\wedge dx^2\wedge\cdots dx^n. \qquad (227)$$

This n-form can be integrated to give the inner product between two forms of the same degree.

• **Contraction of** ϵ: In n-dimensions, if we contract $(n-p)$ indices $(i\cdots k)$ of the ϵ tensor with itself we get

$$\epsilon_{i\cdots kl\cdots m}\epsilon^{i\cdots kq\cdots r} = (n-p)!\delta^{q\cdots r}_{l\cdots m} = (n-p)!\,p!\,\delta^q_{[l}\cdots\delta^r_{m]} \qquad (228)$$

where $[l,..m]$ denotes antisymmetrization. For example,

$$\epsilon^{ijk}\epsilon_{ilm} = (\delta^j_l\delta^k_m - \delta^j_m\delta^k_l) \quad \text{and} \quad \epsilon^{ijk}\epsilon_{ijl} = 2\delta^k_l. \qquad (229)$$

3 Groups, Lie groups and their Lie algebras

3.1 Some references on groups and Lie algebras

1. N Mukunda and S Chaturvedi, *Continuous Groups for Physicists.*

2. Bernard Schutz, *Geometrical Methods of Mathematical Physics.*

3. Govind Krishnaswami, *Classical Mechanics: From Particles to Continua and Regularity to Chaos*, Appendix B.

4. H F Jones, *Groups, Representations and Physics.*

3.2 Concept and definition of a group

A group is a mathematical construct that, among other things, helps us express and work with symmetries. Groups occur in various parts of physics such as crystallography, atomic physics, relativity and particle physics. They help to recognize and organize patterns, but can also enter dynamical principles that constrain or determine the nature of forces. For instance, the angular distribution of possible locations of an electron in a hydrogen atom can be understood using the spherical symmetry of the electric potential felt by the electron. On the other hand, the strong nuclear force among quarks and gluons is determined by a 'gauge principle' based on a so-called color symmetry group. In mechanics, groups typically arise as families of symmetry transformations among states or configurations or solutions of the equations of a system. For example, rotations act on the possible locations of a planet in the Kepler problem while the $x\to -x$ reflection acts as a symmetry of an even harmonic oscillator potential $V(x) = \frac{1}{2}kx^2$ felt by a particle attached to a spring. However, it is advantageous to separate the algebraic concept of a group from its action on a space. Thus, we will begin by defining an ‘abstract’ group and later discuss how it may be realized via an action on an auxiliary space like the state space of a mechanical system. Precisely, a group G is a set of elements $g, h, k, \ldots$ among which a law of composition $G \times G \to G$ is defined: if $g, h \in G$, then their product or composition $gh \in G$. The product must satisfy the following properties. (i) It must be associative, i.e., $g(hk) = (gh)k$ for any $g, h, k \in G$. (ii) G must include an identity element e (sometimes denoted 1 or I) with $ge = eg = g$ for any $g \in G$. (iii) Every element g must have a two-sided inverse g^{-1}, i.e., $gg^{-1} = g^{-1}g = e$. Useful consequences are $(gh)^{-1} = h^{-1}g^{-1}$ and the cancellation law: if $gh = gk$ then $h = k$.

3.3 Cardinality, discrete and continuous groups

The number of elements $|G|$ in a group G is called its order or cardinality. A group of finite order is called a finite group. The ‘trivial’ group has just one element, the identity: $G = \{1\}$ with $1 \cdot 1 = 1$. The set $C_2 = \{1, -1\}$ under multiplication is a group of order two. While 1 is the identity, $(-1)(-1) = (-1)^2 = 1$. We say that -1 generates C_2 since -1 and $(-1)^2$ account for all the distinct elements. C_2 is called the cyclic group of order two. Notice that ± 1 are the two square-roots of unity. More generally, for $n = 1, 2, 3, \ldots$, we have the (multiplicative) cyclic group of order n consisting of the n^{th} roots of unity $\{1, e^{2\pi i/n}, e^{4\pi i/n}, \ldots, e^{2(n-1)\pi i/n}\}$. It is generated by $e^{2\pi i/n}$ and we write $C_n = \langle e^{2\pi i/n} \rangle$. For $n = 1, 2, 3, \ldots$, the set $\mathbb{Z}_n = \{0, 1, \cdots, n-1\}$ with composition given by addition modulo n (e.g., $2 + 3 \equiv 1 \pmod 4$) is also a cyclic group of order n. The identity element is 0 and 1 is its generator. Note that $1 + 1 + \cdots + 1$ (n summands) $= n1 \equiv 0 \pmod n$. We will soon see that $\mathbb{Z}_n$ and C_n are different presentations of the same group: up to ‘isomorphism’ there is just one cyclic group of a given order. Infinite groups could be discrete (like the additive group of integers $\mathbb{Z}$ or the infinite cyclic group) or continuous (like the multiplicative group of complex numbers of unit magnitude). The latter group is denoted $U(1)$ and its elements may be represented as $z(\theta) = e^{i\theta}$ for a real angle θ which is defined modulo 2π (see Fig. 7). Composition is given by $z(\theta_1)z(\theta_2) = e^{i(\theta_1+\theta_2)} = z(\theta_1 + \theta_2)$.

3.4 Subgroup.

A subset H of a group G is called a *subgroup* if it satisfies the group axioms with respect to the operations inherited from G [see Prob. **??**]. The identity subgroup $H = \{e\}$ and $H = G$ are subgroups of any group G. Examples: (i) $C_2 = \{\pm 1\}$ and more generally C_n are subgroups of $U(1)$. (ii) Given any element g of a group G, it generates a (cyclic) subgroup, namely the set of its powers $\langle g \rangle = \{g^0 = e, g, g^{-1}, g^2, g^{-2}, \cdots\}$. If there is a smallest positive integer n such that $g^n = e$, then $\langle g \rangle$ is essentially the same as (i.e., isomorphic to) a cyclic group of order n and otherwise it is an infinite cyclic group. (iii) Every finite group may be realized as a subgroup of a group of permutations (see Prob. **??**).

3.5 Group homomorphisms.

Given a pair of groups, a map $\phi : G \to G'$ is called a *homomorphism* if it preserves products: $\phi(g_1 g_2) = \phi(g_1)\phi(g_2) \; \forall \; g_1, g_2 \in G$. If ϕ preserves products then (see Prob. **??**) ϕ maps the identity in G to that in G' and maps inverses to inverses: $\phi(g^{-1}) = \phi(g)^{-1}$. The homomorphic image $\phi(G) \subseteq G'$ is a subgroup of G'. A homomorphism $\phi : G \to G'$ is an isomorphism if it is bijective (1-1 and onto, and hence invertible). Two groups G and G' are *isomorphic* (denoted $G \cong G'$) if there is an isomorphism between them. Isomorphic groups are algebraically identical but could arise or be presented differently. E.g., C_n and $\mathbb{Z}_n$ are isomorphic, with the isomorphism mapping the generators to each other: $\phi(e^{2\pi i/n}) = 1$, so that $\phi(e^{2\pi ij/n}) = j$ for $j = 0, 1, \cdots, n-1$. The group of unimodular complex numbers $U(1)$ is isomorphic to that of 2 × 2 orthogonal matrices (real A with $A^t A = I$) with unit determinant ($SO(2)$). Composition is given by matrix multiplication. Its elements are $A(\theta) = (\cos\theta, \sin\theta| - \sin\theta, \cos\theta)$ for a real angle θ defined modulo 2π. The isomorphism maps $z = e^{i\theta}$ to $A(\theta)$. Verify that under matrix multiplication, $A(\theta_1)A(\theta_2) = A(\theta_1 + \theta_2)$.

3.6 Isomorphisms and automorphisms

An isomorphism ϕ from a group G to itself is called an *automorphism*. Every group has the identity or trivial automorphism defined by $\phi(g) = g$ for all $g \in G$. $C_2 = \{1, -1\}$ has no nontrivial automorphism since we cannot define $\phi(1) = -1$. Verify that C_3 has just one nontrivial automorphism given by $\phi(1) = 1, \phi(\omega) = \omega^2$ and $\phi(\omega^2) = \omega$ where $\omega = e^{2\pi i/3}$. Since an automorphism must preserve the algebraic structure, it must take a generator to another generator: we check that both ω and ω^2 are generators of C_3. $C_4 = \{1, i, -1, -i\}$ also has one nontrivial automorphism: it exchanges the two generators: $\phi(1) = 1$, $\phi(i) = -i$, $\phi(-1) = -1$ and $\phi(-i) = i$.
• Complex conjugation $z(\theta) \mapsto z^*(\theta)$ (taking $e^{i\theta}$ to $e^{-i\theta}$) is an automorphism of $U(1)$: check this. However, rotation of elements of $U(1)$ is generally not an automorphism since it does not take the identity to the identity. The antipodal map $z(\theta) \mapsto -z(\theta)$ is also not an automorphism of $U(1)$ for the same reason.

3.7 Conjugation and conjugacy classes

Given a group G, we say that $k \in G$ is conjugate to $h \in G$ if $k = ghg^{-1}$ for some[60] $g \in G$. Conjugation $\phi_g(h) = ghg^{-1}$ by a fixed element g defines an automorphism of G. It is called an *inner automorphism*. It is a homomorphism since $\phi_g(h_1 h_2) = gh_1 h_2 g^{-1} = gh_1 g^{-1} gh_2 g^{-1} = \phi_g(h_1)\phi_g(h_2)$. It is $1-1$ since $\phi_g(h_1) = \phi_g(h_2)$ implies $gh_1 g^{-1} = gh_2 g^{-1}$ whence $h_1 = h_2$. It is surjective since given any $k \in G$ we can always find an $h \in G$ such that $\phi_g(h) = k$, in fact $h = g^{-1}kg$. What is more, the inverse of ϕ_g is just $\phi_{g^{-1}}$.

