Methods of Theoretical Physics: II

ABSTRACT

Special Functions. Bessel functions; integral representations; asymptotic expansions. Examples of applications in scattering theory. Hypergeometric functions; integral representations; confluent hypergeometric functions. Laplace and Fourier transforms. Gibbs phenomenon. Integral equations. Conformal mapping, and solution of two-dimensional potential problems. Riemann sphere. Introduction to tensor analysis; general coordinate transformations; covariant differentiation, general relativity. Some introductory group theory; orthogonal groups and their representations; spherical harmonics.

Contents

1 Bessel Functions

In a strictly logical approach we should perhaps, at this stage, begin on a detailed study of the Hypergeometric Equation, and its solutions, since this equation encompasses as special cases many of those that one encounters in physics. However, such a presentation would run the risk of being rather dry and abstract. Instead, we shall adopt the approach of beginning with the Bessel equation, and its solutions. In particular, we shall see how to use the methods of complex analysis in order to determine properties of the solutions. Many of the methods that we use will be generalisable later to other examples, including the hypergeometric equation.

As we saw in part I of the course, Bessel's equation arises when one uses the method of separation of variables to solve an equation such as Laplace's equation in cylindrical polar coordinates. Specifically, it is the radial functions that satisfy the Bessel equation. After appropriate changes of variable, this equation can be cast in the form

$$z^2\, y'' + z\, y' + (z^2 - \nu^2)\, y = 0\,, \tag{1.1}$$

where y is a function of z, and ν is a constant which may be integer on non-integer.

1.1 $J_n(z)$ Bessel Function of Integer Order n

Consider first the case when $\nu = n$, where n is an integer (which can be positive, negative or zero). We can give the following construction of the Bessel function $J_n(z)$, which satisfies (1.1) with $\nu = n$. We define $J_n(z)$ by means of the expansion

$$e^{\frac{1}{2}z(t-t^{-1})} = \sum_{n=-\infty}^{\infty} t^n\, J_n(z)\,. \tag{1.2}$$

This is known as a *generating function* for the Bessel functions. In principle one could expand the left-hand side as a Laurent series in t, and by picking out all the terms proportional to t^n, one reads off the corresponding Bessel function $J_n(z)$. Of course there will be infinitely many terms in this expansion, since each power $(t-t^{-1})^N$ in the Taylor expansion of $e^{\frac{1}{2}z(t-t^{-1})}$ contains all powers of t from t^{-N} to t^N.

Let us begin by verifying that (1.2) does indeed give us a construction of solutions of the Bessel equation. Thus we wish to verify that $J_n(z)$ defined by (1.2) does indeed satisfy

$$z^2\, J_n'' + z\, J_n' + (z^2 - n^2)\, J_n = 0\,. \tag{1.3}$$

To do this, consider

$$
\begin{aligned}
&\sum_{n=-\infty}^{\infty} \Big(z^2\, J_n'' + z\, J_n' + (z^2 - n^2)\, J_n\Big)\, t^n \\
&\quad = \sum_{n=-\infty}^{\infty} \Big(z^2\, \frac{d^2}{dz^2} + z\, \frac{d}{dz} + z^2 - t\, \frac{d}{dt}\, t\, \frac{d}{dt}\Big)\, t^n\, J_n \\
&= \Big(z^2\, \frac{d^2}{dz^2} + z\, \frac{d}{dz} + z^2 - t\, \frac{d}{dt}\, t\, \frac{d}{dt}\Big)\, e^{\frac{1}{2} z (t - t^{-1})}\,, \\
&= \Big(\tfrac{1}{4} z^2\, (t - t^{-1})^2 + \tfrac{1}{2} z\, (t - t^{-1}) + z^2 - \tfrac{1}{4} z\, t^{-2}\, (-2t + 2t^3 + z + 2z\, t^2 + z\, t^4)\Big)\, e^{\frac{1}{2} z (t - t^{-1})} \\
&= 0\,. \qquad (1.4)
\end{aligned}
$$

Note that in the first line, we have used the fact that $n^2\, t^n$ can be written as $t(d/dt)t(d/dt)\, t^n$.

The next step is to observe that (1.2) can be turned into an expression for a single Bessel function, say $J_m(z)$. All we need to do is to multiply (1.2) by t^{-m-1}, and integrate it around a closed contour C encircling the origin. By the theorem of residues, we have

$$
\frac{1}{2\pi\,\mathrm{i}} \oint_C t^{n-m-1}\, dt = \delta_{mn}\,, \tag{1.5}
$$

where the Kronecker delta function δ_{mn} as usual has the meaning that $\delta_{mn} = 0$ unless $m = n$, for which $\delta_{mm} = 1$. Thus from (1.2) we obtain the result that

$$
J_n(z) = \frac{1}{2\pi\,\mathrm{i}} \oint_C t^{-n-1}\, e^{\frac{1}{2} z (t - t^{-1})}\, dt\,, \tag{1.6}
$$

where C is a closed contour that encircles the origin anticlockwise. We can, for example, take C to be C_0, the unit circle, $|t| = 1$. This has furnished us with an integral representation for the Bessel function $J_n(z)$. It is evident that it is analytic for all z in the finite complex plane. The J_n functions are sometimes called *Bessel Functions of the First Kind.* For now, we are assuming that n is an integer.

We can express $J_n(z)$ as a power series in z in the following way. Introduce a new integration variable w, defined by $t = 2w/z$; then

$$
J_n(z) = \frac{1}{2\pi\,\mathrm{i}} \left(\tfrac{1}{2} z\right)^n \oint_C w^{-n-1}\, e^{w - \frac{1}{4} z^2\, w^{-1}}\, dw\,, \tag{1.7}
$$

where again we may take the integration contour to be the unit circle, $|w| = 1$. The factor $e^{-\frac{1}{4} z^2\, w^{-1}}$ can be expanded in a power series,

$$
e^{-\frac{1}{4} z^2\, w^{-1}} = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} \left(\tfrac{1}{2} z\right)^{2r} w^{-r}\,, \tag{1.8}
$$

since this is uniformly convergent on the circle $|w| = 1$. Thus we obtain

$$
J_n(z) = \frac{1}{2\pi\,\mathrm{i}} \sum_{r=0}^{\infty} \frac{(-1)^r}{r!} \left(\tfrac{1}{2} z\right)^{n+2r} \oint_C w^{-n-r-1}\, e^w\, dw\,. \tag{1.9}
$$

As we saw in part I of the course, the residue $\mathcal{R}$ at an N'th-order pole $z = z_0$ of a function $f(z)$ is

$$\mathcal{R} = \frac{1}{(N-1)!} \Big[\frac{d^{N-1}}{dz^{N-1}} \Big((z-z_0)^N\, f(z)\Big)\Big]_{z=z_0}\,. \tag{1.10}$$

Therefore the residue of the integrand in (1.9) at $w = 0$ is given by differentiating e^w $(n+r)$ times, setting $w = 0$, and dividing by $(n + r)!$, when $n + r$ is a positive integer or zero. When $n+r$ is a negative integer (recall that n can be positive, negative or zero), the residue is zero.

Consequently, we find that if n is a positive integer or zero, (1.9) gives

$$J_n(z) = \sum_{r=0}^\infty \frac{(-1)^r\, \big(\frac{1}{2}z\big)^{n+2r}}{r!\,(n+r)!}\,. \tag{1.11}$$

On the other hand if n is a negative integer, $n = -m$, then

$$J_n(z) = \sum_{r=m}^\infty \frac{(-1)^r\, \big(\frac{1}{2}z\big)^{2r-m}}{r!\,(r-m)!} = \sum_{s=0}^\infty \frac{(-1)^{m+s}\, \big(\frac{1}{2}z\big)^{m+2s}}{s!\,(m+s)!}\,, \tag{1.12}$$

where we set $r = m+s$ in the second summation. Evidently, therefore, we have the relation

$$J_{-n}(z) = (-1)^n\, J_n(z)\,, \tag{1.13}$$

where n is any integer.

Notice that by having a variety of ways of representing the Bessel functions available in the armoury, we can pick whichever is most convenient for proving a particular result. In fact the property (1.13) can be seen very easily directly from (1.2). If we send $t \longrightarrow -1/t$ then the effect on the right-hand side is to send $J_n(z) \longrightarrow (-1)^n\, J_{-n}(z)$, while the left-hand side is left unchanged.

Bessel functions have many properties that are analogous to those of trigonometric functions. Recall, for example, the addition formulae such as $\sin(x + y) = \sin x\, \cos y + \cos x\, \sin y$. The analogue for the J_n Bessel functions is

$$J_n(x+y) = \sum_{m=-\infty}^\infty J_m(x)\, J_{n-m}(y)\,. \tag{1.14}$$

We can again prove this very easily from the generating function (1.2). We simply observe that from the elementary properties of the exponential function, it follows that

$$e^{\frac{1}{2}(x+y)(t-t^{-1})} = e^{\frac{1}{2}x\,(t-t^{-1})}\, e^{\frac{1}{2}y\,(t-t^{-1})}\,. \tag{1.15}$$

From (1.2) this implies

$$\sum_{n=-\infty}^\infty t^n\, J_n(x+y) = \Big(\sum_{p=-\infty}^\infty t^p\, J_p(x)\Big)\,\Big(\sum_{q=-\infty}^\infty t^q\, J_q(y)\Big)\,. \tag{1.16}$$

Picking out all the terms associated with $p+q=n$ in the right-hand side, and equating to the term in t^n on the left-hand side, equation (1.14) follows.

Another integral representation for the Bessel function $J_n(z)$ may be obtained as follows. Starting from (1.6), we may write the complex integration variable t, which is taken to run around the unit circle, as $t = e^{\mathrm{i}\,\theta}$. Thus we get

$$J_n(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\mathrm{i}\, n\,\theta + \mathrm{i}\, z\, \sin\theta}\, d\theta. \tag{1.17}$$

By dividing the integration range into two pieces, namely $-\pi \le \theta \le 0$ and $0 \le \theta \le \pi$, and then sending $\theta \longrightarrow -\theta$ in the first of these, we get

$$J_n(z) = \frac{1}{2\pi} \int_0^{\pi} e^{\mathrm{i}\, n\,\theta - \mathrm{i}\, z\, \sin\theta}\, d\theta + \frac{1}{2\pi} \int_0^{\pi} e^{-\mathrm{i}\, n\,\theta + \mathrm{i}\, z\, \sin\theta}\, d\theta\,, \tag{1.18}$$

and hence we arrive at the expression, known as Bessel's integral for $J_n(z)$:

$$J_n(z) = \frac{1}{\pi} \int_0^{\pi} \cos(n\,\theta - z\, \sin\theta)\, d\theta\,. \tag{1.19}$$

To give some idea of what the Bessel functions $J_n(z)$ look like, we give plots below, in Figures 1, 2, 3 and 4, for $J_0(z)$, $J_1(z)$, $J_5(z)$ and $J_{10}(z)$. Like the trigonometric functions they are oscillatory, although they are not periodic as such since the interval between successive zeros changes with z. As we shall see later, at large z they do asymptotically approach a definite period. It is also evident that their magnitudes fall off, in a rather mild way, as z increases.

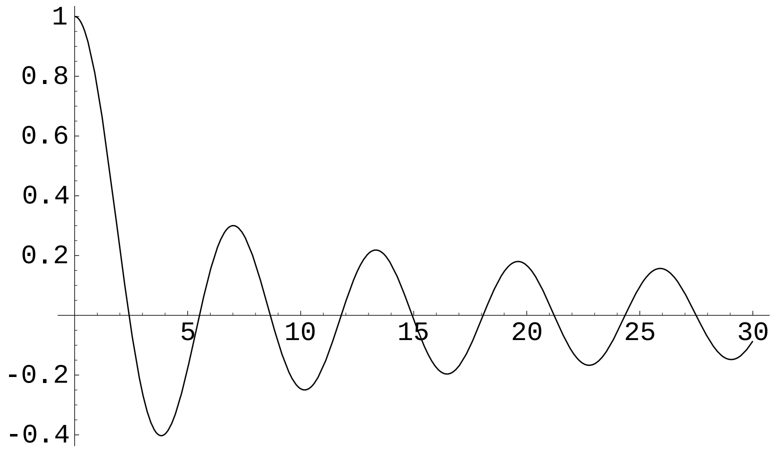

Figure 1: The $J_0(z)$ Bessel Function

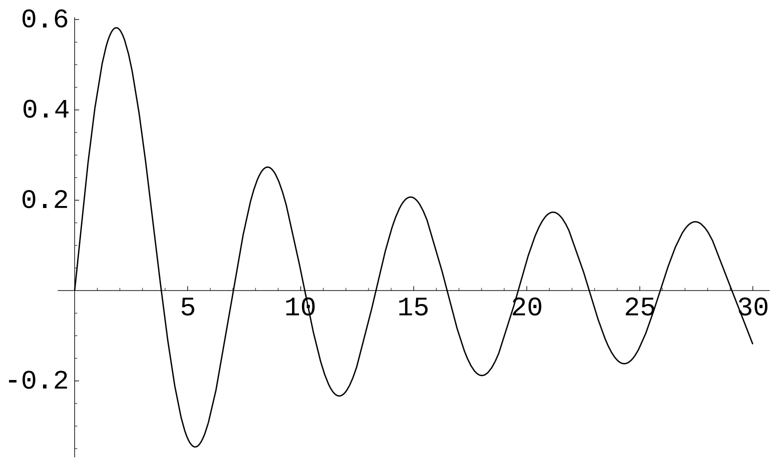

Figure 2: The $J_1(z)$ Bessel Function

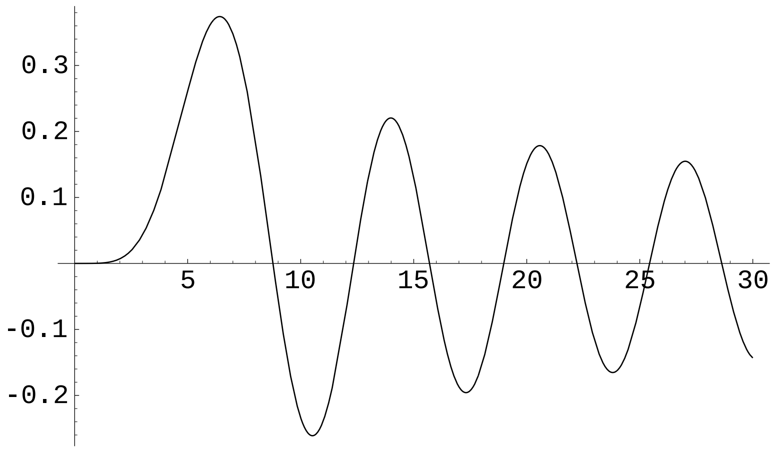

Figure 3: The $J_5(z)$ Bessel Function

1.2 $J_\nu(z)$ Bessel Function of Non-integer Order ν

Until now, we have been assuming that the order n of $J_n(z)$ is an integer. Staying with this assumption for just a moment longer, we may note from the integral representation (1.7) that we can directly substitute it into the Bessel equation (1.3), to obtain