The *conjugacy class* of h is the set $C_h = \{ghg^{-1} | g \in G\}$. The identity element is always in a conjugacy class by itself $C_e = \{e\}$. Conjugacy is an equivalence

[60] If g works, so do gh^n for $n \in \mathbb{Z}$ and more generally gg' for any g' that commutes with h.

relation[61]. This implies G is a disjoint union of conjugacy classes. The conjugacy classes of a pair of elements h and k are either the same or disjoint: they cannot partially overlap.

3.8 Abelian and nonabelian groups

The nature of conjugation and conjugacy classes are related to the notion of a commutative group. We begin by defining the *group commutator* of a pair of elements as $[g,h] = ghg^{-1}h^{-1}$. The commutator measures the extent to which gh and hg differ. If gh and hg are the same, then $[g,h] = e$ and they are said to commute. A group is called *abelian* or *commutative* if $[g,h] = e$ or $gh = hg$ or $ghg^{-1} = h$ for all $g, h \in G$. Otherwise, it is nonabelian. Evidently, a group is abelian iff all conjugacy classes are singleton sets or equivalently, if every inner automorphism is the identity. Roughly, conjugacy classes get longer the more nonabelian a group is. Only the first 4 groups below are abelian.

3.9 Examples

There are many elementary examples of groups that arise in interesting ways, some of which we have met: (i) the multiplicative group $C_2 = \{1, -1\}$ consisting of the identity and reflection symmetry $x \to -x$ of an even potential $V(x)$ in one dimension, (ii) the cyclic group C_5 of order 5, of rotational symmetries[62] of a regular pentagon (if one includes reflection symmetries, one obtains the dihedral group of order 10), (iii) the groups $\mathbb{R}^3$ and $\mathbb{R}$ of translations of 3d Euclidean space and time, (iv) the group $SO(2)$ of rotational symmetries of a circle or an axisymmetric (cylindrically symmetric) potential, (v) the group $SO(3)$ of proper rotations of 3d Euclidean space, (vi) the group $O(3)$ of rotations and reflections of $\mathbb{R}^3$, (vii) the Galilei group and (viii) the groups S_2 and S_3 of permutations of two and three objects encountered in Footnote 46 of Appendix 2.9 and Footnote 50 of Appendix 2.10.

3.10 Lie groups

While examples (i), (ii) and (viii) are discrete groups (with finitely many elements), the rest are examples of continuous groups, where the group elements form continuous families and can be used to model continuous symmetries. Historically, discrete groups arose, in part, in modeling discrete symmetries of algebraic equations, while continuous groups arose via continuous symmetries of differential equations. Prominent among continuous groups are Lie groups, named after the Norwegian mathematician Sophus Lie. A Lie group is a group which is also a differentiable manifold,

[61] Conjugacy is reflexive: $h = ehe^{-1}$ (h is conjugate to h), symmetric: $h = g'kg'^{-1}$ where $g' = g^{-1}$ (h is conjugate to k if k is conjugate to h) and transitive: k conjugate to h and h conjugate to l implies k conjugate to l. A binary relation with these properties is called an equivalence relation. It ensures that conjugacy classes either coincide or do not overlap. For instance, transitivity implies that C_{h_1} and C_{h_2} cannot have a 'partial' overlap.

[62] Cyclic and dihedral groups are *point groups* in 2d. They are symmetries of regular polygons and molecules with a fixed point and are discrete subgroups of the orthogonal group. *Space groups* are symmetries of an infinite crystal and include discrete translations.

with the group operations of composition $(g,h) \mapsto gh$ and inversion $g \mapsto g^{-1}$ being smooth maps from $G \times G \to G$ and $G \to G$ (inversion must be a diffeomorphism). The elements of the group are points on the manifold. The manifold is called the group manifold.

• The dimension of a Lie group G is the dimension of the corresponding group manifold. Note that the Cartesian product $G \times G$ inherits a $2(\dim G)$-dimensional manifold structure from G upon using ordered pairs of charts and transition functions. The concept of smooth maps is as introduced in Appendix 2.2.

• The group $U(1)$ of complex numbers of unit modulus ($|z| = 1$, $z = e^{i\theta}$) is a one-dimensional Lie group. The corresponding group manifold is the circle S^1: elements of the group are points on the unit circle. Group composition $e^{i\theta_3} = e^{i\theta_1} e^{i\theta_2}$ is a smooth map from $S^1 \times S^1$ to S^1 taking $(e^{i\theta_1}, e^{i\theta_2})$ to $e^{i(\theta_1+\theta_2)}$. Inversion is also a smooth map taking $z \to 1/z$. It is a 1-1 onto smooth map from the circle to the circle with smooth inverse, so it is a diffeomorphism.

• The additive group of real numbers is a Lie group with group manifold $\mathbb{R}^1$.

• The additive group of vectors in n-dimensional Euclidean space is a Lie group. The inverse of a vector $\boldsymbol{v}$ is $-\boldsymbol{v}$. The composition of two vectors is their sum $\boldsymbol{v} + \boldsymbol{w}$. The corresponding group manifold is $\mathbb{R}^n$.

3.11 Matrix Lie groups

Natural examples of Lie groups are the matrix groups $GL_n(\mathbb{R})$ and $GL_n(\mathbb{C})$ of invertible $n \times n$ real and complex matrices with composition and inversion given by matrix multiplication and inversion (Nb. GL stands for general linear). For instance, $GL_n(\mathbb{R})$ is an n^2-dimensional submanifold (in fact, an open subset) of the space $\mathbb{R}^{n^2}$ of all $n \times n$ real matrices. Matrix multiplication and inversion may be shown to be smooth maps. Since matrix multiplication is generally noncommutative, these groups for $n > 1$ are nonabelian. Other examples of 'classical' Lie groups[63] such as the special linear, orthogonal, symplectic and unitary groups arise as closed subgroups of $GL_n(\mathbb{R})$ and $GL_n(\mathbb{C})$. The special linear groups $SL_n(\mathbb{R})$ and $SL_n(\mathbb{C})$ consist of invertible matrices with unit determinant. The orthogonal and special orthogonal groups $O(n)$ and $SO(n)$ consist of orthogonal matrices ($A^t A = I$) in $GL_n(\mathbb{R})$ and $SL_n(\mathbb{R})$ respectively. Similarly, the unitary and special unitary groups $U(n)$ and $SU(n)$ consist of unitary matrices ($U^\dagger U = I$) in $GL_n(\mathbb{C})$ and $SL_n(\mathbb{C})$. The symplectic group $Sp(2n, \mathbb{R})$ consists of $2n \times 2n$ matrices M that preserve the canonical symplectic structure: $M^t \omega M = \omega$ where $\omega = (0, -I | I, 0)$ and I is the $n \times n$ identity matrix. They are the linear canonical transformations of the phase space $\mathbb{R}^{2n}$. Later, we will discuss some basic properties of Lie groups in the context of the orthogonal group.

3.12 Transformation group acting on a set

Groups often arise as families of (often symmetry) transformations of a space M such as a configuration or phase space or the space of solutions of equations of motion. For instance, the rotation group $SO(3)$ acts on the configuration space $\mathbb{R}^3$ of

[63] 'Classical' here is used to mean that these were among the first Lie groups to be studied.

a particle in a spherically symmetric potential by rotating the radius vector about the force center: $\boldsymbol{r} \mapsto R\boldsymbol{r}$ for $R \in SO(3)$. Evidently, such a 'transformation group' is to be regarded as an action of an abstract group G on a set M. Precisely, an action of G on M is a map from $G \times M \to M$ taking $m \in M$ to $g \cdot m \in M$ such that $e \cdot m = m$ and $g \cdot (h \cdot m) = (gh) \cdot m$ for all $g, h \in G$ and $m \in M$.

• **Orbit.** The set of points that a given point $m \in M$ can be mapped to, $\mathcal{O}_m = \{g \cdot m | g \in G\}$ is called the orbit of m under the action of G. The orbit of a vector $\boldsymbol{v} \in \mathbb{R}^3$ under the action of the rotation group SO(3) is the sphere centered at the origin with radius $|\boldsymbol{v}|$. The orbit of the origin in the origin alone.

• **Transitive action.** The action is said to be *transitive* if every point of M can be mapped to every other point of M by the action of some group element. In other words, the action is transitive if M is the orbit of any of its points. The action of rotations on $\mathbb{R}^3$ is not transitive: for instance, the origin cannot be mapped to any other point by a rotation. On the other hand, translations act transitively on $\mathbb{R}^3$: any point can be translated to any other point.

• **Stabilizer:** The stabilizer of a point $m \in M$ is the subset of elements of G that take m to itself. Given any $m \in M$, show that the stabilizer of m is a subgroup of G. For the action of SO(3) on $\mathbb{R}^3$, the stabilizer of the origin is the whole of SO(3). The stabilizer of a nonzero vector $\boldsymbol{v} \in \mathbb{R}^3$ consists of rotations about $\boldsymbol{v}$, this forms the subgroup SO(2) of rotations in the plane perpendicular to $\boldsymbol{v}$.

• The stabilizer of m is also called the **isotropy subgroup** or **little group** of m. The Lorentz group $SO(3,1)$ acts on Minkowski space via rotations, boosts and their compositions. It is the group of linear transformations that preserve the Minkowskian inner product between 4-vectors. The action is not transitive. The little groups of time-like, space-like and light-like momentum 4-vectors are of interest in relativistic physics. They are isomorphic to SO(3), SO(2,1), and the Euclidean group E(2).

3.13 Coset spaces

The idea of a group acting on itself is extremely useful and can be used to 'subdivide' a group. Given a subgroup H of G, we may consider all its left translates, i.e., the subsets $gH = \{gh | h \in H\}$ where g ranges over elements of G. The subsets gH are called left cosets of G by H. Note that distinct elements of G may produce the same coset. For instance, all elements $h_1, h_2, \ldots$ of H give rise to the same coset[64] $h_1 H = h_2 H = eH = H$. Moreover, all cosets have the same cardinality as H. In fact, the elements of the list gH are all distinct. Additionally, two cosets are either the same or disjoint: $g_1 H = g_2 H$ or $g_1 H \cap g_2 H = \{\}$. The former happens if $g_1 = g_2 h$ for some $h \in H$ and the latter happens if there is no such $h \in H$. Thus, a group is a disjoint union of (left) cosets[65]. The set of left cosets forms the left coset space

[64] In formulae such as $h_1 H = h_2 H$ we mean that the two sets are the same, although the order of elements in the two lists may differ.

[65] (Left) cosets may be interpreted as equivalence classes. For any two elements of G, define the relation $g \sim g'$ if there is an $h \in H$ such that $gh = g'$. This relation is reflexive ($g \sim g$ since $ge = g$), symmetric ($g \sim g' \Rightarrow g' \sim g$ since $gh = g'$ implies $g'h^{-1} = g$) and transitive ($g \sim g'$ and $g' \sim g''$ implies $g \sim g''$ since $gh = g'$ and $g'h' = g''$ implies $ghh' = g''$) and therefore an equivalence relation. Evidently, the

denoted G/H and pronounced 'G mod H'. Similarly, the right translates Hg of the subgroup by elements of G leads to the right coset space, denoted $H\backslash G$. It is often convenient to pick an element from each coset and use it as a representative for the coset. For example, the even integers $2\mathbb{Z}$ form a subgroup of the additive group of integers $\mathbb{Z}$. There are only two cosets: the sets of even and odd integers: $2\mathbb{Z}$ and $2\mathbb{Z}+1$ (left and right cosets coincide since addition is commutative). In this case, we could pick 0 and 1 as the two coset representatives.

3.14 Normal subgroup and quotient or factor group

In general, neither the space of left nor right cosets is a group. However, they acquire the structure of a group if H is a so-called normal subgroup of G. Precisely, N is a normal subgroup (denoted $N \lhd G$) if each left coset gN is also a right coset Ng for the same $g \in G$. Such an N is also called an invariant subgroup as it is one that is invariant under conjugation: $gNg^{-1} = N$ for any $g \in G$. It is then easy to see that the set of left (or right) cosets of G by N is a group with identity given by the coset $eN = N$. Indeed, the group multiplication and inversion (for left cosets) are given by

$$(gN)\,(g'N) = (gg')N \quad \text{and} \quad (gN)^{-1} = g^{-1}N \tag{230}$$

Here, we used the formulae: $gNg'N = gg'NN = gg'N$ and $(gN)^{-1} = N^{-1}g^{-1} = Ng^{-1} = g^{-1}N$ since $N = N^{-1}$ on account of it being closed under inverses. G/N is called the quotient group or factor group. Some elementary properties are worth noting. (i) Every subgroup of an abelian group is a normal subgroup. (ii) If G is finite, then the cardinality of the coset space G/H is $|G|/|H|$ (Lagrange's theorem). (iii) If N is an invariant Lie subgroup of the Lie group G, then the dimension of the coset space G/N is the difference between the dimensions of G and N. (iv) The kernel K (inverse image $\phi^{-1}(e')$ of the identity $e' \in G'$) of a group homomorphism $\phi : G \to G'$ is always a normal subgroup of G (see Prob. **??**) and the image $\phi(G)$ is isomorphic to G/K. (v) The center $Z(G)$ of a group G, consisting of elements that commute with all other elements, is an abelian normal subgroup. It is normal since every group element commutes with elements in the center: $Z(G)g = gZ(G)$.

3.15 Simple and semisimple groups

A simple group G is one that does not have any normal subgroups other than G and $\{e\}$. Simple groups are like prime numbers, they do not admit any nontrivial factor groups and can serve as building blocks for other groups. The cyclic group C_p for prime p is simple as every nontrivial element is a generator. By contrast, $C_4 = \{\pm 1, \pm i\}$ is not simple: $C_2 = \{\pm 1\}$ is a normal subgroup. More generally, G is semisimple if G has no nontrivial abelian invariant subgroups. If G is simple, then it is automatically semisimple. A connected nonabelian Lie group is called simple if it does not have any proper connected normal Lie subgroups (it can have discrete normal subgroups). $SO(3)$, $SU(2)$ and $SL_2(\mathbb{R})$ are simple Lie groups while $SO(4)$

equivalence class of g is the left coset gH.

is semisimple but not simple. The unitary groups $U(n)$ and general linear groups $GL_n(\mathbb{R})$ for $n \geq 2$ are neither simple nor semisimple since multiples of the identity ($e^{i\theta} I$ for real θ and λI for nonzero real λ) form nontrivial abelian connected invariant Lie subgroups. In fact, $U(2)$ also admits $SU(2)$ as a normal subgroup.

We now introduce two ways in which we may combine a pair of groups to synthesize a larger one: the direct and semidirect products.

3.16 Direct product

Suppose H and N are a pair of groups with identity elements e_H and e_N. Then the Cartesian product $H \times N$ consisting of all ordered pairs (h, n) with $h \in H$ and $n \in N$ can be given the structure of a group called the direct product of H and N. The composition law is defined as $(h, n) \cdot (h', n') = (hh', nn')$ and $(h, n)^{-1} = (h^{-1}, n^{-1})$. The subgroups consisting of elements of the form (e_H, n) and (h, e_N) are isomorphic to N and H respectively. $H \times N$ and $N \times H$ are isomorphic groups. When the groups are abelian, one tends to use additive rather than multiplicative notation. For example, the group $\mathbb{R}^2$ of translations of the Euclidean plane is the direct product (or sum) of two copies of the group $\mathbb{R}$ of translations of the real line: $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

3.17 Semidirect product.

The semidirect product is a generalization of the direct product. Here, we suppose that we are given an action of H on N. More precisely, for each $h \in H$, we have an automorphism $\varphi_h : N \to N$ such that $\varphi_{h'} \varphi_h = \varphi_{h'h}$ and $\varphi_h^{-1} = \varphi_{h^{-1}}$. We may use this to define the composition law $(h, n) \cdot (h', n') = (hh', n\varphi_h(n'))$. We verify in Prob. **??** that $H \times N$ with this composition law is a group. It is called the semidirect product of H acting on N via φ and is denoted $H \rtimes_\varphi N$. Evidently, the semidirect product reduces to the direct product if H acts trivially on N, i.e., $\varphi_h(n') = n'$ for all $h \in H$ and $n' \in N$. What is more, the set of elements (e_H, n) forms a normal subgroup of $H \rtimes_\varphi N$ isomorphic to N. The Euclidean group is a semidirect product of 3d rotations acting on space translations. The Galilei and Poincaré groups are semidirect products of the group of rotations and boosts acting on space-time translations.

3.18 Permutation group.

The permutation group S_n or symmetric group on n letters is the set of all permutations of n distinct objects, usually denoted $1, 2, \cdots, n$, with group multiplication given by composition of permutations. A permutation σ may be written in two-row notation as $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & \cdots \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots \end{smallmatrix}\right)$. The group has order $n!$ since $\sigma(1)$ can be chosen in n ways followed by $\sigma(2)$ in $n-1$ ways and so on. A permutation may also be written as a product of disjoint cycles: its cycle decomposition. For $k = 0, 1, \ldots,$ a $(k+1)$-cycle is of the form $(i \; \sigma(i) \; \sigma^2(i) \cdots \sigma^k(i))$ with $\sigma^{k+1}(i) = i$. For example, S_2 consists of 2 elements: the identity $\sigma = e$ [with $e(1) = 1, e(2) = 2$] and exchange transposition $\sigma = \tau$ [$\tau(1) = 2, \tau(2) = 1$] with $\tau^2 = e$. Thus,

$$e = \left(\begin{smallmatrix} 1 & 2 \\ 1 & 2 \end{smallmatrix}\right) = (1)(2) \quad \text{and} \quad \tau = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right) = (12). \tag{231}$$

$g\downarrow,\ \ h\rightarrow$	e	(12)	(23)	(31)	(123)	(132)
e	e	(12)	(23)	(31)	(123)	(132)
(12)	(12)	e	(123)	(132)	(23)	(31)
(23)	(23)	(132)	e	(123)	(31)	(12)
(31)	(31)	(123)	(132)	e	(12)	(23)
(123)	(123)	(31)	(12)	(23)	(132)	e
(132)	(132)	(23)	(31)	(12)	e	(123)

Table 1: Multiplication table of gh for $g, h \in S_3$

The group S_3 has 6 elements. The identity $\sigma(i) = i$ is denoted $(1)(2)(3)$. There are three pairwise transpositions[66] $(12)(3), (1)(23)$ and $(2)(31)$. Here, $(1)(23)$ means $\sigma(1) = 1, \sigma(2) = 3, \sigma(3) = 2$. There are also two cyclic permutations $(123) = (12)(23) = (13)(12)$ and $(132) = (12)(13)$ which have been written as products of pairwise exchanges composed from right to left. Here (132) means $\sigma(1) = 3, \sigma(3) = 2$ and $\sigma(2) = 1$. In the composition $\sigma = (12)(13)$, 3 is mapped to 1 which is then mapped to 2, so that $\sigma(3) = 2$. On the other hand, $\sigma(2) = 1$ and $\sigma(1) = 3$.