$$\begin{aligned} J_n'' + \frac{1}{z}\, J_n' + \Big(1 - \frac{n^2}{z^2}\Big)\, J_n &= \frac{1}{2\pi\,\mathrm{i}}\, \big(\tfrac{1}{2}z\big)^n \oint_C w^{-n-1} \Big[1 - \frac{n+1}{w} + \frac{z^2}{4w^2}\Big]\, e^{w - \frac{1}{4}z^2\, w^{-1}}\, dw\,, \\ &= -\frac{1}{2\pi\,\mathrm{i}}\, \big(\tfrac{1}{2}z\big)^n \oint_C \frac{d}{dw}\Big[w^{-n-1}\, e^{w-\frac{1}{4}z^2\, w^{-1}}\Big]\, dw\,, \\ &= 0\,. \end{aligned} \tag{1.20}$$

This last step follows from the fact that $w^{-n-1}\, e^{w-\frac{1}{4}z^2\, w^{-1}}$ is single valued, and so it returns to its original value after completing the trip around the closed contour C, which was taken to be the unit circle C_0. This gives a direct proof that the integral repsesentation (1.7) for

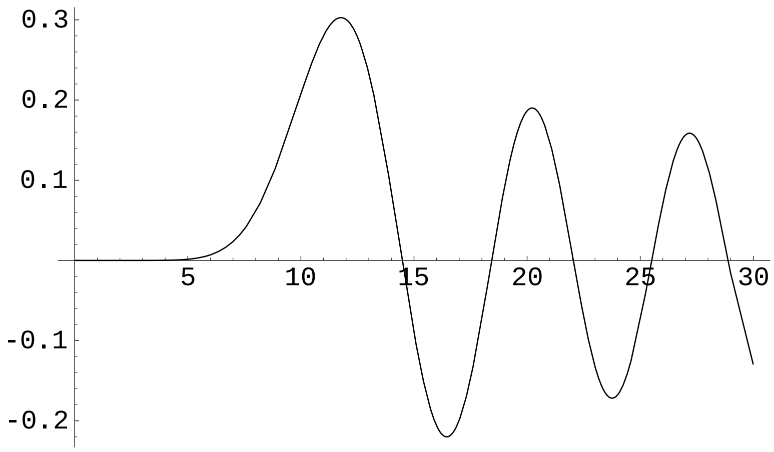

Figure 4: The $J_{10}(z)$ Bessel Function

the Bessel function of integral order satisfies Bessel's equation.

Now, a straightforward modification allows us to adopt (1.7) as an integral representation for the Bessel function $J_\nu(z)$, where now ν is not restricted to being an integer. It is evident that a manipulation identical to (1.20) can be carried out for $J_\nu(z)$ defined by

$$J_\nu(z) = \frac{z^\nu}{2^{\nu+1}\,\pi\,\mathrm{i}} \int_C w^{-\nu-1}\, e^{w-\frac{1}{4}z^2\,w^{-1}}\, dw\,, \tag{1.21}$$

provided that we make an appropriate different choice for the contour C. (We shall keep the same symbol C, but it will now mean something different.) Thus we substitute (1.21) into (1.1), deducing that $J_\nu(z)$ does indeed satisfy this equation as long as

$$\int_C \frac{d}{dw}\left[w^{-\nu-1}\, e^{w-\frac{1}{4}z^2\,w^{-1}}\right] dw = 0\,. \tag{1.22}$$

This will be true provided that the quantity

$$w^{-\nu-1}\, e^{w-\frac{1}{4}z^2\,w^{-1}} \tag{1.23}$$

returns to its initial value after following round from the beginning to the end of the path described by C. Clearly, when ν is not an integer, we cannot take C to be the unit circle any more. Instead, we can take C to be very like the Hankel contour that we used in part I of the course, only now reflected across the imaginary axis. Thus we take a contour that starts at $-\infty$ just below the real axis, loops anticlockwise around the origin, and exits to the west again just above the real axis; see Figure 7 below . At both the starting and finishing points, therefore, the real part of w is $-\infty$, and so the e^w factor ensures that (1.23) vanishes at both ends. To be precise, we take $|\arg w| \le \pi$ on the contour.

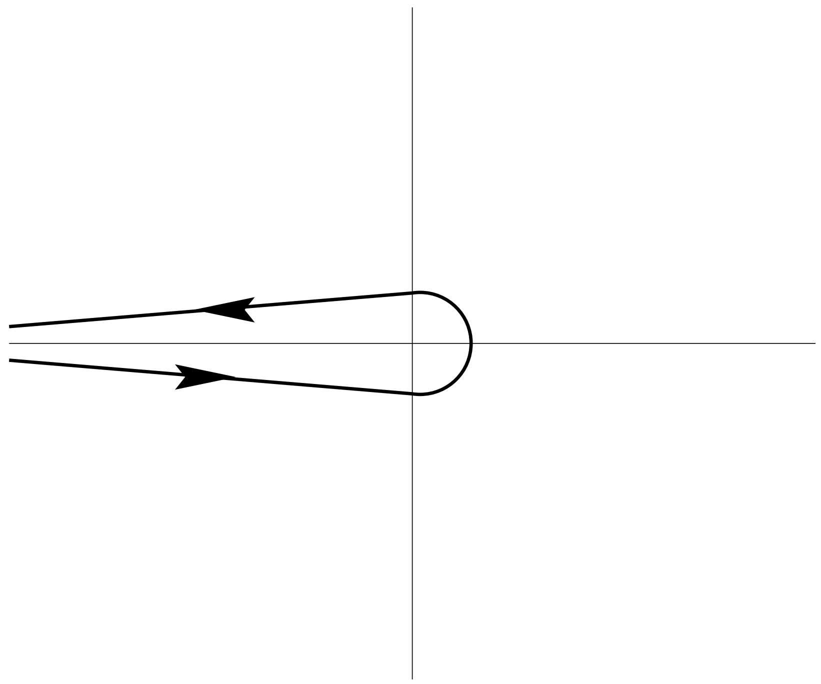

Figure 5: The contour of integration for the integral (1.21) for $J_\nu(z)$

This integral representation for $J_\nu(z)$ can be expressed as a power series. We may note that the integral itself in (1.21) defines an analytic function z, and so it must admit a Taylor expansion. In fact, the integral has a series expansion in powers of $q \equiv z^2$, which can be obtained by differentiating under the integral sign, to construct the Taylor expansion. Definining

$$h(q) \equiv \int_C w^{-\nu-1}\, e^{w - \frac{1}{4} q\, w^{-1}}\, dw\,, \tag{1.24}$$

we construct the series expansion

$$\begin{aligned} h(q) &= h(0) + q\, h'(0) + \tfrac{1}{2} q^2\, h''(0) + \tfrac{1}{6} q^3\, h'''(0) + \cdots = \sum_{r=0}^\infty \frac{q^r}{r!}\, h^{(r)}(0)\,,\\ &= \sum_{r=0}^\infty \frac{(-q)^r}{4^r\, r!} \int_C w^{-\nu-r-1}\, e^w\, dw\,,\\ &= 2\pi\,\mathrm{i} \sum_{r=0}^\infty \frac{(-q)^r}{4^r\, r!\, \Gamma(\nu+r+1)}\,. \end{aligned} \tag{1.25}$$

This last result comes from the contour-integral expression for the Gamma function that we derived in part I of the course, namely

$$\frac{1}{\Gamma(z)} = -\frac{1}{2\pi\,\mathrm{i}} \int_\gamma e^{-t}\, (-t)^{-z}\, dt\,, \tag{1.26}$$

where γ denotes the Hankel contour, which runs from $+\infty$ just above the real axis, swings in around the origin, and goes out east again just below the real axis. (This is just the reflection of our current contour C across the imaginary axis.) Thus we arrive at the result that $J_\nu(z)$ has the series expansion

$$J_\nu(z) = \sum_{r=0}^\infty \frac{(-1)^r\, z^{\nu+2r}}{2^{\nu+2r}\, r!\, \Gamma(\nu+r+1)}\,. \tag{1.27}$$

It is easy to see that this expansion agrees with the one that we derived in (1.11), in the case that ν is a non-negative integer. It also coincides with (1.12) in the case that ν is a negative integer. In general, for arbitrary ν we take (1.21) as the integral representation defining $J_\nu(z)$, and (1.27) as the series representation for $J_\nu(z)$.

Notice that since $J_\nu(z)$ satisfies Bessel's equation (1.1), and this equation is invariant under sending $\nu \longrightarrow -\nu$, it follows that $J_\nu(z)$ and $J_{-\nu}(z)$ generically give us the two linearly-independent solutions of the Bessel equation. This argument would break down, of course, if it were the case that $J_{-\nu}(z)$ were simply a constant multiple of $J_\nu(z)$. We know that this is precisely what *does* happen if ν is an integer, since then we have the relation (1.13) which tells us that $J_{-n}(z) = (-1)^n\, J_n(z)$. This is, however, a peculiarity of integer values for ν. When $\nu \ne$ integer, it is clear from (1.27) that $J_{-\nu}(z)$ cannot be a constant multiple of $J_\nu(z)$. (The powers of z in the expansions of $J_\nu(z)$ and $J_{-\nu}(z)$ will be completely different.) Thus when $\nu \ne$ integer, the general solution of the Bessel equation (1.1) is given by

$$\alpha\, J_\nu(z) + \beta\, J_{-\nu}(z)\,, \tag{1.28}$$

where α and β are constants. We shall see later how to obtain the second independent solution to (1.1) when ν is an integer.

Here are a couple of sample plots of Bessel functions $J_\nu(z)$ with non-integer order ν. We present the cases $\nu = \frac{1}{3}$ and $\nu = -\frac{1}{3}$, in Figures 5 and 6 below.