S_3 can be realized as the group of rigid motion *symmetries of an equilateral triangle* Δ with vertices labelled v_1, v_2, v_3, say counterclockwise, with horizontal base $v_1 v_2$ and apex v_3. The symmetries of Δ are counterclockwise rotations R_θ about the center by angles $\theta = 0, 2\pi/3, 4\pi/3$ and reflections about the perpendiculars through the vertices v_1, v_2 and v_3. The transformation group consisting of these 6 symmetries is called the *dihedral group* of order 6 and is isomorphic to S_3 via the following map. To R_0 we associate the identity element e. $R_{2\pi/3}$ is mapped to (123) since it takes $v_1 \to v_2$, $v_2 \to v_3$, $v_3 \to v_1$. Similarly, $R_{4\pi/3}$ corresponds to (132) as it takes $v_1 \to v_3$, $v_3 \to v_2$, $v_2 \to v_1$. In the same spirit, reflection through the perpendicular through v_1 is mapped to (23) and so on. Notice that the square of any reflection is the identity and that $R_{2\pi/3}^2 = R_{4\pi/3}$ while $R_{4\pi/3}^2 = R_{8\pi/3} = R_{2\pi/3}$. Correspondingly, the square of any transposition is the identity while $(123)^2 = (132)$ and $(132)^2 = (123)$. The 'multiplication table' of S_3 is displayed in Table. 1. Evidently, it is a nonabelian group. In general, reflections do not commute $[(12)(23) = (123)$ while $(23)(12) = (132)]$ nor do rotations commute with reflections: $(123)(12) = (31)$ while $(12)(123) = (23)$.

By Lagrange's theorem, since the order of a subgroup must divide that of the group, S_3 can only have *subgroups* of order $1, 2, 3$ and 6. There are 4 nontrivial subgroups, each is cyclic and is generated by a transposition or cyclic permutation:

$$\{e, (12)\}, \quad \{e, (23)\}, \quad \{e, (31)\} \quad \text{and} \quad \{e, (123), (132)\}. \tag{232}$$

The first 3 are reflection symmetries while the fourth consists of rotations of Δ. Pairwise transpositions are the building blocks: any permutation can be expressed as a product of transpositions, although the expression is not unique. However, a permutation σ requires an even or odd number of transpositions to be expressed this way. Thus, we define the *sign (or signature or parity) of a permutation* $\operatorname{sgn}(\sigma)$ as ± 1 in

[66] When clear from context, we suppress 1-cycles. So in S_3, (23) is short for $(1)(23)$.

the even and odd cases. The identity has sign $+1$ and any exchange has sign -1. For S_3, cyclic permutations have sign $+1$ as $(123) = (31)(12)$ and $(132) = (12)(31)$.

The sign of a permutation gives a homomorphism: $S_n \to C_2$. The kernel is the *alternating group* A_n of even permutations, a normal subgroup[67] of S_n. For $n = 3$, $A_3 = \{e, (123), (132)\}$ consists of rotational symmetries of Δ. It has 2 left cosets

$$\begin{aligned} &(12)A_3 = (23)A_3 = (31)A_3 = \{(12), (23), (31)\} \\ \text{and} \qquad &eA_3 = (123)A_3 = (132)A_3 = \{e, (123), (132)\} = A_3. \end{aligned} \tag{233}$$

As expected, the left cosets are also right cosets, i.e., $(12)A_3 = A_3(12)$, etc.

All members of a *conjugacy class* have cycle decompositions of the same structure. Cycle structure refers to the number of 1-cycles, 2-cycles, etc. Hence, we should expect S_3 to have three conjugacy classes: the identity, the transpositions and the cyclic permutations: $\{e\}$, $\{(12), (23), (31)\}$ and $\{(123), (132)\}$. The members of a conjugacy class must have the same parity. For instance, the conjugates of (12) are

$$\begin{aligned} (23)(12)(23)^{-1} = (31), &\qquad (31)(12)(31)^{-1} = (23), \\ (123)(12)(123)^{-1} = (23) \quad \text{and} &\quad (132)(12)(132)^{-1} = (13). \end{aligned} \tag{234}$$

S_3 can be realized as a *semidirect product* $H \rtimes N$ of H acting on N, where H and N are cyclic groups of order 2 and 3. For instance, we take $H = \{e, (12)\}$ and $N = A_3$ regarded as subgroups of S_3 and consider the action of H on A_3 via conjugation: $\varphi_h(n') = hn'h^{-1}$. Thus, the semidirect product is

$$(h, n) \cdot (h', n') = (hh', nhn'h^{-1}). \tag{235}$$

The Cartesian product has six elements

$$(e, e),\ (e, (123)),\ (e, (132)),\ ((12), e),\ a = ((12), (123)) \quad \& \quad b = ((12), (132)). \tag{236}$$

The first 4 elements are identified with $e, (123), (132), (12) \in S_3$. If $a \leftrightarrow (31)$ and $b \leftrightarrow (23)$ then one finds that (235) agrees with the S_3 composition law. For instance,

$$((12), (123)) \cdot ((12), (123)) = ((12)^2, (123)(12)(123)(12)) = (e, (31)^2) = (e, e), \tag{237}$$

which agrees with $(31)^2 = e$ in S_3.

3.19 Lie group as a homogeneous manifold

A manifold M (or even just a topological space or set) is homogeneous for a group G if it carries a transitive action of G. To be meaningful, one needs to specify the nature of the manifold M (topological, smooth or geometrically rigid like a Riemannian manifold) and the action of the group must respect that structure. Roughly, all points of a homogeneous manifold look locally the same. For example, the unit circle $x^2 + y^2 = 1$ on the plane is homogeneous under the action of the group $SO(2)$

[67] Conjugation by any element ($g\sigma g^{-1}$) cannot change the parity of σ, so A_n invariant.

of rotations about the z axis. The time axis $\mathbb{R}$ is homogeneous under the action of the group of time-translations $t \mapsto t + s$. Euclidean space $\mathbb{R}^3$ is homogeneous under the action of the group of space-translations $\boldsymbol{r} \mapsto \boldsymbol{r} + \boldsymbol{s}$. The term *homogeneous* should ring a bell: recall the homogeneity of time and space which implied that the results of experiments with identical external conditions do not depend on when or where they are performed. By contrast, the rigid toroidal surface of an inflated tube of a car tyre is *not* homogeneous for the action of the group of rotations about the axle. This action is not transitive since it cannot change the distance of a point on the tyre from the axle. In fact, near a point on the inner rim, the tubular surface looks like a saddle or mountain pass while near a point on the outer rim, it looks like a hill, so neighborhoods of points do not all look the same. Though not a group[68], the round unit sphere S^2 is homogeneous for the rotation group SO(3) in 3d as it carries a transitive action of the latter: any point can be rotated to any other point on S^2. S^2 is also homogeneous for the action of $O(3)$. An ellipsoid of revolution $E = \{x^2 + y^2 + 2z^2 = 1\}$ regarded as a rigid surface in $\mathbb{R}^3$ is not homogeneous under 3d rotations since they do not preserve E. Rotations about the z-axis act on the ellipsoid, though not transitively.

Any group is homogeneous under its own action: G acts on itself transitively via both left and right multiplication. We define the left action of G on itself via $L_g h = gh$ for any $g, h \in G$. The action is transitive since, given any $h, k \in G$, we have $L_{kh^{-1}} h = k$. The right action $R_g h = hg$ is similarly transitive. The right and left actions coincide if G is abelian. For a Lie group, L_g and R_g are diffeomorphisms of G. Thus, a Lie group G is a homogeneous manifold under the action of G.

3.20 Lie algebra of a Lie group

Among the points of G, the identity is distinguished by its simplicity. It makes sense to begin a detailed study of G by focusing on the linear neighborhood of the identity[69]. This leads to the idea of the Lie algebra $\underline{G}$, which, as a vector space, is the tangent space at the identity $T_e G$. Each tangent vector at the identity is a Lie algebra element. We may use left translations L_g (by all elements of G) to pushforward (172) any fixed tangent vector $u \in T_e G$ to obtain a 'left-invariant' vector field $L_{g*} u$ on G. Thus, for each $u \in \underline{G}$ we get an associated left-invariant vector field on G. Consequently, the Lie algebra may also be regarded as the space of left-invariant (or right-invariant) vector fields on G. The algebraic structure of the group endows $T_e G$ (or the space of left-invariant vector fields) with the additional structure of a Lie algebra, i.e., with a bilinear antisymmetric product or 'Lie bracket' satisfying the Jacobi identity. In fact, the Lie bracket is simply the commutator of left-invariant vector fields. We know that the commutator of vector fields is antisymmetric and satisfies the Jacobi identity; one needs to show that the commutator of two left-invariant vector

[68] Any Lie group has at least one nonvanishing vector field: the left-invariant vector field obtained by pushing forward a nonzero tangent vector at the identity. However, as noted in Fig. 6b, S^2 does not admit a nonvanishing vector field.

[69] By homogeneity, the linear neighborhood $T_g G$ of any other point $g \in G$ may be studied by left- or right-translating the tangent space at the identity via L_g or R_g. This idea is also used in studying the rotational dynamics of a rigid body.

fields is again left-invariant. For the Lie algebra of a matrix Lie group, the Lie bracket may be realized concretely in terms of the commutator of matrices, as we shall soon see in the context of the rotation group (243). In fact, we may write a group element in the infinitesimal neighborhood of the identity as $g = e^{su} \approx I + su + s^2u^2/2$ for a real s with $|s| \ll 1$. The matrix u (or su) is then an element of the Lie algebra. The group commutator $[g, h]$ of two such elements $g = e^{su}$ and $h = e^{tv}$ may be shown to be $[g, h] \approx I + st[u, v]$. Thus, the matrix commutator of Lie algebra elements is the first nontrivial approximation to the group commutator.

3.21 Circle group $U(1)$

The 'smallest' Lie group is the group $U(1)$ of unimodular complex numbers $\{z \in \mathbb{C} | z^*z = 1\}$. As shown in Fig. 7, the group elements lie on the unit circle in the complex plane, so the group is also called the circle group S^1. This establishes that it is a differentiable manifold. It is called $U(1)$ since it is also the set of 1×1 unitary matrices[70].

Any unimodular z may be expressed as $z = e^{i\theta}$ where θ is defined modulo 2π. The identity element is $z = 1$, corresponding to $\theta \equiv 0$ modulo 2π. The multiplication law is abelian $e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)} = e^{i\theta_2}e^{i\theta_1}$. The inverse of $z = e^{i\theta}$ is the reciprocal $1/z = e^{-i\theta}$. Since $(\theta_1, \theta_2) \mapsto \theta_1 + \theta_2$ and $\theta \mapsto -\theta$ modulo 2π are smooth maps, $U(1)$ is a one-dimensional Lie group. It is compact (closed and bounded as a subset of the complex plane) and path connected though not simply connected. Its Lie algebra $\underline{U(1)}$ is the tangent space at $z = 1$, which is isomorphic to $\mathbb{R}$. $\underline{U(1)}$ can be taken to be the 1d vector space of imaginary numbers iy for $y \in \mathbb{R}$. A basis for the Lie algebra may be chosen as i. The name **generators** is given to basis elements of the Lie algebra. The U(1) Lie algebra has only one generator, which we have taken as i. We notice that exponentiating a Lie algebra element such as πi gives us a group element $e^{\pi i} = \cos\pi + i\sin\pi = -1$. This map from Lie algebra to Lie group is called the exponential map. More generally, given a nonzero Lie algebra element (say i), exponentiating all its real multiples iy, we get a 1-parameter subgroup e^{iy}. In this case, the exponential map surjects onto the group but is many-to-one: $e^{iy} = e^{i(2n\pi+y)}$ for any $n \in \mathbb{Z}$.

• Since U(1) is abelian, its Lie algebra is also abelian. There is only one generator i and it commutes with itself $[i, i] = 0$.

• Since $U(1)$ is abelian, all its subgroups are also abelian. Its finite subgroups are the cyclic groups C_n for $n = 1, 2, \cdots$. Here, $C_n = \{e^{2\pi ij/n} | j = 0, 1, 2, \cdots n - 1\}$, so the elements of C_n lie at the vertices of a regular n-gon centered at the origin of the complex plane, with the identity as one of its vertices. U(1) also has infinite discrete subgroups. An example is provided by the infinite cyclic subgroup generated by a rotation such as $g = e^{i\sqrt{2}}$. Since there is no integer power such that $g^n = 1$, the powers $\{g^0, g^{\pm 1}, g^{\pm 2}, \cdots\}$ form an infinite cyclic subgroup of U(1).

[70] The unitary group $U(n)$ consists of $n \times n$ complex matrices with $U^\dagger U = I$ where $U^\dagger = (U^t)^*$. Equivalently, it consists of linear maps on an n-dimensional complex vector space that preserve a Hermitian positive-definite inner product. It is a real Lie group of dimension n^2: transition functions are smooth real (not complex) functions.

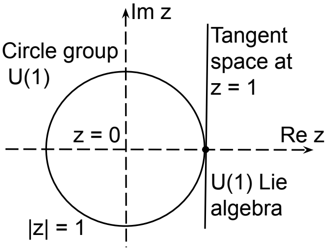


Figure 7: The group $U(1)$ of unimodular complex numbers and its Lie algebra $\underline{U(1)} \cong \mathbb{R}$.

3.22 The orthogonal group O(3)

The orthogonal group $G = O(3)$ consists of the set of 3 × 3 real orthogonal matrices, i.e., matrices A that satisfy $A^t A = I$. The generalization[71] to $n \times n$ orthogonal matrices for $n = 1, 2, 3, 4, \ldots$ is called $O(n)$. The group composition law is associative matrix multiplication: note that $(AB)^t AB = B^t A^t AB = I$ if A and B are orthogonal. The identity element is the identity matrix while the inverse of A is simply its transpose A^t. It is a nonabelian group since $AB \neq BA$ in general for a pair of orthogonal matrices. The orthogonal group is a matrix group, it is a subgroup of the general linear group of all invertible real 3 × 3 matrices. The orthogonal group is important as it is the group of rotations and reflections of 3d Euclidean space. It frequently arises as a group of symmetries (e.g., of a spherically symmetric potential) or as a configuration space (e.g., of the rigid body). We will soon view $O(3)$ as a manifold. First, what is its dimension? The condition $A^t A = I$ implies that a 3 × 3 orthogonal matrix is one whose columns furnish an orthonormal basis $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ for $\mathbb{R}^3$ (see below). The first basis vector $\boldsymbol{a}$ is any unit vector. The latter are parametrized by points on the unit sphere $S^2 \subset \mathbb{R}^3$. Thus, 2 real parameters are needed to specify $\boldsymbol{a}$. Having picked $\boldsymbol{a}$, the second basis vector $\boldsymbol{b}$ can be any unit vector in the plane orthogonal to $\boldsymbol{a}$ and is specified by a point on the unit circle S^1 on this plane. Thus, one additional parameter is needed to specify $\boldsymbol{b}$. Pictorially, we may view the possible choices of $\boldsymbol{a}$ and $\boldsymbol{b}$ as giving us a unit circle bundle over the unit 2 sphere. Having chosen $\boldsymbol{a}$ and $\boldsymbol{b}$, the third basis vector $\boldsymbol{c}$ must be perpendicular to both: $\boldsymbol{c} = \pm \boldsymbol{a} \times \boldsymbol{b}$. Thus, there is no additional continuous real parameter needed to specify $\boldsymbol{c}$. The two choices of $\boldsymbol{c}$ lead to orthogonal matrices with determinant ± 1. We conclude that $O(3)$ is a 3-parameter family of matrices. In fact, we may view it as a 3d submanifold of $\mathbb{R}^9$. Suppose we write A in terms of its columns,

$$A = \begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{pmatrix} \quad \text{so that} \quad A^t = \begin{pmatrix} \boldsymbol{a}^t \\ \boldsymbol{b}^t \\ \boldsymbol{c}^t \end{pmatrix}. \qquad (238)$$

[71] Alternatively, suppose V is an n dimensional real vector space with positive-definite inner product $\langle \cdot, \cdot \rangle$. Then $O(n)$ is the group of linear maps $A : V \to V$ that preserve the inner product $\langle Au, Av \rangle = \langle u, v \rangle$ for all $u, v \in V$ with product given by composition of maps. The definition is independent of the choice of V and inner product, it only depends on n.

The constraint $A^t A = I$, becomes 6 conditions on the 9 matrix elements of A:

$$\boldsymbol{a}^t\boldsymbol{a} - 1 = \boldsymbol{b}^t\boldsymbol{b} - 1 = \boldsymbol{c}^t\boldsymbol{c} - 1 = \boldsymbol{a}^t\boldsymbol{b} = \boldsymbol{b}^t\boldsymbol{c} = \boldsymbol{c}^t\boldsymbol{a} = 0. \tag{239}$$

Thus, $O(3)$ is the common zero locus of these six independent quadratic functions of nine real variables. Hence, we may view $O(3)$ as a 3d algebraic submanifold of $\mathbb{R}^9$. It is bounded since $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ must each be a unit vector. It is closed[72] since it is the intersection of the inverse images of the closed one-element set $\{0\}$ under the continuous maps $\boldsymbol{a}^t\boldsymbol{a} - 1, \cdots, \boldsymbol{c}^t\boldsymbol{a}$ from $\mathbb{R}^9 \to \mathbb{R}$. Thus, $O(3)$ is a compact 3d manifold. However, it is not path connected. Taking the determinant of $A^t A = I$, we find $(\det A)^2 = 1$, so $\det A = \pm 1$. We check that there are orthogonal matrices with either sign of determinant. Since the determinant cannot jump discontinuously from 1 to -1 along a continuous path, we conclude that $O(3)$ is disconnected. It has two connected components. The identity I lies in the connected component where $\det A = 1$ and comprises proper rotations of $\mathbb{R}^3$. In fact, the connected component of the identity is a closed subgroup of $O(3)$ and is a Lie group in its own right, the special orthogonal group $SO(3)$ which is also the kernel of the determinant homomorphism from $O(3)$ to $C_2 = \{\pm 1\}$. The subgroup $SO(3)$ and its Lie algebra play a key role in the study of angular momentum and the rigid body problem. The other component where $\det A = -1$ is not a subgroup as it is not closed under composition and does not include the identity matrix. It consists of so-called improper rotations and is a coset of $SO(3)$ by a reflection: product of a reflection and a proper rotation.

3.23 The Lie algebra of O(3)

The Lie algebra as a vector space is defined as the tangent space to the group at the identity. To identify the orthogonal Lie algebra $\underline{O(3)}$, we suppose $A \approx I + u$ and treat u to linear order. The orthogonality condition

$$(I + u^t)(I + u) \approx I \quad \text{becomes} \quad u + u^t = 0. \tag{240}$$

Thus, the Lie algebra of the orthogonal group consists of 3×3 real antisymmetric matrices[73]. A real linear combination of antisymmetric matrices $\alpha u + \beta v$ remains antisymmetric, so this is indeed a vector space. The entries above the diagonal are the only linearly independent entries of an antisymmetric matrix, so $\underline{O(3)}$ is a 3-dimensional real vector space isomorphic to $\mathbb{R}^3$. We say that $\underline{O(3)}$ is a 3-dimensional Lie algebra. It is no surprise that G and $\underline{G}$ have the same dimension. A convenient basis $\{e_1, e_2, e_3\}$ for $\underline{O(3)}$ is furnished by the matrices with ab-entries $(e_i)_{ab} = -\epsilon_{iab}$:

$$e_1 = -\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad e_2 = -\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad e_3 = -\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{241}$$

This means any antisymmetric 3×3 matrix u may be expressed as $u = u^1 e_1 + u^2 e_2 + u^3 e_3$ where u^1, u^2, u^3 are real coefficients.