We may generalise the Bessel integral (1.19) for the integer-order Bessel functions to the case where the order is non-integral. First, we note that by performing the transformation $w = \frac{1}{2} z\, t$, we can cast the integral representation (1.21) into the form

$$J_\nu(z) = \frac{1}{2\pi\, \mathrm{i}} \int_C t^{-\nu-1}\, e^{\frac{1}{2} z(t - t^{-1})}\, dt\,. \tag{1.29}$$

This will be an analytic function of z provided that $\mathrm{Re}(z\, t)$ is negative when t heads of to $-\infty$ at the beginning and end of the contour. We shall deform the contour so that it consists of a line running from $-\infty$ to -1 just below the real axis, then a unit circle running anticlockwise around the origin, and finally a line running from -1 to $-\infty$ just above the

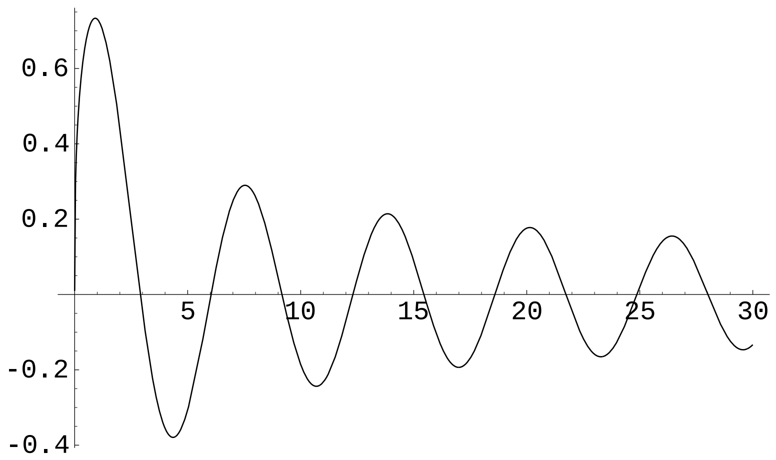

Figure 6: The $J_{\frac{1}{3}}(z)$ Bessel Function

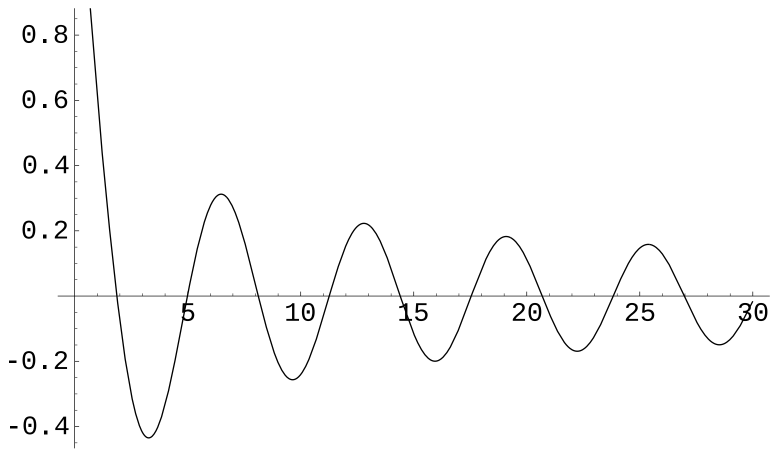

Figure 7: The $J_{-\frac{1}{3}}(z)$ Bessel Function

real axis. (See Figure 8 below.) Initially we shall take z to be real and positive, but by analytic continuation we may then allow z to be any complex number with $\mathrm{Re}(z) > 0$.

The part of the contour comprising the unit circle can be handled precisely as in the case of the integer-order result (1.19). The two line integrals give additional contributions

$$\Big[\frac{e^{(\nu+1)\,\pi\,\mathrm{i}}}{2\pi\,\mathrm{i}} - \frac{e^{-(\nu+1)\,\pi\,\mathrm{i}}}{2\pi\,\mathrm{i}}\Big] \int_1^\infty x^{-\nu-1}\, e^{\frac{1}{2}z(-x+x^{-1})}\, dx\,, \tag{1.30}$$

where we have written $t = e^{\mp\mathrm{i}\,\pi}\, x$ for the ingoing and outgoing pieces respectively. Thus writing $x = e^\theta$, we arrive at the result, due to Schläfli, that

$$J_\nu(z) = \frac{1}{\pi} \int_0^\pi \cos(\nu\,\theta - z\,\sin\theta)\, d\theta - \frac{\sin\nu\pi}{\pi} \int_0^\infty e^{-\nu\,\theta - z\,\sinh\theta}\, d\theta\,. \tag{1.31}$$

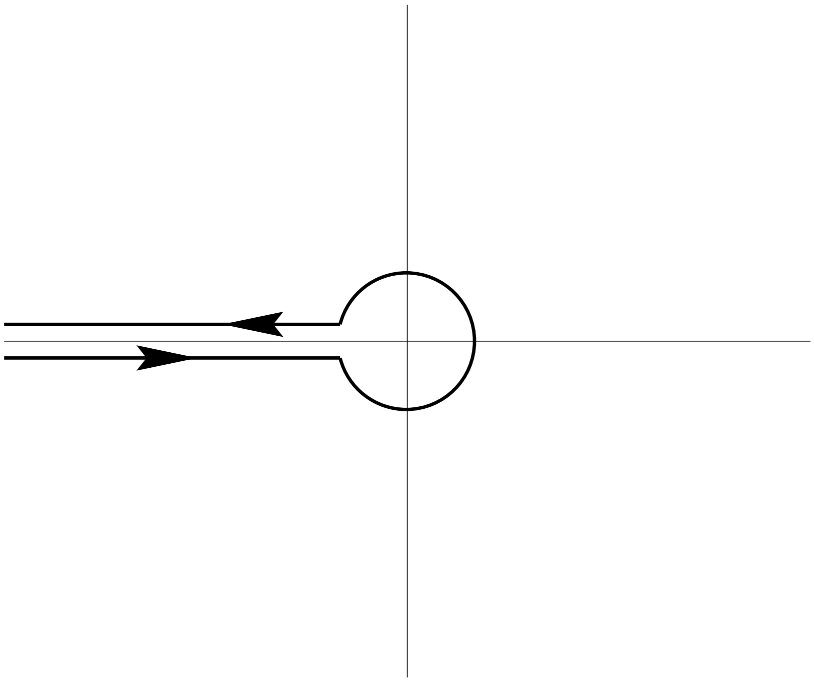

Figure 8: The deformed contour for deriving Schläfli's integral

Notice that in the special case where ν is an integer, this reduces immediately to the previous result (1.19).

1.3 Recurrence Formulae for the Bessel Functions

Notice that from the integral representation (1.21) for the Bessel function $J_\nu(z)$, we can derive a simple expression for obtaining $J_{\nu+1}(z)$ in terms of $J_\nu(z)$. To do this, multiply (1.21) by $z^{-\nu}$ and differentiate with respect to z, to get

$$\begin{aligned} \frac{d}{dz}\Big(z^{-\nu}\,J_\nu(z)\Big) &= \frac{1}{2^{\nu+1}\,\pi\,\mathrm{i}}\,\frac{d}{dz}\int_C w^{-\nu-1}\,e^{w-\frac{1}{4}z^2\,w^{-1}}\,dw\,,\\ &= -\frac{z}{2^{\nu+2}\,\pi\,\mathrm{i}}\int_C w^{-\nu-2}\,e^{w-\frac{1}{4}z^2\,w^{-1}}\,dw\,,\\ &= -z^{-\nu}\,J_{\nu+1}(z)\,. \end{aligned} \tag{1.32}$$

In other words, we have

$$J_{\nu+1}(z) = -z^\nu\,\frac{d}{dz}\Big(z^{-\nu}\,J_\nu(z)\Big)\,, \tag{1.33}$$

which can trivially be written also as

$$J_{\nu+1}(z) = -z^{\nu+1}\,\frac{d}{z\,dz}\Big(z^{-\nu}\,J_\nu(z)\Big)\,, \tag{1.34}$$

Iterating (1.34) once, we get

$$\begin{aligned} J_{\nu+2}(z) &= -z^{\nu+2}\,\frac{d}{z\,dz}\Big(z^{-\nu}\,J_{\nu+1}(z)\Big) \\ &= z^{\nu+2}\,\frac{d}{z\,dz}\Big(\frac{d}{z\,dz}\Big(z^{-\nu}\,J_\nu(z)\Big)\Big)\,. \end{aligned} \tag{1.35}$$

Clearly we can repeat this as many times as we wish, to obtain the recurrence formula

$$J_{\nu+r}(z) = (-1)^r\,z^{\nu+r}\,\Big[\frac{d}{z\,dz}\Big]^r\,\Big(z^{-\nu}\,J_\nu(z)\Big)\,, \tag{1.36}$$

where r is any non-negative integer.

Another recurrence formula can be obtained by considering $J_{\nu+1}(z)+J_{\nu-1}(z)$, which, from (1.21), can be written as

$$\begin{aligned} J_{\nu+1}(z)+J_{\nu-1}(z) &= \frac{z^\nu}{2^{\nu+1}\,\pi\,\mathrm{i}}\int_C (\tfrac12 z\,w^{-1}+2w\,z^{-1})\,w^{-\nu-1}\,e^{w-\frac14 z^2\,w^{-1}}\,dw\,, \\ &= \frac{2}{z}\,\frac{z^\nu}{2^{\nu+1}\,\pi\,\mathrm{i}}\int_C w^{-\nu}\,\Big(1+\frac{z^2}{4w^2}\Big)\,e^{w-\frac14 z^2\,w^{-1}}\,dw\,, \\ &= \frac{2}{z}\,\frac{z^\nu}{2^{\nu+1}\,\pi\,\mathrm{i}}\int_C w^{-\nu}\,\frac{d}{dw}\,e^{w-\frac14 z^2\,w^{-1}}\,dw\,, \\ &= \frac{2\nu}{z}\,\frac{z^\nu}{2^{\nu+1}\,\pi\,\mathrm{i}}\int_C w^{-\nu-1}\,e^{w-\frac14 z^2\,w^{-1}}\,dw\,, \end{aligned} \tag{1.37}$$

where in the last line we integrated by parts, and made use of the fact that the "boundary term" in the integration by parts vanishes. (This is the same property that we used previously in order to show that $J_\nu(z)$ defined by (1.21) satisfied the Bessel equation.) Thus we have obtained the recurrence formula

$$J_{\nu+1}(z)+J_{\nu-1}(z) = \frac{2\nu}{z}\,J_\nu(z)\,. \tag{1.38}$$

1.4 Bessel Functions of Half-integer Order

The Bessel functions $J_\nu(z)$ take on a particularly simple form when ν is half an odd integer. Consider the case when $\nu=\frac12$. In general we have the series expansion (1.27), namely

$$J_\nu(z) = \sum_{r=0}^\infty \frac{(-1)^r\,z^{\nu+2r}}{2^{\nu+2r}\,r!\,\Gamma(\nu+r+1)}\,. \tag{1.39}$$

Setting $\nu=\frac12$, we may observe first that

$$\begin{aligned} \Gamma(\tfrac12+r+1) &= (\tfrac12+r)\,\Gamma(\tfrac12+r) = (\tfrac12+r)(\tfrac12+r-1)\,\Gamma(\tfrac12+r-1)\,, \\ &= (\tfrac12+r)(\tfrac12+r-1)\cdots\tfrac12\,\cdot\Gamma(\tfrac12)\,, \\ &= 2^{-r-1}\,(2r+1)(2r-1)(2r-3)\cdots 3\cdot 1\,\cdot\Gamma(\tfrac12)\,. \end{aligned} \tag{1.40}$$

Furthermore, we may write

$$r! = 2^{-r}\,(2r)\,(2r-2)(2r-4)\cdots 4\cdot 2\,. \tag{1.41}$$

Combined with the fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we therefore have that

$$r!\,\Gamma(\tfrac{1}{2}+r+1) = 2^{-2r-1}\,\sqrt{\pi}\,(2r+1)!\,. \tag{1.42}$$

Substituting into (1.39), we therefore obtain

$$J_{\frac{1}{2}}(z) = \sqrt{\frac{2z}{\pi}}\,\sum_{r=0}^{\infty}\frac{(-1)^r\,z^{2r}}{(2r+1)!}\,, \tag{1.43}$$

whence

$$J_{\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi\,z}}\,\sin z\,. \tag{1.44}$$

From our previous recurrence formula (1.36), it then immediately follows that

$$\begin{aligned} J_{r+\frac{1}{2}}(z) &= \sqrt{\frac{2}{\pi}}\,z^{r+\frac{1}{2}}\,\Big[\frac{d}{z\,dz}\Big]^r\,\Big(\frac{\sin z}{z}\Big)\,,\\ &= \frac{1}{\sqrt{\pi}}\,(2z)^{r+\frac{1}{2}}\,\Big[\frac{d}{dz^2}\Big]^r\,\Big(\frac{\sin z}{z}\Big)\,, \end{aligned} \tag{1.45}$$

where r is any non-negative integer. It is clear after a moment's thought that this means that

$$J_{r+\frac{1}{2}}(z) = P_r(z)\,\sin z + Q_r(z)\,\cos z\,, \tag{1.46}$$

where $P_r(z)$ and $Q_r(z)$ are polynomials in $z^{-\frac{1}{2}}$.

1.5 The Second Solution of Bessel's Equation

We saw previously that if ν is not an integer, the Bessel functions $J_\nu(z)$ and $J_{-\nu}(z)$ are linearly independent, and both solve the Bessel equation (1.1). Being a second-order differential equation, the Bessel equation has exactly two linearly independent solutions, and so they may be taken to be $J_\nu(z)$ and $J_{-\nu}(z)$ when ν is non-integral.

When ν is an integer n the above reasoning fails because, as we saw in (1.13), $J_n(z)$ and $J_{-n}(z)$ *are* now linearly dependent; $J_n(z) = (-1)^n\,J_{-n}(z)$. As is often the case when the "second solution" of a differential degenerates for some special values of the parameters, one can in fact still extract the second solution by taking an appropriately rescaled limit. In the present case, we do this by a construction in which we take the difference between the $J_\nu(z)$ and $J_{-\nu}(z)$ solutions, divide by a quantity that vanishes appropriately at $\nu =$ integer, and then take the limit where ν tends to the integer n. The idea is that the vanishing denominator scales up the numerator that is otherwise tending to zero, so that a finite and non-zero result is obtained.