[72] For our purposes, a closed set C contained in Euclidean space is one that contains all its limit points with respect to the Euclidean distance function. The inverse image $f^{-1}(C)$ of a closed set under a continuous map f is closed. The intersection of a finite number of closed sets is closed. A closed and bounded subset of Euclidean space is called compact.

[73] Since $SO(3)$ is the identity component of $O(3)$, they have the same Lie algebra.

3.24 Exponential map.

As with $U(1)$, given a nonzero Lie algebra element u, it determines a 1-parameter subgroup of the Lie group via the matrix exponential $u \mapsto A = e^{su}$ where $s \in \mathbb{R}$. Indeed, we verify that if u is antisymmetric, then A is orthogonal:

$$A^t = e^{su^t} = e^{-su} = A^{-1}. \tag{242}$$

More generally, by exponentiating[74] a linear combination $u^i e_i$ of basis elements, we obtain a three-parameter family of group elements: $e^{su^i e_i}$. All elements of the identity component of $O(3)$ [proper rotations, i.e., elements of $SO(3)$] may be obtained this way. However, improper rotations cannot be reached by exponentiating elements u of the Lie algebra, since $\det e^{su} = 1$. We could of course compose elements of $SO(3)$ by reflections to obtain the improper rotations. Thus, we see that the Lie algebra contains enough data to recover a finite neighborhood of the group identity. In favorable cases (e.g., compact connected groups such as $U(1)$ or $SO(3)$), the exponential map takes the Lie algebra surjectively onto the group.

3.25 Lie bracket and structure constants

As noted earlier, the group structure of a Lie group G endows the tangent vector space $\underline{G}$ at the group identity with a bilinear antisymmetric product $\underline{G} \times \underline{G} \to \underline{G}$ called the Lie bracket, satisfying the Jacobi identity. For the Lie algebra of a matrix group, the Lie bracket is simply the commutator[75] of matrices and the Jacobi identity is automatically satisfied since it is a property of the matrix commutator. For $\underline{O(3)}$, we verify that the Lie brackets among the basis elements are

$$[e_1, e_2] = e_3, \quad [e_2, e_3] = e_1 \quad \text{and} \quad [e_3, e_1] = e_2 \quad \text{or} \quad [e_i, e_j] = \epsilon_{ijk} e_k. \tag{243}$$

This Lie algebra should remind us of the cross products $\hat{x} \times \hat{y} = \hat{z}, \hat{y} \times \hat{z} = \hat{x}, \hat{z} \times \hat{x} = \hat{y}$ of the orthonormal basis vectors $\hat{x}, \hat{y}, \hat{z}$ of $\mathbb{R}^3$. In fact, $\underline{O(3)}$ is isomorphic[76] to the cross product Lie algebra of vectors in 3d Euclidean space $\mathbb{R}^3$. The isomorphism takes vectors r_a to antisymmetric matrices via $u_{ab} = \epsilon_{abc} r_c$ and conversely

[74] While the matrix exponential gives a map from the Lie algebra to the group for matrix groups, there is a more general way of defining the exponential map. Given an element of the Lie algebra $u \in T_e G$, it defines a left-invariant vector field X_u on G via the pushforward of u through the left-translation map to all points of G, $L_{g_*} : T_e G \to T_g G$. Now, consider the integral curve of X_u (with parameter s) that begins at the identity in the direction of u. This is the solution of the system of 1st order ODEs $\dot{x} = X_u$ subject to the IC $x(0) = I$. The exponential map, by definition, maps u to the point $x(1) \in G$ (i.e., put $s = 1$). The appearance of e^u in matrix groups is natural since the solution of this initial value problem in that case is $\exp(su)$.

[75] Note that while the commutator of matrices in $\underline{G}$ is an element of $\underline{G}$, the product of matrices in the Lie algebra of a matrix group is not regarded as an element of the Lie algebra. With reference to $\underline{O(3)}$, the product of antisymmetric matrices is not antisymmetric in general.

[76] The group $SU(2)$ of 2×2 unitary matrices with unit determinant also has a Lie algebra isomorphic to $\underline{O(3)}$. Indeed, the Lie brackets among $(1/2i)\times$ the Pauli matrices [which furnish a basis for $\underline{SU(2)}$] are the same as those among e_1, e_2 and e_3.

$r_c = (1/2)\epsilon_{abc}u_{ab}$. In general, the Lie brackets among basis elements of a Lie algebra can be expressed as a linear combination of basis elements: $[e_i, e_j] = c_{ijk}e_k$. The coefficients c_{ijk} are called the structure constants of the Lie algebra (in the chosen basis). They must be antisymmetric in i and j. The Jacobi identity is a quadratically nonlinear condition on the structure constants. For $SO(3)$ or $O(3)$, the structure constants in the above basis are given by the components of the Levi-Civita symbol $c_{ijk} = \epsilon_{ijk}$.

3.26 The group SU(2) and its Lie algebra

SU(2) as the 3-sphere. The Lie group SU(2) consists of 2×2 unitary matrices of unit determinant, $gg^\dagger = g^\dagger g = I$ and $\det g = I$. The condition of unitarity means the rows are orthonormal vectors in $\mathbb{C}^2$. Writing $g = (a, b|c, d)$ and imposing unitarity and the unit determinant condition, we find that any element of SU(2) may be expressed as

$$g = \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix} \quad \text{with} \quad a, b \in \mathbb{C} \quad \text{and} \quad |a|^2 + |b|^2 = 1. \tag{244}$$

If we write a and b in terms of their real and imaginary parts, $a = a_1 + ia_2$ and $b = b_1 + ib_2$ then $|a|^2 + |b|^2 = 1$ becomes the condition $a_1^2 + a_2^2 + b_1^2 + b_2^2 = 1$. This is the equation for a 3-sphere embedded in $\mathbb{R}^4$. Thus, $SU(2)$ as a manifold is the 3-sphere S^3. Since composition and inversion are compatible with the manifold structure, SU(2) is a three-dimensional Lie group.

• Let us show how we arrive at the above parametrization (244). Note that

$$gg^\dagger = I \quad \Rightarrow \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a^* & c^* \\ b^* & d^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \det g = ad - bc = 1. \tag{245}$$

These imply the four equations

$$|a|^2 + |b|^2 = 1, \quad |c|^2 + |d|^2 = 1, \quad ac^* + bd^* = 0 \quad \text{and} \quad ad - bc = 1. \tag{246}$$

The conditions following from $g^\dagger g = I$ are not independent of these (they have the same solutions). These four equations are actually 5 real conditions. The first and second are real conditions. The 3rd is two conditions: both the real and imaginary parts of $ac^* + bd^*$ must vanish. The $\det g = 1$ equation is just one real condition. This is because unitarity already implies that $|\det g| = 1$ so that $\det g$ must be a point on the unit circle. The condition $\det g = 1$ then picks one point on this circle. These 5 conditions on the 8 real variables in $a, b, c, d \in \mathbb{C}$ leave behind a 3 dimensional manifold. Let us now try to solve the equations.

• Assuming $a \neq 0$, we use $\det g = 1$ to eliminate $d = (1 + bc)/a$. Putting this in the third condition, we get

$$ac^* + b\left(\frac{1 + b^*c^*}{a^*}\right) = 0 \quad \text{or} \quad |a|^2 c^* + b + |b|^2 c^* = 0. \tag{247}$$

Using $|a|^2 + |b|^2 = 1$ this allows us to eliminate $c = -b^*$. Putting this in $d = (1 + bc)/a$ we get

$$d = (1 - |b|^2)/a = |a|^2/a = a^*. \tag{248}$$

Thus we have eliminated c and d in favor of a and b. The remaining condition $|c|^2 + |d|^2 = 1$ tells us that $|a|^2 + |b|^2 = 1$. This leads to the advertised parametrization (244) when $a \neq 0$.

• The special case $a = 0$ is dealt with separately. If $a = 0$ then $|b|^2 = 1$ so $b \neq 0$. Consequently, $ac^* + bd^* = 0$ implies that $d = 0$. Finally, $ad - bc = 1$ implies that $c = -1/b = -b^*$. Combining, we get the advertised parametrization (244) of g.

The Lie algebra of SU(2). The Lie algebra of SU(2) consists of traceless 2×2 antihermitian matrices. To see this, consider an SU(2) group element near the identity: $g \approx I + u$. Then $g^\dagger \approx I + u^\dagger$. Putting this in $g^\dagger g = I$ gives $(I + u)(I + u^\dagger) \approx I$ or $I + u + u^\dagger \approx I$, whence $u^\dagger = -u$, so that u is antihermitian. Moreover, $\det g \approx \det(I + u) \approx 1 + \operatorname{tr} u = 1$ implies $\operatorname{tr} u = 0$. So the Lie algebra of $SU(2)$ consists of 2×2 traceless antihermitian matrices. A convenient basis consists of the three matrices $\tau_j = \sigma_j/2i$ where $\sigma_1 = (0, 1|1, 0)$, $\sigma_2 = (0, -i|i, 0)$ and $\sigma_3 = (1, 0|0, -1)$ are the 3 Pauli matrices. Division by i ensures that the τ_j are antihermitian while the division by 2 is to make the structure constants simple. Show that the Lie brackets (commutators) among the generators $\tau_{1,2,3}$ are

$$[\tau_i, \tau_j] = \epsilon_{ijk}\tau_k. \tag{249}$$

So the structure constants of the SU(2) Lie algebra in this basis are given by the components of the Levi-Civita symbol. Any traceless antihermitian matrix can be expressed as $u = x_1\tau_1 + x_2\tau_2 + x_3\tau_3 = \boldsymbol{x} \cdot \boldsymbol{\tau}$ for three real coefficients x_1, x_2, x_3. We notice that these structure constants are the same as those of the SO(3) Lie algebra in the basis $(e_i)_{ab} = -\epsilon_{iab}$. Thus, the SU(2) and SO(3) Lie algebras are isomorphic.