To be precise this second solution, known, not surprisingly, as the Bessel function of the second kind, and denoted by $Y_\nu(z)$, is defined by

$$Y_\nu(z) = \frac{J_\nu(z)\,\cos\nu\pi - J_{-\nu}(z)}{\sin\nu\pi}\,. \tag{1.47}$$

First, note that for a generic (non-integer) value of z, $Y_\nu(z)$ is just a certain linear combination of $J_\nu(z)$ and $J_{-\nu}(z)$, with the coefficients of both terms being finite and non-zero. Thus when ν is non-integral, $Y_\nu(z)$ is a perfectly good choice for the second solution of the Bessel equation.[1]

Now, consider what happens when ν is taken to be an integer, n. The numerator becomes precisely the combination $(-1)^n\,J_n(z) - J_{-n}(z)$ that vanishes by virtue of the relation (1.13). However, as promised, the denominator vanishes too. We end up, as ν is sent to n, with a "zero divided by zero" expression that actually has a regular limit. Of course given that this limit exists, which we shall show in a moment, it follows that $Y_n(z)$ solves the Bessel equation, since $Y_\nu(z)$ solves it for all non-integer ν, and this will continue to be true as ν approaches the integer n. So it remains to show that the limit does indeed exist, and that the resulting function $Y_n(z)$ is linearly independent of $J_n(z)$.

We can show both of these properties together, in fact. Recall that the Wronskian of two solutions y_1 and y_2 of a second-order linear differential equation is defined by

$$\Delta(y_1, y_2) \equiv y_1\,y_2' - y_2\,y_1'\,. \tag{1.48}$$

Recall also that the Wronskian of the two solutions is non-vanishing if and only if the solutions are linearly independent.

For the Bessel equation, if

$$\begin{aligned} z^2\,y_1'' + z\,y_1' + (z^2-\nu^2)\,y_1 &= 0\,,\\ z^2\,y_2'' + z\,y_2' + (z^2-\nu^2)\,y_2 &= 0\,, \end{aligned} \tag{1.49}$$

then multiplying the second equation by y_1 and subtracting the first equation multiplied by y_2 from it, we get

$$z^2\,(y_1\,y_2'' - y_2\,y_1'') + z\,(y_1\,y_2' - y_2\,y_1') = 0\,, \tag{1.50}$$

whence

$$z\,\Delta' + \Delta = 0\,. \tag{1.51}$$

[1]Sometimes $Y_\nu(z)$ is known as the Neumann function, and is denoted instead by $N_\nu(z)$.

This can be immediately solved for the Wronskian, giving $\log\Delta + \log z = \text{constant}$, or in other words

$$\Delta = \frac{c}{z}\,, \tag{1.52}$$

where c is a constant. So the question of linear independence comes down to whether in a particular case the constant c turns out to be zero or not.

Let us first consider the Wronskian of $J_\nu(z)$ and $J_{-\nu}(z)$. We expect to find that it is non-zero when ν is not an integer, but that it becomes zero when ν is an integer. Let's see if this is what happens. Since we have established the result (1.52), we have only to determine the constant c (which we expect to be dependent on ν, but, of course, independent of z.) We can fix c for the case $y_1 = J_\nu(z)$, $y_2 = J_{-\nu}(z)$ by looking at any convenient range of the coordinate z; the most convenient thing is to look at the place where z is very small, since this allows us to use just the leading-order terms in the series expansions of the Bessel functions.

We have from (1.27) that

$$\begin{aligned} J_\nu(z) &= \frac{2^{-\nu}}{\Gamma(1+\nu)}\, z^\nu + O(z^{\nu+2})\,,\\ J_{-\nu}(z) &= \frac{2^{\nu}}{\Gamma(1-\nu)}\, z^{-\nu} + O(z^{-\nu+2})\,, \end{aligned} \tag{1.53}$$

Therefore, substituting into (1.48), we find that

$$\Delta(J_\nu, J_{-\nu}) = -\frac{2\nu}{z\,\Gamma(1+\nu)\Gamma(1-\nu)} + O(1)\,. \tag{1.54}$$

Of course since we know that $J_\nu(z)$ and $J_{-\nu}(z)$ satisfy the Bessel equation, and that Δ must be of the form (1.52) for any two solutions, this means that the higher-order terms represented by $O(1)$ are actually zero. The point is, though, that we can be sure that *only* the leading-order terms that we displayed explicitly in (1.53) contribute to the $O(1/z)$ result. (The higher terms from (1.53) would obviously contribute to Δ at orders z^s with $s \ge 0$.)

Now, we use some standard properties of the Gamma function that were proved in Part I of the course, namely

$$x\,\Gamma(x) = \Gamma(x+1)\,, \qquad \Gamma(x)\,\Gamma(1-x) = \frac{\pi}{\sin\pi x}\,. \tag{1.55}$$

Putting these together, we learn that $\Gamma(1+\nu)\,\Gamma(1-\nu) = \nu\,\pi/\sin(\nu\,\pi)$, and so (1.54) becomes

$$\Delta(J_\nu, J_{-\nu}) = -\frac{2\sin\nu\pi}{\pi\,z}\,. \tag{1.56}$$

So, comparing with (1.52), we have

$$c = -\frac{2\sin\nu\pi}{\pi}\,. \tag{1.57}$$

Thus we have found the expected result, namely that J_ν and $J_{-\nu}$ are linearly independent for all ν except when ν is an integer.

Now consider the Wronskian $\Delta(J_\nu, Y_\nu)$ of J_ν and Y_ν, defined in (1.47). Clearly since $\Delta(J_\nu, J_\nu)$ is *always* zero, this will simply be given by the contribution from the second term in Y_ν:

$$\Delta(J_\nu, Y_\nu) = -\frac{1}{\sin\nu\pi}\,\Delta(J_\nu, J_{-\nu}) = \frac{2}{\pi\, z}\,. \tag{1.58}$$

In the final stage here, we have substituted our previous result for $\Delta(J_\nu, J_{-\nu})$.

Our expression (1.58) shows that $J_\nu(z)$ and $Y_\nu(z)$ are linearly-independent for *all* values of ν, integer and non-integer. This is what we wanted to show. Also, the fact that the Wronskian in (1.58) has turned out to be a finite and non-zero constant multiple of $1/z$ shows that our limiting procedure to construct $Y_\nu(z)$ at integer ν is a good one; it has produced a function that has neither diverged nor vanished.

Let us investigate the properties of $Y_\nu(z)$ a little further. For now, we shall restrict attention to looking at the behaviour near $z = 0$. We have already seen how the $J_\nu(z)$ Bessel function behaves, in the power-series expansion (1.27). Writing out the first few terms for $J_\nu(z)$, we see that it is

$$J_\nu(z) = \frac{z^\nu}{2^\nu\,\Gamma(\nu+1)}\Big[1 - \frac{z^2}{4(\nu+1)} + \frac{z^4}{4^2\,(\nu+1)(\nu+2)} - \frac{z^6}{4^3\,(\nu+1)(\nu+2)(\nu+3)} + \cdots\Big]\,. \tag{1.59}$$

Now, in Part I of the course, we discussed how one in general constructs the second independent solution of a second-order linear ODE in terms of a given original solution. In particular, we saw that given a solution $y_1(z)$, and Wronskian Δ, then the second solution $y_2(z)$ is obtained as

$$y_2(z) = y_1(z)\int^z \frac{\Delta(t)}{y_1(t)^2}\,dt\,. \tag{1.60}$$

Of course if one takes different values for the constant lower limit of integration here, one gets different constant multiples of the original solution $y_1(z)$ added to the second solution $y_2(z)$. This is to be expected; if $y_2(z)$ is a solution linearly independent of $y_1(z)$, then so is $y_2(z) + \alpha\, y_1(z)$ for any constant α.

From this discussion, it follows that with an appropriate choice of the lower limit of integration, we must have that

$$Y_\nu(z) = \frac{2}{\pi}\,J_\nu(z)\int^z \frac{1}{t\,J_\nu(t)^2}\,dt\,. \tag{1.61}$$

Here, we have substituted the result (1.58) for the Wronskian of $J_\nu(z)$ with $Y_\nu(z)$. Now, we may take the series expansion for $J_\nu(z)$ given in (1.59), and substitute it into (1.61):

$$Y_\nu(z) = \frac{2^{2\nu+1}\,\Gamma(\nu+1)^2\,J_\nu(z)}{\pi} \int^z t^{-2\nu-1}\Big[1 + \frac{t^2}{2(\nu+1)} + \frac{(2\nu+5)\,t^4}{16(\nu+1)^2\,(\nu+2)} + \cdots\Big]\,. \quad (1.62)$$

For generic (i.e. non-integer) values of ν, it is clear that term-by-term integration of the integral in (1.62) will just generate powers of z of the form $z^{-2\nu}$, $z^{-2\nu+2}$, $z^{-2\nu+4}$, *etc.*. In fact, we know that at the end of the day the result must be that the entire expression in (1.62) just produces some linear combination of $J_\nu(z)$ and $J_{-\nu}(z)$, since these are the two linearly independent solutions of Bessel's equation when ν is not an integer.

However, when $\nu = n =$ integer, it is evident that there will always be a particular term in the integrand in (1.62) that is of the form t^{-1}. For example, if $\nu = 0$ it will be the first term in the square brackets that gives t^{-1}. If $\nu = 1$, it will be the second term that gives t^{-1}, and so on. The point is that whenever ν is an integer, we are finding that the integral in (1.62) yields a logarithm, since

$$\int^z t^{-1}\,dt = \log z\,. \quad (1.63)$$

Thus we have learned that when $\nu = n$ is an integer, the second solution $Y_n(z)$ always has a logarithmic divergence as z tends to zero. This logarithmic behaviour is in fact precisely what is expected from a general analysis of the properties of the second solution of a differential equation expanded around a regular singular point (see the discussion in Part 1 of the course).

In order to obtain the full structure of the small-z series expansion for $Y_\nu(z)$, it is easiest to go back to the original definition (1.47). As we have seen above, the nature of the expansion will depend significantly on whether or not ν is an integer, since there will be logarithims involved if ν is an integer, but not otherwise. In fact, we are really only interested in finding the series expansion when ν *is* an integer, since for non-integer ν, $Y_\nu(z)$ is nothing but a non-singular linear combination of $J_\nu(z)$ and $J_{-\nu}(z)$, each of which can be expanded straightforwardly using (1.27).

We need, therefore, to study $Y_\nu(z)$ given by (1.47) as ν approaches an integer n. We may write $\nu = n+\epsilon$, where ϵ will be sent to zero. We can assume, without loss of generality, that n is a non-negative integer. We have

$$\begin{aligned} \cos\nu\pi &= \cos(n+\epsilon)\pi \approx (-1)^n\,, \\ \sin\nu\pi &= \sin(n+\epsilon)\pi \approx (-1)^n \sin\epsilon\pi \approx (-1)^n\,\epsilon\,\pi\,. \end{aligned} \quad (1.64)$$

Therefore from (1.47) we find that

$$Y_n(z) = \frac{1}{\epsilon\,\pi}\Big(J_{n+\epsilon}(z) - (-1)^n\, J_{-n-\epsilon}(z)\Big)\,, \tag{1.65}$$

in the limit where ϵ is sent to zero. In other words, we have to pick out the $O(\epsilon)$ term in $(J_{n+\epsilon}(z) - (-1)^n\, J_{-n-\epsilon}(z))$. (We know, of course, that there is no ϵ-independent term, by virtue of the relation $J_n(z) = (-1)^n\, J_{-n}(z)$ that we derived earlier.)

Some useful *lemmata* are the following:

$$\begin{aligned}
\Big(\frac{z}{2}\Big)^{n+\epsilon} &= \Big(\frac{z}{2}\Big)^n\, e^{\epsilon\,\log(\frac{1}{2}z)} = \Big(\frac{z}{2}\Big)^n\,(1+\epsilon\,\log\frac{z}{2}+\cdots)\,,\\
\frac{1}{\Gamma(p+\epsilon+1)} &= \frac{1}{\Gamma(p+1)}\Big(1-\epsilon\,\psi(p+1)+\cdots\Big)\,,\\
\frac{1}{\Gamma(q-\epsilon+1)} &= -\frac{\sin(q-\epsilon)\pi}{\pi}\,\Gamma(-q+\epsilon) = (-1)^q\,\epsilon\,\Gamma(-q)+\cdots
\end{aligned} \tag{1.66}$$

where p is a non-negative integer, q is a negative integer, and in all cases the terms represented by $\cdots$ are of order ϵ^2 or higher, and are therefore not needed in our limiting procedure. The function $\psi(z)$ is the *digamma function*, defined by

$$\psi(z) \equiv \frac{\Gamma'(z)}{\Gamma(z)}\,. \tag{1.67}$$

One can show that for an integer argument m, it is given by

$$\psi(m) = -\gamma + \sum_{r=1}^{m-1}\frac{1}{r}\,, \tag{1.68}$$

where $\gamma = 0.5772157\ldots$ is the Euler-Mascheroni constant, defined as the limit when $m \longrightarrow \infty$ of