• **Isomorphism between SU(2) and $\mathbb{R}^3$ cross product Lie algebras.** In addition, $\mathbb{R}^3$ is also a 3d Lie algebra with Lie bracket given by the cross product of vectors. The cross products of the basis vectors $\hat{r}_i = \{\hat{x}, \hat{y}, \hat{z}\}$ are

$$\hat{x} \times \hat{y} = \hat{z}, \quad \hat{y} \times \hat{z} = \hat{x}, \quad \hat{z} \times \hat{x} = \hat{y}. \tag{250}$$

The structure constants of the cross product Lie algebra in this basis,

$$\hat{r}_i \times \hat{r}_j = \epsilon_{ijk}\hat{r}_k \tag{251}$$

are again given by the Levi-Civita symbol. Thus the SU(2) Lie algebra is isomorphic to the cross product Lie algebra on $\mathbb{R}^3$ with the isomorphism taking τ_i to $\hat{r}_i$ and extended by linearity to other traceless antihermitian matrices. We can write an explicit formula for this isomorphism. Given a vector $\boldsymbol{x} \in \mathbb{R}^3$, we have the associated traceless antihermitian matrix

$$u(\boldsymbol{x}) = \boldsymbol{\tau} \cdot \boldsymbol{x} = \frac{1}{2i}\begin{pmatrix} x_3 & x_1 - ix_2 \\ x_1 + ix_2 & -x_3 \end{pmatrix}. \tag{252}$$

We note that $\det u(\boldsymbol{x}) = \frac{1}{4}(x_1^2 + x_2^2 + x_3^2) = \frac{1}{4}|\boldsymbol{x}|^2$ is one-fourth the squared length of the vector $\boldsymbol{x}$.

- Conversely, the vector $\boldsymbol{x}(u)$ associated to the $\underline{SU(2)}$ Lie algebra element u has the components $(\boldsymbol{x}(u))_j = -2\,\mathrm{tr}\,(u\tau_j)$. Check this formula.
- This isomorphism between $\underline{SU(2)}$ and $\mathbb{R}^3$ takes the commutator of traceless antihermitian matrices to the cross product of vectors. Show that this is the case. Thus, we have an isomorphism of Lie algebras.

3.27 Adjoint action or representation of group and Lie algebra

Any group acts on itself by conjugation, which is an inner automorphism. Given any fixed $g \in G$, we have the automorphism $A_g : G \to G$ given by $A_g(h) = ghg^{-1}$. If h is in the linear neighborhood of the identity I, then then we can turn this into a linear action of G on its Lie algebra. Let us suppose that G is a matrix Lie group and put $h = I + v$ with v treated to linear order. Then

$$A_g(h) \approx g(I+v)g^{-1} = I + gvg^{-1}. \tag{253}$$

Thus, we get the so-called group adjoint action of G on its Lie algebra $\underline{G}$:

$$Ad_g(v) = gvg^{-1} \quad \text{for any} \quad g \in G \quad \text{and} \quad v \in \underline{G}. \tag{254}$$

Check that this is indeed a group action. Moreover, the Lie algebra is a vector space (T_eG, having dimension equal to that of G) and the action is linear. An action of a group on a vector space is called a group representation. This is why the group adjoint action on $\underline{G}$ is also called the adjoint representation of the group. The adjoint representation of a Lie group has the same dimension as the group itself.

- We may further suppose that g lies close to the identity of G. Then the group adjoint action reduces to an action of the Lie algebra on itself. Find a formula for this Lie algebra adjoint representation.

3.28 Two-to-one homomorphism from SU(2) to SO(3)

- $SU(2)$ is the group of 2×2 complex unitary matrices of unit determinant. It is a 3d Lie group. Its Lie algebra consists of traceless antihermitian 2×2 matrices.
- O(3) consists of orthogonal transformations of Euclidean 3-space, i.e., $\boldsymbol{x} \to R\boldsymbol{x}$ where $R^tR = I$. SO(3) consists of proper rotations of Euclidean 3-space, i.e., $\boldsymbol{x} \to R\boldsymbol{x}$ where $R^tR = I$ with $\det R = 1$. Its Lie algebra consists of 3×3 traceless antisymmetric matrices.
- Given an element g of SU(2) we will define an associated orthogonal transformation ϕ_g of $\mathbb{R}^3$. This will lead us to a homomorphism from SU(2) to O(3). We will indicate why ϕ is a 2:1 homomorphism.
- We will argue that SU(2) is path connected, while O(3) has two connected components: the identity component SO(3) and the other component consisting of improper rotations. Continuity of the map ϕ will imply that the image of SU(2) lies in the identity component. Thus we will get a 2:1 homomorphism from SU(2) to SO(3). This homomorphism can be shown to be surjective with kernel consisting of $\{\pm I_{2\times 2}\} \cong C_2$. Thus $SU(2)/C_2 \cong SO(3)$.

• **Map from SU(2) to O(3).** To define the 3d orthogonal transformation ϕ_g, we will exploit the adjoint action of SU(2) on its Lie algebra. Using the isomorphism between $\underline{SU(2)}$ and $\mathbb{R}^3$, we will get an action of SU(2) on $\mathbb{R}^3$, which we will show to be orthogonal.

• Now $\mathrm{SU}(2)$ acts on its Lie algebra via the adjoint representation: $Ad_g(u) = gug^{-1} = gug^\dagger$. This transformation clearly preserves antihermiticity ($(gug^\dagger)^\dagger = gu^\dagger g^\dagger = -gug^\dagger$ if $u^\dagger = -u$) and tracelessness ($\operatorname{tr} gug^\dagger = \operatorname{tr} u = 0$). Using the map from $\underline{SU(2)}$ to $\mathbb{R}^3$ we convert this into an action of SU(2) on $\mathbb{R}^3$. In other words, for any $g \in SU(2)$, we get a map

$$\phi_g : \mathbb{R}^3 \to \mathbb{R}^3 \quad \text{taking } \boldsymbol{x} \mapsto \boldsymbol{x}' = \phi_g(\boldsymbol{x}) \quad \text{with} \quad (\phi_g(\boldsymbol{x}))_i = -2\operatorname{tr}(g\, u(\boldsymbol{x})\, g^\dagger\, \tau_i). \tag{255}$$

Since $\det gu(\boldsymbol{x})g^\dagger = \det u(\boldsymbol{x})$, it follows that ϕ_g preserves lengths:

$$|\boldsymbol{x}'|^2 = |\phi_g(\boldsymbol{x})|^2 = 4\det gu(\boldsymbol{x})g^\dagger = 4\det u(\boldsymbol{x}) = |\boldsymbol{x}|^2. \tag{256}$$

But length preserving linear transformations are the same as orthogonal transformations. So if $g \in \mathrm{SU}(2)$ then $\phi_g \in \mathrm{O}(3)$.

• **Formula for orthogonal transformation ϕ_g.** Using the parametrization of $\mathrm{SU}(2)$ group elements

$$g = \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix}, \quad g^\dagger = \begin{pmatrix} a^* & -b \\ b^* & a \end{pmatrix} \quad \text{with} \quad |a|^2 + |b|^2 = 1. \tag{257}$$

we may obtain an explicit formula for the effect of the orthogonal transformation ϕ_g on a vector $\boldsymbol{x} = (x, y, z)$. Denoting $\phi_g(\boldsymbol{x}) = \boldsymbol{x}' = (x', y', z')$, we find

$$gu(\boldsymbol{x}) = \frac{1}{2i}\begin{pmatrix} az + b(x+iy) & a(x-iy) - bz \\ -b^*z + a^*(x+iy) & -b^*(x-iy) - a^*z \end{pmatrix}. \tag{258}$$

Right multiplying by $g^\dagger$ and comparing with the definition

$$gu(\boldsymbol{x})g^\dagger = u(\boldsymbol{x}') = \frac{1}{2i}\begin{pmatrix} z' & x' - iy' \\ x' + iy' & -z' \end{pmatrix}, \tag{259}$$

we find that

$$\begin{aligned} x' &= \Re\left[a^2(x-iy) - b^2(x+iy) - 2abz\right] \\ y' &= \Im\left[{a^*}^2(x+iy) - {b^*}^2(x-iy) - 2a^*b^*z\right] \quad \text{and} \\ z' &= (|a|^2 - |b|^2)z + 2\Re(a^*b(x+iy)). \end{aligned} \tag{260}$$

This is clearly a linear transformation. Write this out as $\boldsymbol{x}' = R(g)\boldsymbol{x}$ and identify the entries of the 3×3 orthogonal matrix $R(g) = \phi_g$ in terms of the entries of g. Verify that $R^tR = I$.

• **Group homomorphism property:** This map $\phi : g \mapsto \phi_g$ is a group homomorphism from $\mathrm{SU}(2)$ to $\mathrm{O}(3)$. To show ϕ is a homomorphism, we must show that $\phi_{gg'}$ is the same orthogonal transformation of $\mathbb{R}^3$ as $\phi_g \phi_{g'}$. In other words, we must show that for all $\boldsymbol{x} \in \mathbb{R}^3$ and all $g, g' \in \mathrm{SU}(2)$, we have

$$\phi_{gg'}(\boldsymbol{x}) = \phi_g(\phi_{g'}(\boldsymbol{x})). \tag{261}$$

We check by explicit calculation that this is true. Details are suppressed.