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{m}-\log m\,. \tag{1.69}$$

Using the *lemmata*, we find that

$$\begin{aligned}
&J_{n+\epsilon}(z) - (-1)^n\, J_{-n-\epsilon}(z)\\
&= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\Big(\frac{z}{2}\Big)^{n+2r}\,(1+\epsilon\,\log\frac{z}{2}+\cdots)(1-\epsilon\,\psi(n+r+1)+\cdots)\\
&\quad -(-1)^n\,\epsilon\sum_{r=0}^{n-1}\frac{(n-r-1)!}{r!}\Big(\frac{z}{2}\Big)^{-n+2r}+\cdots\\
&\quad -(-1)^n\sum_{r=n}^{\infty}\frac{(-1)^r}{r!}\Big(\frac{z}{2}\Big)^{-n+2r}\,(1-\epsilon\,\log\frac{z}{2}+\cdots)(1+\epsilon\,\psi(-n+r+1)+\cdots)\,,
\end{aligned} \tag{1.70}$$

where the second and third lines come from splitting the r summation for $J_{-n-\epsilon}(z)$ into the range where $r-n$ is negative, and the remainder, where $r-n \ge 0$. After making a shift of the summation variable in the third line, $r \longrightarrow r+n$, one immediately sees that, as expected, all the ϵ-independent terms cancel out, and what remains can be written as

$$\begin{aligned} J_{n+\epsilon}(z) - (-1)^n\, J_{-n-\epsilon}(z) &= \epsilon \sum_{r=0}^{\infty} \frac{(-1)^r}{r!\,(n+r)!} \Big(\frac{z}{2}\Big)^{n+2r} \Big[2\log\frac{z}{2} - \psi(n+r+1) - \psi(r+1)\Big] \\ &\quad -\epsilon \sum_{r=0}^{n-1} \frac{(-1)^r\,(n-r-1)!}{r!} \Big(\frac{z}{2}\Big)^{-n+2r} + O(\epsilon^2)\,. \end{aligned} \tag{1.71}$$

Finally, therefore, we find by substituting into (1.65) and sending ϵ to zero that $Y_n(z)$ has the series expansion

$$\begin{aligned} Y_n(z) &= \frac{1}{\pi} \sum_{r=0}^{\infty} \frac{(-1)^r}{r!\,(n+r)!} \Big(\frac{z}{2}\Big)^{n+2r} \Big[2\log\frac{1}{2}z - \psi(n+r+1) - \psi(r+1)\Big] \\ &\quad -\frac{1}{\pi} \sum_{r=0}^{n-1} \frac{(-1)^r\,(n-r-1)!}{r!} \Big(\frac{z}{2}\Big)^{-n+2r}\,. \end{aligned} \tag{1.72}$$

1.6 Asymptotic Expansions of $J_\nu(z)$ and $Y_\nu(z)$

So far, we have studied the expansions for $J_\nu(z)$ and $Y_n(z)$, expressed as power series around $z=0$. The resulting expression (1.27) for $J_\nu(z)$ is convergent for all finite z, since $J_\nu(z)$ is analytic in the finite complex plane. For $Y_n(z)$, the series (1.72) has a branch point and poles at $z=0$, as signalled by the occurrence of the logarithms and inverse powers of z, but otherwise it is analytic in the finite complex plane. These series are, in particular, useful and usable for answering all questions about the small-z behaviour of the Bessel functions.

We should also like to know how the Bessel functions behave at large values of their argument z. For example, in a scattering problem, where z might parameterise the radial coordinate that measures the distance from the scattering-centre, one would like to know how the scattered waves depend on z at large distance. We shall in fact study an example of such a scattering problem later.

Finding the large-z behaviour of a function is the kind of problem that we studied at the end of Part 1 of the course, under the heading of *Asymptotic Expansions.* In a typical example, and indeed the Bessel functions are no exception, one cannot obtain convergent power-series expansions at large z, owing to the fact that they have essential singularities at infinity. Another example of such a function is the exponential e^z. Transforming from the complex variable z to $w=1/z$, we see that in the vicinity of $z=\infty$ the exponential looks like $e^{1/w}$ with w close to zero. This has a singularity at $w=0$ that is "worse" than any power-law $1/w^n$, no matter how large n is. This is what is called an essential singularity.

We saw in Part I of the course that in such circumstances, when there is an essential singularity, one may still be able to construct a useful series expansion that approximates a function $F(z)$ at large z. However, it will no longer be a convergent series; instead, it is an asymptotic expansion. We refer the reader to Part 1 of the lecture notes for details. A brief summary of the idea is as follows.

An ordinary convergent power series approximates $F(z)$ to better and better accuracy, at fixed z, as more and more terms are included in the sum. Eventually, the agreement becomes perfect as the number of terms is taken to infinity. By contrast, an asymptotic expansion is actually *divergent*; if one sums up all the terms at a fixed value of z, the sum diverges. However, instead what we do is to look at a *fixed* number of terms in the series; the first N terms, let us say. Then, as z is made larger and larger, the N-term series gives a better and better approximation to $F(z)$, becming perfect in the limit when z becomes infinite. For any given finite value of z there is a limit to how good an approximation we can get; beyond a certain point, adding in more terms in the series makes things *worse*, not better. Nonetheless, the asymptotic expansion is a very useful approximation that gives all the required information about the large-z asymptotic behaviour of the function.

We have obtained the integral representation (1.29) for the Bessel function $J_\nu(z)$. A very useful technique for constructing the asymptotic expansion of a function defined by an integral representation is by means of the *Method of Steepest Descent*. This was discussed in detail in Part 1 of the course, and we shall not present all the details again here. The general idea, expressed in the notation of variables that we are using in this section, is that one has an integral representation of the form

$$F(z) = \int_C g(t)\, e^{z\, f(t)}\, dt\,, \tag{1.73}$$

where $f(t)$ is such that $\mathrm{Re}(z\, f(t))$ goes to $-\infty$ at both ends of the range of integration along the contour C. The idea is that as z is taken very large, the integrand becomes dominated by the point (or points) in the complex t-plane where $f(t)$ is stationary, $f'(t) = 0$. The function $g(t)$ is assumed to have such a form that it varies only slowly in the vicinity of the point, which is at, let us say, $t = t_0$. Then, what one does is to deform the contour so that it passes through the stationary point at $t = t_0$, and swing it around so that it follows the path of steepest descent as one moves away from $t = t_0$ in either direction along the contour. To a good approximation, since one has

$$f(t) = f(t_0) + \frac{1}{2}(t-t_0)^2\, f''(t_0) + \cdots\,, \tag{1.74}$$

the integral is now just dominated by a Gaussian integrand of the form

$$e^{-\frac{1}{2}u^2}\,, \tag{1.75}$$

where u is the renamed integration variable after having deformed the contour so that it follows the path of steepest descent. All other factors in the integrand can just be taken outside the integration, with their original argument t replaced by the value t_0 at the stationary point. If there is more than one stationary point, we just repeat the procedure at each, and add up the contributions.

Without further ado, let us now use the method of steepest descent to calculate the asymptotic behaviour of the Bessel function $J_\nu(z)$. We have, from (1.29),

$$J_\nu(z) = \frac{1}{2\pi\,\mathrm{i}} \int_C t^{-\nu-1}\, e^{\frac{1}{2}z(t-t^{-1})}\, dt\,, \tag{1.76}$$

and so comparing with (1.73) we have

$$f(t) = \tfrac{1}{2}(t - t^{-1})\,. \tag{1.77}$$

This has stationary points at $f'(t) = \frac{1}{2}(1 + t^{-2}) = 0$, in other words at $t = \pm\mathrm{i}$. Note that we have $f(\mathrm{i}) = \mathrm{i}$, and $f(-\mathrm{i}) = -\mathrm{i}$. The first thing we do now is to deform the contour C so that it passes through the points $t = \pm\mathrm{i}$.

Consider the contribution from $t = +\mathrm{i}$ first. Expanding $f(t)$ in a Taylor series around $t = +\mathrm{i}$, we have

$$f(t) = \mathrm{i} - \frac{\mathrm{i}}{2}\,(t-\mathrm{i})^2 + \cdots\,. \tag{1.78}$$

(The first term is just $f(\mathrm{i})$, and of course there is no linear term since $f'(\mathrm{i}) = 0$.) To deform the contour so that it follows the path of steepest descent, it is useful to introduce a new integration coordinate u in place of t, which will be real along the steepest-descent path. We do this by defining it to be such that

$$-\frac{\mathrm{i}}{2}\,(t-\mathrm{i})^2 = -\frac{u^2}{2z}\,. \tag{1.79}$$

(Take z to be real and positive for now.) Thus we have

$$(t-\mathrm{i})^2 = \frac{u^2}{z}\, e^{-\frac{1}{2}\mathrm{i}\pi}\,. \tag{1.80}$$

Taking the square root, we get

$$t - \mathrm{i} = -\frac{u}{\sqrt{z}}\, e^{-\frac{1}{4}\mathrm{i}\pi}\,. \tag{1.81}$$

We have chosen the square root with the minus sign here because we want the contour to run in the natural anticlockwise direction as u runs from negative to positive values. Thus for negative u, the contour approaches $t = \mathrm{i}$ from the south-east, and as u goes positive it leaves $t = \mathrm{i}$ in a north-westerly direction (the slope of the line being precisely -1). Note that to change integration variable from t to u, we shall have

$$dt = \frac{dt}{du}\, du = -\frac{1}{\sqrt{z}}\, e^{-\frac{1}{4}\mathrm{i}\pi}\,. \tag{1.82}$$

Let us call I_+ the contribution to $J_\nu(z)$ from this stationary point at $t = +\mathrm{i}$. Thus from (1.76) we shall have

$$I_+ \sim -\frac{1}{2\pi\,\mathrm{i}} \left(e^{\frac{1}{2}\mathrm{i}\pi}\right)^{-\nu-1} \frac{1}{\sqrt{z}}\, e^{-\frac{1}{4}\mathrm{i}\pi}\, e^{\mathrm{i}\, z} \int e^{-\frac{1}{2}u^2}\, du\,. \tag{1.83}$$

The factors sitting out at the front come from taking $t^{-\nu-1}$ outside the integral, setting $t = \mathrm{i} = e^{\frac{1}{2}\mathrm{i}\pi}$ as we do so; making the transformation from dt to du using (1.82); and taking out the factor $e^{z\, f(\mathrm{i})} = e^{\mathrm{i}\, z}$ that comes from

$$e^{z\, f(t)} \approx e^{z\, f(t_0) - \frac{1}{2}u^2}\,. \tag{1.84}$$

The integration over u can be excellently approximated by allowing the limits to be $-\infty$ and $+\infty$, since we are assuming that z is large. (See (1.79); when z is large, u can be large while t is still rather close to $t = \mathrm{i}$.) Thus the integral is just a Gaussian, which gives a factor of $\sqrt{2\pi}$. Putting it all together, we therefore have

$$I_+ \sim \frac{1}{\sqrt{2\pi\, z}}\, e^{\mathrm{i}(z - \frac{1}{2}\nu\,\pi - \frac{1}{4}\pi)}\,. \tag{1.85}$$

Now we consider the contribution I_- to $J_\nu(z)$ from the other stationary point, at $t = -\mathrm{i}$. Expanding around this point we have

$$f(t) = -\mathrm{i} + \frac{\mathrm{i}}{2}\,(t + \mathrm{i})^2 + \cdots\,, \tag{1.86}$$

and so we choose our real integration variable u that parameterises the path of steepest descent to be such that

$$(t + \mathrm{i})^2 = \frac{u^2}{z}\, e^{\frac{1}{2}\mathrm{i}\pi}\,. \tag{1.87}$$

This time, the square root will be

$$t + \mathrm{i} = \frac{u}{\sqrt{z}}\, e^{\frac{1}{4}\mathrm{i}\pi}\,, \tag{1.88}$$

so that the contour comes in from the south-west, and head onwards to the north-east, as it should. The slope here is precisely $+1$. Thus we find by a similar calculation to the above that

$$I_- \sim \frac{1}{\sqrt{2\pi\, z}}\, e^{\mathrm{i}(-z + \frac{1}{2}\nu\,\pi + \frac{1}{4}\pi)}\,. \tag{1.89}$$

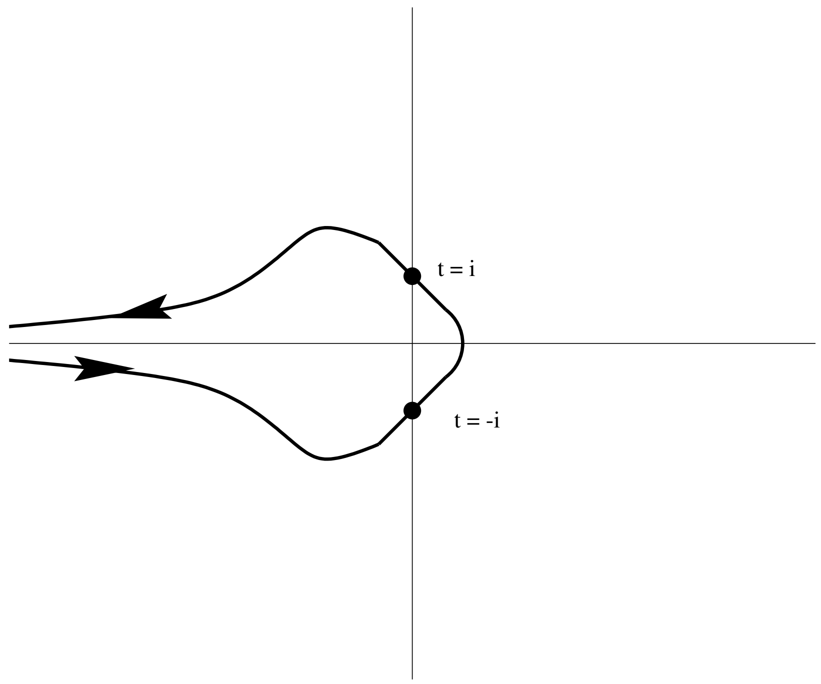

Figure 9: The deformed Bessel contour that follows the paths of steepest descent at $t = \pm \mathrm{i}$.