• **Map $\phi : SU(2) \to O(3)$ is 2:1.** Note that $(-g)^\dagger = -g^\dagger$. It follows that $\phi_g = \phi_{-g}$, so antipodal points of SU(2) are mapped to the same orthogonal transformation. Hence, this map from SU(2) to O(3) is at least two-to-one. To show that it is precisely 2:1 we check that the kernel of the homomorphism consists of the 2 element subgroup $\{\pm I\}$ which is isomorphic to the cyclic group of order 2. In fact,

$$\ker \phi = \{g \in \mathrm{SU}(2) \mid \phi_g = I_3 \in \mathrm{O}(3)\} \tag{262}$$

For g to lie in $\ker \phi$, we need $gug^\dagger = u$ for all $u \in \underline{\mathrm{SU}(2)}$. Why? Since $\underline{SU(2)}$ is isomorphic to $\mathbb{R}^3$ this is a necessary and sufficient condition for ϕ_g to act as the identity on $\mathbb{R}^3$. So we need to find all g such that $gu = ug$ for all u. We find that

$$\begin{aligned} gu &= \begin{pmatrix} az + b(x+iy) & a(x-iy) - bz \\ -b^* z + a^*(x+iy) & -b^*(x-iy) - a^* z \end{pmatrix} \quad \text{and} \\ ug &= \begin{pmatrix} za - b^*(x-iy) & zb + a^*(x-iy) \\ a(x+iy) + zb^* & b(x+iy) - za^* \end{pmatrix}. \end{aligned} \tag{263}$$

Then $gu = ug$ for all x, y, z implies that $b = 0$ and $a \in \mathbb{R}$. Along with $|a|^2 + |b|^2 = 1$, we get $a = \pm 1$ and so $g = \pm I$ and $\ker \phi = \{\pm I\}$. $\ker \phi$ is of course a normal subgroup of SU(2). If $\phi_g = R$ then $\phi_{-g} = \phi_{-I} \phi_g = R$. So ϕ is a 2:1 homomorphism from $\mathrm{SU}(2)$ onto its image, which must be isomorphic to $\mathrm{SU}(2)/C_2$.

• **Image of SU(2) lies in SO(3).** Next we argue that the image of SU(2) under the homomorphism ϕ lies in SO(3). In other words, ϕ_g cannot be an improper rotation. To see this, we first note that SU(2) is path connected.

• **SU(2) is path connected.** To show this, given any $g \in \mathrm{SU}(2)$ it suffices to display a path g_t (for $0 \le t \le 1$) in $\mathrm{SU}(2)$ such that $g_0 = g$ and $g_1 = I$. Now any such g can be diagonalized by a unitary matrix V. The eigenvalues of a unitary matrix are complex numbers of unit magnitude[77]. Since the determinant is one, the eigenvalues must be λ and $1/\lambda$ with $|\lambda| = 1$. In other words,

$$g = VDV^\dagger = V \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix} V^\dagger \tag{264}$$

Now define the 1-parameter family of unitary matrices

$$g_t = V \begin{pmatrix} \lambda_t & 0 \\ 0 & 1/\lambda_t \end{pmatrix} V^\dagger \tag{265}$$

[77] The eigenvalue problem $U\psi = \lambda\psi$ and its adjoint $\psi^\dagger U^\dagger = \lambda^* \psi^\dagger$ together give $\psi^\dagger U^\dagger U \psi = \lambda^* \lambda \psi^\dagger \psi$. Since $U^\dagger U = I$ and $\psi^\dagger \psi \neq 0$, we get $\lambda^* \lambda = 1$. Alternatively, the rows of a unitary matrix are orthonormal vectors. Imposing this condition in the basis where the matrix is diagonal, we find that the eigenvalues must have unit magnitude.

where λ_t is any curve on the unit circle of the complex plane with $\lambda_0 = \lambda$ and $\lambda_1 = 1$. Then $g_0 = g$ and $g_1 = I$. So we have joined g to I by a continuous curve in $SU(2)$. To join two points g and g' in $SU(2)$ just join each to I and concatenate the paths. Thus we showed that $SU(2)$ is path connected.

• Now $\phi : SU(2) \to O(3)$ maps the identity to the identity in $O(3)$ which lies in the component of proper rotations $SO(3)$. As $SU(2)$ is path connected, the image of $SU(2)$ under the continuous map ϕ must also be path connected and therefore must lie in $SO(3)$.

• **ϕ maps SU(2) onto SO(3).** We will now argue that the image of SU(2) under ϕ is the whole of $SO(3)$. To do so we use a result of Euler that says that any rotation in 3 dimensions can be expressed as a product of three rotations about the z, y and z axes:

$$R = R^z_\varphi R^y_\theta R^z_\psi \tag{266}$$

The angles ψ, θ, φ are called Euler angles. Bearing this in mind, if we can show that rotations about the z and y axes lie in the image of SU(2) under ϕ, then we would have shown that any SO(3) element lies in the homomorphic image of SU(2). We will do this below.

• **Rotations about z lie in the image of SU(2).** Consider the SU(2) group elements of the form $g_\theta = \begin{pmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta} \end{pmatrix}$ where θ is an angle. This is a 1-parameter subgroup of SU(2). Then under the adjoint action,

$$\phi_g : u(\boldsymbol{x}) \mapsto g_\theta u(\boldsymbol{x}) g_\theta^\dagger = \begin{pmatrix} z & e^{-2i\theta}(x - iy) \\ e^{2i\theta}(x + iy) & -z \end{pmatrix} = \begin{pmatrix} z' & x' - iy' \\ x' + iy' & -z' \end{pmatrix}. \tag{267}$$

So $z' = z$. Thus g_θ is mapped by ϕ to the rotation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{268}$$

So $\phi_{g_\theta} = R^z_{2\theta}$ is a counterclockwise rotation about the z-axis by an angle 2θ. Thus every rotation about the z-axis lies in the image of SU(2) under this homomorphism. Moreover, as θ goes from 0 to π, g_θ goes from $g_0 = I$ to $g_\pi = -I$ tracing an open curve in SU(2), but the image $R^z_0 = R^z_{2\pi} = I$ traces a closed curve in O(3). On the other hand, the closed curve g_θ for $0 \leq \theta \leq 2\pi$ on SU(2) is mapped to a closed curve $R_{2\theta}$ that traverses itself twice. This is a manifestation of $\phi_{-g} = \phi_g$ and the 2:1 nature of the homomorphism.

• **Rotations about y lie in the image of SU(2).** Let the 1-parameter subgroup of elements $\tilde{g}_\alpha$ of SU(2) be defined as

$$\tilde{g}_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \quad \text{with} \quad \tilde{g}_\alpha^\dagger = \tilde{g}_{-\alpha}. \tag{269}$$

Then using the abbreviations $c = \cos\alpha$ and $s = \sin\alpha$,

$$\tilde{g}_\alpha u(\boldsymbol{x}) \tilde{g}_{-\alpha} = \frac{1}{2i} \begin{pmatrix} (c^2 - s^2)z - 2csx & 2csz + (c^2 - s^2)x - iy \\ 2csz + (c^2 - s^2)x + iy & -(c^2 - s^2)z + 2csx \end{pmatrix}. \tag{270}$$

Comparing this with

$$u(\boldsymbol{x}') = \frac{1}{2i}\begin{pmatrix} z' & x' - iy' \\ x' + iy' & -z' \end{pmatrix}, \tag{271}$$

we deduce that $y' = y$ and that $\phi_{\tilde{g}_\alpha}$ is a counterclockwise rotation about the y-axis by an angle 2α:

$$\begin{pmatrix} z' \\ x' \end{pmatrix} = \begin{pmatrix} \cos 2\alpha & -\sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{pmatrix}\begin{pmatrix} z \\ x \end{pmatrix} = R^y_{2\alpha}\begin{pmatrix} z \\ x \end{pmatrix} \quad \Rightarrow \quad \phi_{\tilde{g}_\alpha} = R^y_{2\alpha}. \tag{272}$$

So we have shown that any rotation about the y axis of $\mathbb{R}^3$ lies in the image of SU(2) under the homomorphism ϕ.

• $\boldsymbol{SO(3) \cong SU(2)/\{\pm I\}}$. Combining with Euler's theorem, we deduce that the homomorphism ϕ maps SU(2) in a 2:1 manner onto SO(3). We say that SU(2) is the double cover of SO(3). Since the kernel of ϕ is $\{\pm I\}$, we have shown that SO(3) is a quotient of SU(2) by an invariant cyclic subgroup of order two: $SO(3) \cong SU(2)/\{\pm I\}$.

• **SO(3) is not simply connected.** We may use the 2:1 homomorphism ϕ to learn a little about the topology of SO(3). We will argue that SO(3) is not simply connected. $R^z_{2\theta}$ for $0 \le \theta \le \pi$ is a closed curve on $SO(3)$ joining I_3 to I_3. We ask whether we can shrink it to the identity I_3 by a smooth deformation. Now $R^z_{2\theta}$ is the image of the curve $g_\theta = \begin{pmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta} \end{pmatrix}$ in SU(2). g_θ for $0 \le \theta \le \pi$ defines an open curve joining I_2 to $-I_2$. Now any deformation of $R^z_{2\theta}$ holding $R_0 = R_{2\pi} = I$ will correspond to a deformation of g_θ holding $g_0 = I, g_\pi = -I$ fixed. So $R^z_{2\theta}$ cannot be continuously shrunk to the constant curve at the point I_3 since that would correspond to the constant curve at I_2 on SU(2). Thus, SO(3) cannot be simply connected. On the other hand, one can show that SU(2) (the 3-sphere) is simply connected. This is done in topology textbooks using the Seifert-Van Kampen theorem.