The deformed contour that we have used in the steepest-descent integrals is depicted in Figure 9. Notice that the contour is running at precisely the 45-degree angles implied by (1.81) and (1.88) as it passes through the points $t = +\mathrm{i}$ and $t = -\mathrm{i}$ respectively.

Finally, we put the two results together, $J_\nu(z) = I_+ + I_-$, giving

$$J_\nu(z) \sim \sqrt{\frac{2}{\pi\, z}}\, \cos(z - \tfrac{1}{2}\nu\,\pi - \tfrac{1}{4}\pi)\,. \tag{1.90}$$

This is our asymptotic formula for the large-z behaviour of the Bessel function $J_\nu(z)$.

Notice that this result fits very nicely with what we saw in the various graphs of Bessel functions, in Figures 1 to 6. One can see from the plots that the intervals between successive zeros seem to be settling down to equal steps, precisely as is implied by the asymptotically cosine form appearing in (1.90). Furthermore, one can see from the graphs that the amplitude of the oscillation is falling off in a rather mild way as z gets larger. This also is understandable from the asymptotic expression (1.90), which has a $1/\sqrt{z}$ prefactor to the cosine function.

The asymptotic formula that we have obtained here is the leading term in the full asymptotic expansion. As was discussed in Part 1 of the course, there is a systematic procedure for constructing the expansion to any desired number of terms. Essentially, what one does is to replace the truncated Taylor series for $f(t)$ in (1.74) by the full series, or at least as many terms as one wishes to work with. The redefined integration coordinate u is then given by the corresponding full expression, rather than the truncated one (1.80). Other than that, and the associated complications that now arise from having to invert so as to express dt/du in terms of u, things proceed pretty much as before. The result, which we shall derive later, can be shown to be

$$J_\nu(z) \sim \sqrt{\frac{2}{\pi\, z}}\, \Big[\cos(z - \tfrac{1}{2}\nu\,\pi - \tfrac{1}{4}\pi) \sum_{r=0}^\infty a_r\, z^{-2r} + \sin(z - \tfrac{1}{2}\nu\,\pi - \tfrac{1}{4}\pi) \sum_{r=0}^\infty b_r\, z^{-2r-1}\Big]\,, \tag{1.91}$$

where $a_0 = 1$ and

$$\begin{aligned} a_r &= \frac{(-1)^r}{(2r)!\, 2^{6r}}\, \Big((4\nu^2 - 1^2)(4\nu^2 - 3^2)\cdots(4\nu^2 - (4r-1)^2)\Big)\,, \\ b_r &= \frac{(-1)^{r+1}}{(2r+1)!\, 2^{6r+3}}\, \Big((4\nu^2 - 1^2)(4\nu^2 - 3^2)\cdots(4\nu^2 - (4r+1)^2)\Big)\,. \end{aligned} \tag{1.92}$$

Our result above corresponds to the leading-order term with the coefficient $a_0 = 1$ in this asymptotic expansion. In practice, (1.90) is commonly quite sufficient.

Having struggled to obtain the asymptotic form of $J_\nu(z)$, it is, fortunately, now a relative triviality to get the analogous formula for $Y_\nu(z)$. We need only refer back to the original definition of $Y_\nu(z)$, given in (1.47), and plug in the result (1.90). After an elementary use of the identities for the product of two trigonometric functions, we get the result:

$$Y_\nu(z) \sim \sqrt{\frac{2}{\pi\, z}}\, \sin(z - \tfrac{1}{2}\nu\,\pi - \tfrac{1}{4}\pi)\,. \tag{1.93}$$

1.7 The Hankel Functions $H_\nu^{(1)}(z)$ and $H_\nu^{(2)}(z)$

We have seen that asymptotically, $J_\nu(z)$ and $Y_\nu(z)$ become very similar to certain cosine and sine functions. Not surprisingly, perhaps, it turns out that it is often convenient to introduce complex combinations of $J_\nu(z)$ and $Y_\nu(z)$, which have the property of approaching complex exponentials of the form $e^{\pm \mathrm{i}\, z}$ asymptotically. In particular, these are very convenient combinations to use when considering solutions of a wave equation. Accordingly, one defines the so-called Hankel functions of the first and second kind, denoted by $H_\nu^{(1)}(z)$ and $H_\nu^{(2)}(z)$ respectively, by

$$H_\nu^{(1)}(z) = J_\nu(z) + \mathrm{i}\, Y_\nu(z)\,, \qquad H_\nu^{(2)}(z) = J_\nu(z) - \mathrm{i}\, Y_\nu(z)\,. \tag{1.94}$$

Clearly, from (1.90) and (1.93), when z is large they have the asymptotic behaviour

$$H_\nu^{(1)}(z) \sim \sqrt{\frac{2}{\pi\, z}}\, e^{\mathrm{i}(z - \frac{1}{2}\nu\,\pi - \frac{1}{4}\pi)}\,, \qquad H_\nu^{(2)}(z) \sim \sqrt{\frac{2}{\pi\, z}}\, e^{-\mathrm{i}(z - \frac{1}{2}\nu\,\pi - \frac{1}{4}\pi)}\,. \tag{1.95}$$

The Hankel functions can be obtained elegantly from the contour integral representation (1.29), by making suitable changes to the choice of contour. Specifically, we can show that they are given by

$$\begin{aligned} H_\nu^{(1)}(z) &= \frac{1}{\pi\,\mathrm{i}} \int_{C_1} t^{-\nu-1}\, e^{\frac{1}{2}z\,(t-t^{-1})}\, dt\,, \\ H_\nu^{(2)}(z) &= \frac{1}{\pi\,\mathrm{i}} \int_{C_2} t^{-\nu-1}\, e^{\frac{1}{2}z\,(t-t^{-1})}\, dt\,, \end{aligned} \tag{1.96}$$

where the contours C_1 and C_2 are chosen as follows. The contour C_2 starts out like the original contour in Figure 7, just below the real axis out west at $t = -\infty$. It heads in and swings half way around the origin, and then dives directly in to the origin along the positive real axis. The contour C_1 is the reflection of this across the real axis; it comes out from the origin, swings up and around, and heads off to the west, just above the real axis, eventually reaching $t = -\infty$. The two contours are depicted in Figure 10 below.

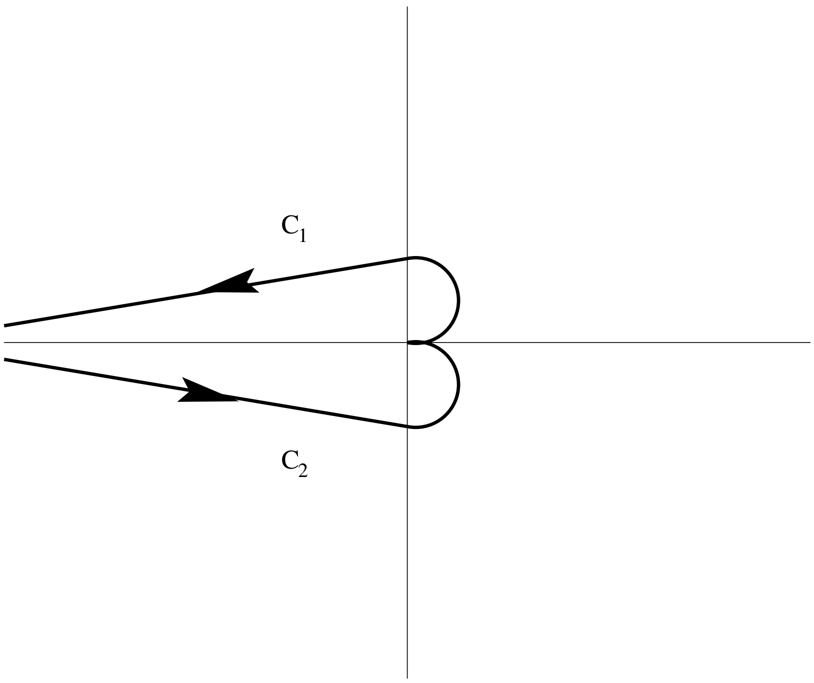

Figure 10: The contours C_1 and C_2 for the Hankel functions $H_\nu^{(1)}(z)$ and $H_\nu^{(2)}(z)$.

The reason why such contours are allowed is that as t heads in to the origin along the real axis, the factor $e^{-\frac{1}{2}z\,t^{-1}}$ in the integrand goes to zero (when the real part of z is positive.) Thus we again have the situation that when one substitutes into the Bessel equation, the "boundary term" arising from integration by parts vanishes at both ends of the contour, just like it did in our earlier discussion of the integral representation for $J_\nu(z)$. Thus with either of the contours C_1 or C_2, the integral defines a function that satisfies Bessel's equation.

Let us now verify that indeed the expressions for $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$ in (1.96) are in agreement with the definitions (1.94). It is clear that the sum of the contours C_1 and C_2 is equivalent, up to allowed deformations, to the contour C used in the integral representation (1.29) for $J_\nu(z)$. Therefore we can immediately verify from (1.96) and (1.29) that indeed we shall have

$$J_\nu(z) = \tfrac{1}{2}(H^{(1)}_\nu(z) + H^{(2)}_\nu(z))\,. \tag{1.97}$$

It remains to show from (1.96) that

$$Y_\nu(z) = \frac{1}{2\mathrm{i}}(H^{(1)}_\nu(z) - H^{(2)}_\nu(z))\,, \tag{1.98}$$

which is what is required by (1.94). To do this, we first make the change of integration variable $t = e^{\mathrm{i}\,\pi}/s$ in the expression for $H^{(1)}_\nu(z)$ in (1.96). Note that since the imaginary part of t is positive on the contour C_1, it follows that this maps into a contour for s where again its imaginary part is positive.[2] In fact for this reason, the contour for the transformed integral using s can again be taken to be just C_1. The starting point $t = 0$ becomes $s = -\infty$, while the endpoint $t = -\infty$ becomes $s = 0$. This reversal of the direction is compensated by the fact that $dt/t = -ds/s$. The fact that the contour has been mapped back onto itself is crucial, because it means that we can again interpret the integral as giving a Hankel function of the first kind; this time, with order $-\nu$. Thus we find that

$$\begin{aligned} H^{(1)}_\nu(z) &= \frac{1}{\pi\,\mathrm{i}}\,e^{-\mathrm{i}\,\nu\,\pi}\int_{C_1} s^{\nu-1}\,e^{\frac{1}{2}z\,(-s^{-1}+s)}\,ds\,,\\ &= e^{-\mathrm{i}\,\nu\,\pi}\,H^{(1)}_{-\nu}(z)\,. \end{aligned} \tag{1.99}$$

By a similar argument, in which we change the integration variable in the expression for $H^{(2)}_\nu(z)$ in (1.96) by $t = e^{-\mathrm{i}\,\pi}/s$, we deduce also that

$$H^{(2)}_\nu(z) = e^{\mathrm{i}\,\nu\,\pi}\,H^{(2)}_{-\nu}(z)\,. \tag{1.100}$$

(The change of variable here ensures that t, whose imaginary part is negative on the contour C_2, maps into s that also has negative imaginary part. Again, this means that s can be integrated along the same contour as was t.)

Having established these two results we can now not only express $J_\nu(z)$ in terms of $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$ using (1.97), but also $J_{-\nu}(z)$ in terms of $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$. These

[2]Consider a point on the contour C_1 in the complex t plane. Since t lies in the upper half plane, it has the form $t = r\,e^{\mathrm{i}\,\theta}$, where $0 < \theta < \pi$. Therefore $s = e^{\mathrm{i}\,\pi}/t = r^{-1}\,e^{\mathrm{i}(\pi-\theta)}$, and so s lies in the upper half plane too.

can then be plugged into the original definition of $Y_\nu(z)$ in terms of $J_\nu(z)$ and $J_{-\nu}(z)$ as given in (1.47). This gives

$$Y_\nu(z) = \frac{1}{2\sin\nu\pi}\Big(\cos\nu\pi\,(H^1_\nu(z) + H^{(2)}_\nu(z)) - e^{\mathrm{i}\,\nu\,\pi}\,H^{(1)}_\nu(z) - e^{-\mathrm{i}\,\nu\,\pi}\,H^{(2)}_\nu(z)\Big)\,. \tag{1.101}$$

Collecting terms, we see that this produces precisely the expression (1.98). This completes the demonstration that the original definitions (1.94) of the Hankel functions agree precisely with the integral representations given in (1.96).

Notice that we can easily repeat the previous derivation of the asymptotic behaviour of the $J_\nu(z)$ Bessel function, for the case of the Hankel functions $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$. In fact, we have already obtained all the necessary results in section 1.6. When we applied the method of steepest descent there, we found that the contour C passed through two stationary points, at $t = +\mathrm{i}$ and $t = -\mathrm{i}$, and so we obtained two contributions which, when added, gave the asymptotic form of $J_\nu(z)$. For the Hankel functions we have the same integrand (multiplied by a factor of 2), but now with the contour C_1 or C_2. In fact in the method of steepest descent the contour C_1 will be deformed to one that passes just through the single stationary point at $t = +\mathrm{i}$. Likewise, C_2 will be deformed to a contour passing just through the $t = -\mathrm{i}$ stationary point. Thus the asymptotic forms of $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$ will be precisely equal to $2I_+$ and $2I_-$ respectively, where $I_\pm$ are the contributions coming from the steepest-descent integrations around $t = \pm\mathrm{i}$ respectively in section 1.6. Sure enough, we see that the asymptotic forms of $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$ given in (1.95) are precisely in agreement with $2I_+$ and $2I_-$ respectively, where $I_\pm$ were obtained in (1.85) and (1.89).

1.8 Orthogonality of Bessel functions

If the Bessel equation (1.1) is divided by z, it assumes the self-adjoint form

$$(z\,y')' + \Big(z - \frac{\nu^2}{z}\Big)\,y = 0\,. \tag{1.102}$$

From the general discussion of Sturm-Liouville problems (see Part 1 of the lecture course), this means that, with respect to suitable boundary conditions, the Bessel functions will satisfy orthogonality relations. These will be useful, for example, when we analyse problems that involve solving Laplace's equation or the wave equation in situations with cylindrical symmetry, where Bessel functions arise in the solutions.

Recall, for example, that Laplace's equation in cylindrical polar coordinates (ρ, ϕ, z) is

$$\frac{1}{\rho}\,\frac{\partial}{\partial\rho}\Big(\rho\,\frac{\partial\psi}{\partial\rho}\Big) + \frac{1}{\rho^2}\,\frac{\partial^2\psi}{\partial\phi^2} + \frac{\partial^2\psi}{\partial z^2} = 0\,. \tag{1.103}$$

Separating variables by writing $\psi = R(\rho)\,\Phi(\phi)\,Z(z)$, we get

$$\frac{d^2 Z}{dz^2} - k^2\, Z = 0\,, \qquad \frac{d^2\Phi}{d\phi^2} + \nu^2\,\Phi = 0\,, \tag{1.104}$$

$$\frac{d^2 R}{d\rho^2} + \frac{1}{\rho}\,\frac{dR}{d\rho} + \Big(k^2 - \frac{\nu^2}{\rho^2}\Big)\, R = 0\,, \tag{1.105}$$

where k^2 and ν^2 are separation constants. Rescaling the radial coordinate by defining $x = k\,\rho$, and renaming R as y, the last equation takes the standard Bessel form

$$x^2\,\frac{d^2 y}{dx^2} + x\,\frac{dy}{dx} + (x^2 - \nu^2)\,y = 0\,. \tag{1.106}$$

Thus the radial functions $R(\rho)$ are of the form

$$R(\rho) = J_\nu(k\,\rho) \qquad \text{or} \qquad Y_\nu(k\,\rho)\,. \tag{1.107}$$

In a typical electrostatics problem, the potential ψ will be required to be regular on the axis at $\rho = 0$. For now, consider an example where in addition $\psi = 0$ on a cylindrical surface at some radius $\rho = a$. This implies that the general solution of Laplace's equation will be expressed in terms of the $J_\nu(z)$ and $Y_\nu(z)$ Bessel functions.[3] The requirement of regularity at $\rho = 0$ implies that the $Y_\nu(z)$ Bessel functions are excluded (as indeed, if ν is not an integer, are the $J_\nu(z)$ Bessel functions for $\nu < 0$). So for now, let us just consider $J_\nu(z)$ as the expansion functions.

We have seen from the plots of the Bessel functions, and from their asymptotic behaviour, that $J_\nu(z)$ has a discrete infinite set of zeros, at points on the real z axis that asymptotically approach an equal spacing. Let us say that the m'th zero of $J_\nu(z)$ occurs at

$$z = \alpha_{\nu m}\,, \qquad \text{so} \qquad J_\nu(\alpha_{\nu m}) = 0\,. \tag{1.108}$$

So $m = 1$ is the location of the first zero, $m = 2$ is the location of the second, and so on, as z increases from 0. They occur at definite values of $\alpha_{\nu m}$, though it is not easy to give explicit expressions for $\alpha_{\nu m}$.

If we are wanting to impose the requirement that the potential ψ vanishes on a cylindrical surface at $\rho = a$, then we shall want to expand ψ in terms of Bessel functions $J_\nu(k\,\rho)$ for which $k\,a$ is equal to one of the quantities $\alpha_{\nu m}$ defined above. In other words, this determines

[3] If the boundary conditions were different, we could instead have a situation where the separation constant k above were imaginary, in which case we would be dealing with Bessel functions of the form $J_\nu(\mathrm{i}\,z)$, *etc.*These are given different names (just like hyperbolic as opposed to trigonometric functions), and we shall discuss them later. Like the hyperbolic functions, they have real-exponential rather than oscillatory behaviour.

the set of values for the separation constant k that can arise in this boundary-value problem. Thus we shall consider the Bessel function expressions

$$J_\nu(\alpha_{\nu m}\,\rho/a)\,; \tag{1.109}$$

these will form our expansion fuctions for the radial function $R(\rho)$. Substituting such an $R(\rho)$ into (1.105), and multiplying by ρ, we get

$$\rho\,\frac{d^2}{d\rho^2}\,J_\nu(\alpha_{\nu m}\,\rho/a) + \frac{d}{d\rho}\,J_\nu(\alpha_{\nu m}\,\rho/a) + \Big(\frac{\alpha_{\nu m}^2\,\rho}{a^2} - \frac{\nu^2}{\rho}\Big)\,J_\nu(\alpha_{\nu m}\,\rho/a) = 0\,. \tag{1.110}$$

Now we follow the usual story for proving orthogonality, of muliplying (1.110) by $J_\nu(\alpha_{\nu n}\,\rho/a)$, and on the other hand writing the equivalent equation to (1.110) but with m replaced by n, multiplying *it* by $J_\nu(\alpha_{\nu m}\,\rho/a)$, and subtracting the latter from the former. This gives

$$\begin{aligned} J_\nu(\alpha_{\nu n}\,\rho/a)\,\frac{d}{d\rho}\Big(\rho\,\frac{d}{d\rho}J_\nu(\alpha_{\nu m}\,\rho/a)\Big) - J_\nu(\alpha_{\nu m}\,\rho/a)\,\frac{d}{d\rho}\Big(\rho\,\frac{d}{d\rho}J_\nu(\alpha_{\nu n}\,\rho/a)\Big) \\ = \frac{\alpha_{\nu n}^2 - \alpha_{\nu m}^2}{a^2}\,\rho\,J_\nu(\alpha_{\nu m}\,\rho/a)\,J_\nu(\alpha_{\nu n}\,\rho/a)\,. \end{aligned} \tag{1.111}$$

Next, we integrate this from $\rho = 0$ to $\rho = a$. On the left-hand side we integrate by parts, finding that there is now a cancellation of the resulting two integrands, leaving only the "boundary terms." Thus we have

$$\begin{aligned} \Big|\rho\,J_\nu(\alpha_{\nu n}\,\rho/a)\,\frac{d}{d\rho}J_\nu(\alpha_{\nu m}\,\rho/a)\Big|_0^a - \Big|\rho\,J_\nu(\alpha_{\nu m}\,\rho/a)\,\frac{d}{d\rho}J_\nu(\alpha_{\nu n}\,\rho/a)\Big|_0^a \\ = \frac{\alpha_{\nu n}^2 - \alpha_{\nu m}^2}{a^2}\,\int_0^a J_\nu(\alpha_{\nu m}\,\rho/a)\,J_\nu(\alpha_{\nu n}\,\rho/a)\,\rho\,d\rho\,. \end{aligned} \tag{1.112}$$

Recalling from (1.27) that near $\rho = 0$, $J_\nu(\alpha_{\nu n}\,\rho/a)$ is proportional to ρ^ν, we see that with our assumption that $\nu \ge 0$ the lower limits on the left-hand side of (1.112) will give zero. Furthermore, the upper limits will also give zero, since by construction $J_\nu(\alpha_{\nu m}) = 0$. Thus we arrive at the conclusion that for $m \ne n$ (which implies $\alpha_{\nu m} \ne \alpha_{\nu n}$), we shall have

$$\int_0^a J_\nu(\alpha_{\nu m}\,\rho/a)\,J_\nu(\alpha_{\nu n}\,\rho/a)\,\rho\,d\rho = 0\,. \tag{1.113}$$

Having established orthogonality when $m \ne n$, it remains to determine the normalisation of the integral that we get when instead we take $m = n$. To do this, let $x = \alpha_{\nu m}\,\rho/a$, so that

$$\int_0^a J_\nu(\alpha_{\nu n}\,\rho/a)^2\,\rho\,d\rho = \frac{a^2}{\alpha_{\nu n}^2}\,\int_0^{\alpha_{\nu n}} J_\nu(x)^2\,x\,dx\,. \tag{1.114}$$

To evaluate the integral on the right-hand side, we integrate by parts, by writing $J_\nu(x)^2\, x = \frac{1}{2} d/dx(x^2\, J_\nu(x)^2) - \frac{1}{2} x^2\, d/dx(J_\nu(x)^2)$, so that

$$\int_{x_1}^{x_2} J_\nu(x)^2\, x\, dx = \Big[\tfrac{1}{2} x^2\, J_\nu^2\Big]_{x_1}^{x_2} - \int_{x_1}^{x_2} x^2\, J_\nu\, J_\nu'\, dx\,. \tag{1.115}$$

We have also allowed rather more general upper and lower limits of integration x_1 and x_2 here, since then the resulting formula will be of wider applicability. Now use the Bessel equation (1.1) to write $x^2\, J_\nu$ as $\nu^2\, J_\nu - x\, J_\nu' - x^2\, J_\nu''$, so that we get

$$\begin{aligned}
\int_{x_1}^{x_2} J_\nu(x)^2\, x\, dx &= \Big[\tfrac{1}{2} x^2\, J_\nu^2\Big]_{x_1}^{x_2} - \int_{x_1}^{x^2} \Big(\nu^2\, J_\nu\, J_\nu' - x\, J_\nu'^{\,2} - x^2\, J_\nu'\, J_\nu''\Big)\, dx\,, \\
&= \Big[\tfrac{1}{2} x^2\, J_\nu^2\Big]_{x_1}^{x_2} - \int_{x_1}^{x_2} \Big(\tfrac{1}{2}\nu^2\, (J_\nu^2)' - \tfrac{1}{2}(x^2\, J_\nu'^{\,2})'\Big)\, dx \\
&= \tfrac{1}{2}\Big[x^2\, J_\nu^2 - \nu^2\, J_\nu^2 + x^2\, J_\nu'^{\,2}\Big]_{x_1}^{x_2}\,.
\end{aligned} \tag{1.116}$$

In our specific case we have integration limits $x_1 = 0$, $x_2 = \alpha_{\nu n}$. Therefore the first two terms in the final line vanish at both our endpoints (recall that $\alpha_{\nu n}$ are precisely the values of argument for which $J_\nu(\alpha_{\nu n}) = 0$). For the final term, we use (1.33), expanded out to give

$$J_\nu'(z) = \frac{\nu}{z}\, J_\nu(z) - J_{\nu+1}(z)\,. \tag{1.117}$$

Thus, with our assumption that $\nu \ge 0$ we see that $x^2\, J_\nu'^{\,2}$ will vanish at $x = 0$. Also, from (1.117) we see that $J_\nu'(\alpha_{\nu n}) = -J_{\nu+1}(\alpha_{\nu n})$, and so

$$\int_0^{\alpha_{\nu n}} J_\nu(x)^2\, x\, dx = \tfrac{1}{2}\alpha_{\nu n}^2\, J_{\nu+1}(\alpha_{\nu n})^2\,, \tag{1.118}$$

implying finally that

$$\int_0^a J_\nu(\alpha_{\nu m}\, \rho/a)\, J_\nu(\alpha_{\nu n}\, \rho/a)\, \rho\, d\rho = \tfrac{1}{2} a^2\, J_{\nu+1}(\alpha_{\nu n})^2\, \delta_{mn}\,. \tag{1.119}$$

With this orthogonality relation, it is now a simple matter to determine the coefficients in an expansion for solutions of Laplace's equation, expressed in terms of the J_ν Bessel functions, so as to match a given boundary condition. The essential point is that, just like a Fourier series, a suitable function can be expanded as a Fourier-Bessel series, i.e. a sum over a complete set of Bessel functions. Specifically, in the present case we can expand any well-behaved function $f(\rho)$ that is regular at $\rho = 0$ and that vanishes at $\rho = a$ as a sum of the form

$$f(\rho) = \sum_{n=1}^{\infty} c_n\, J_\nu(\alpha_{\nu n}\, \rho/a)\,. \tag{1.120}$$

Multiplying by $J_\nu(\alpha_{\nu m}\, \rho/a)\, \rho$ and integrating, the orthogonality relation (1.119) gives us

$$\int_0^a f(\rho)\, J_\nu(\alpha_{\nu m}\, \rho/a)\, \rho\, d\rho = \tfrac{1}{2} a^2\, J_{\nu+1}(\alpha_{\nu m})^2\, c_m\,, \tag{1.121}$$

thus determining the expansion coefficients c_m.

Consider the following example. A conducting cylinder of height h and radius a is held at zero potential. A flat conductor closes off the cylinder at $z = 0$, and is also at zero potential. The top face, at $z = h$, is held at some specified potential

$$\psi(\rho, \phi, h) = \Psi(\rho, \phi)\,. \tag{1.122}$$

The problem is to determine the potential everywhere inside the cavity.

From (1.104) we see that the z dependence and ϕ dependence of the separation functions $Z(z)$ and $\Phi(\phi)$ will be

$$\begin{aligned} Z(z) &\sim \sinh kz\,, \qquad \cosh kz\,, \\ \Phi(\phi) &\sim \cos\nu\phi\,, \qquad \sin\nu\phi\,. \end{aligned} \tag{1.123}$$

The vanishing of the potential on the plate at $z = 0$ means that for $Z(z)$, we shall have only the $\sinh kz$ solution. The periodicity in ϕ means that ν must be an integer.

Thus the general solution of Laplace's equation for this problem will be

$$\psi(\rho, \phi, z) = \sum_{m=0}^\infty \sum_{n=1}^\infty J_m(\alpha_{mn}\,\rho/a)\,(a_{mn}\,\sin m\phi + b_{mn}\,\sin m\phi)\,\sinh(\alpha_{mn}\,z/a)\,. \tag{1.124}$$

The expansion coefficients a_{mn} and b_{mn} are determined by matching this solution to the specified boundary condition (1.122) at $z = h$. Thus we have

$$\Psi(\rho, \phi) = \sum_{m=0}^\infty \sum_{n=1}^\infty J_m(\alpha_{mn}\,\rho/a)\,(a_{mn}\,\sin m\phi + b_{mn}\,\sin m\phi)\,\sinh(\alpha_{mn}\,h/a)\,. \tag{1.125}$$

The orthogonality relation (1.119) for the Bessel functions, together with the standard orthogonality for the trigonometric functions, means that all we need to do is to multiply (1.125) by $J_p(\alpha_{pq}\,\rho/a)\,\sin p\phi$ or $J_p(\alpha_{pq}\,\rho/a)\,\sin p\phi$ and integrate over ρ and ϕ in order to read off the integrals that determine the individual coefficients a_{pq} and b_{pq}. It is easy to see that the result is

$$\begin{aligned} a_{pq} &= \frac{2}{\pi\, a^2\,\sinh(\alpha_{pq} h/a)\, J_{p+1}(\alpha_{pq})^2} \int_0^{2\pi} d\phi \int_0^a \rho\, d\rho\, \Psi(\rho, \phi)\, J_p(\alpha_{pq}\,\rho/a)\, \sin p\phi\,, \\ b_{pq} &= \frac{2}{\pi\, a^2\,\sinh(\alpha_{pq} h/a)\, J_{p+1}(\alpha_{pq})^2} \int_0^{2\pi} d\phi \int_0^a \rho\, d\rho \Psi(\rho, \phi)\, J_p(\alpha_{pq}\,\rho/a)\, \cos p\phi\,. \end{aligned} \tag{1.126}$$

In this section, we have seen how to make an expansion of solutions of Laplace's equation or the wave equation in terms of the Bessel functions J_ν, appropriate to a system with cylindrical symmetry. Furthermore, we made the assumption that the field we were solving for (for example, the electrostatic potential) was required to be non-singular on the axis of symmetry, and vanishing at radius $\rho = a$. Another example where such boundary conditions would be appropriate is a streched membrane forming a circular drum, for which the oscillations would vanish on the rim of the drum, and, of course, they would be non-singular in the middle of the membrane.

In different circumstances one might want to consider a situation with a different boundary condition at $\rho = a$. For example, in an electrostatics problem one might require that the electric field, rather than the potential, vanish at $\rho = a$. In this case one would instead want to impose that the derivative of the potential vanish at $\rho = a$. This example could be handled by a very similar method to the one we used, and only some of the fine details would change. Essentially, one would now be changing the boundary conditions in the Sturm-Liouville problem (see the lecture notes for Part 1 of the course). Again we would be working with orthogonal sets of Bessel eigenfunctions but now in (1.112) the boundary terms that arise from integration by parts when proving orthogonality would vanish for slightly different reasons. For example, if we require $\partial\psi/\partial\rho = 0$ at $\rho = a$, then we would change our choice of the constants $\alpha_{\nu\mu}$ so that instead of being defined as the zeros of $J_\nu(z)$, they would instead be defined as the zeros of $J'_\nu(z)$. With appropriate such changes, the discussion would then go through in a very similar vein.

Another modification that might arise in a slightly different kind of problem is that we might need also to make use of the "second solution" of the Bessel equation. The general series expansion after separating variables in Laplace's equation or the wave equation would involve both the J_ν and the $J_{-\nu}$ (or Y_ν, if ν is an integer) Bessel functions. In other words, Bessel functions that are singular at $\rho = 0$ might be needed too. This could happen either because one for some reason needed to allow the field ψ to be singular there, or else because $\rho = 0$ might not be within the region under consideration. An example would be if we were solving an electrostatics problem in the region between two concentric cylinders of radii a and b. Now, we would in general need the second-solution Bessel functions as well. Again, it is not too much of an extension of the methods developed already in this section to cope with such a circumstance. One would need to establish appropriate orthogonality properties for the extended set of Bessel functions, and to establish normalisation results analogous to (1.116).

Going through the details of such modifications and generalisations would really be "more of the same." There are more interesting things to pursue, so let's move on.

1.9 Modified Bessel Functions of the First and Second Kind

A familiar feature of the equation for simple harmonic motion, $y''(z) + \omega^2\, y(z) = 0$ is that its oscillatory solutions $\sin \omega z$ and $\cos \omega z$ become instead the non-oscillatory hyperbolic functions $\sinh \omega z$ and $\cosh \omega z$ if the sign of the ω^2 term is reversed, to give $y''(z) - \omega^2\, y(z) = 0$. Of course another way of achieving this sign reversal is by sending $z \longrightarrow \mathrm{i}\, z$ in the original simple harmonic equation, and hence also in its solutions. One has the familiar relations that

$$\sin \mathrm{i} z = \mathrm{i}\, \sinh z\,, \qquad \cos \mathrm{i} z = \cosh z\,. \tag{1.127}$$

The differential equation with the hyperbolic functions as solutions also commonly arises in physics. For example, in a solution by separation of variables, it might be that a separation constant has one sign for certain types of boundary condition, and the opposite sign for other types of boundary condition. And this sign change could precisely manifest itself in taking us from trigonometric to hyperbolic functions.

The story is very similar for the Bessel functions. We have seen that the solutions $J_\nu(z)$ and $Y_\nu(z)$ of Bessel's equation

$$z^2\, y'' + z\, y' + (z^2 - \nu^2)\, y = 0 \tag{1.128}$$

are oscillatory (for real z), at least when $|z|$ is large enough. If we now make the replacement $z \longrightarrow \mathrm{i}\, z$, then the equation takes the form, known as the *Modified Bessel Equation*,

$$z^2\, y'' + z\, y' - (z^2 + \nu^2)\, y = 0\,. \tag{1.129}$$

Clearly its solutions will follow from those of (1.128) by making the replacement $z \longrightarrow \mathrm{i}\, z$ in the arguments of $J_\nu(z)$ and $Y_\nu(z)$.

Actually, our use of the word "clearly" here was perhaps a little optimistic. The problem is that although the basic facts are clear, there is a lot of confusion caused by different notations in the literature. Let's make an uncontroversial definition first. All authors agree to define a "modified Bessel function of the first kind," called $I_\nu(z)$, as follows

$$I_\nu(z) \equiv e^{-\frac{1}{2}\pi\,\nu\,\mathrm{i}}\, J_\nu(z\, e^{\frac{1}{2}\pi\,\mathrm{i}})\,. \tag{1.130}$$

The controversy comes with the choice of definition for the "modified Bessel function of the second kind," called[4] $K_\nu(z)$. Here, we shall define $K_\nu(z)$ as follows:

$$K_\nu(z) \equiv \tfrac{1}{2}\pi\, e^{\frac{1}{2}(\nu+1)\,\pi\,\mathrm{i}}\, H_\nu^{(1)}(z\, e^{\frac{1}{2}\pi\,\mathrm{i}})\,, \tag{1.131}$$

[4]It seems that everybody agrees on its name, and its symbol, if not its definition. It's not clear whether one should regard that as a good thing or a bad thing!

where $H_\nu^{(1)}(z)$ is the first Hankel function, introduced earlier. From our previous definitions, it follows that alternative (equivalent) ways of writing $K_\nu(z)$ are

$$\begin{aligned} K_\nu(z) &= \tfrac{1}{2}\pi\, e^{\frac{1}{2}(\nu+1)\,\pi\,\mathrm{i}} \Big(J_\nu(z\, e^{\frac{1}{2}\pi\,\mathrm{i}}) + \mathrm{i}\, Y_\nu(z\, e^{\frac{1}{2}\pi\,\mathrm{i}})\Big)\,, \\ &= \frac{\pi\,(I_{-\nu}(z) - I_\nu(z))}{2\sin\nu\pi}\,. \end{aligned} \tag{1.132}$$

Obviously, from our previous discussions for $J_\nu(z)$ and $Y_\nu(z)$, it is the case that $I_\nu(z)$ and $K_\nu(z)$ constitute two linearly-independent solutions of the modified Bessel equation.

We shall stick with these definitions. Just as a parenthetic remark, we may note that the chief "rival" to this definition is one where our $K_\nu(z)$ is multiplied by a factor of $\cos\nu\pi$. The logic for this extra factor is that then, the I_ν and the K_ν modified Bessel functions will satisfy identical recurrence relations. Without the $\cos\nu\pi$, there will be slightly different formulae for I_ν and K_ν. The price to be paid, however, for making them uniform in this respect is that the $\cos\nu\pi$ factor will kill off the K_ν function completely if ν is half an odd integer. For that reason, the "rival" definition has fallen into disfavour. Another reason for preferring the definition we are using here is that it is the one used in the algebraic computing language *Mathematica*, which is an immensely powerful tool for analytic mathematical computation.

Having settled on the notation, now let us move on to the more substantial items on the agenda. First, we can immediately write down a power-series expansion for $I_\nu(z)$, valid for small z, by substituting the definition (1.130) into (1.27), to get

$$I_\nu(z) = \sum_{r=0}^{\infty} \frac{1}{r!\,\Gamma(\nu+r+1)}\,\Big(\frac{z}{2}\Big)^{\nu+2r}\,. \tag{1.133}$$

Notice how the phase factor in (1.130) has precisely removed the phase factor arising from replacing z by $z\, e^{\frac{1}{2}\pi\,\mathrm{i}}$ in (1.27), and furthermore, how the $(-1)^r$ factor is also removed.

Recall that we had previously determined that the series expansion (1.27) is convergent in the entire finite complex plane. Since all we have really done is to rotate z through 90 degrees, it follows that the series expansion (1.133) is also convergent in the entire finite complex plane. This does not, however, necessarily mean that it will remain small! Indeed, it is obvious from (1.133) that if we take z to be real and positive, then the series for $I_\nu(z)$ is a sum of positive terms. Therefore, if we take z to be very large and positive, then it follows that $I_\nu(z)$ will get very large. (This does not contradict the convergence of the series. Think of the series for e^z,

$$e^z = \sum_{r=0}^{\infty} \frac{1}{r!}\, z^r\,. \tag{1.134}$$

Again, for real positive z this is the sum of positive terms, and again it follows that for large positive z it gets very large. But we know from kindergarten that the series converges for all finite z.) Keep this fact in mind as we move on to the next stage in the development.

In a moment, we shall present an extremely useful integral representation for $K_\nu(z)$. Before doing so, we shall establish a property of $K_\nu(z)$ which characterises it as being quite distinct in its behaviour from $I_\nu(z)$. We saw in (1.95) how the Hankel function $H_\nu^{(1)}(z)$ behaves at large values of $|z|$. It follows, given the definition (1.131) for $K_\nu(z)$, that at large z we shall have that

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2z}}\, e^{-z}\,. \tag{1.135}$$

Notice again how all the phase factors have nicely cancelled, upon substitution of (1.131) into (1.95). The key point to notice from this is that as z tends to $+\infty$, $K_\nu(z)$ tends to zero.

Now, we can present the integral representation for $K_\nu(z)$. It is

$$K_\nu(z) = \frac{\sqrt{\pi}}{\Gamma(\nu+\frac{1}{2})}\, \Big(\frac{z}{2}\Big)^\nu \int_1^\infty e^{-z\,x}\,(x^2-1)^{\nu-\frac{1}{2}}\,dx\,, \quad \nu > -\tfrac{1}{2}\,, \quad -\tfrac{1}{2}\pi < \arg z < \tfrac{1}{2}\pi\,. \tag{1.136}$$

The proof that this integral really does give $K_\nu(z)$ consists of three parts. First, we prove that it satisfies the modified Bessel equation, which shows that it *must* be some linear combination of $K_\nu(z)$ and $I_\nu(z)$. Next, we prove that in fact it is *purely* a multiple of $K_\nu(z)$, with no contamination from $I_\nu(z)$. Finally, we test its normalisation, to show that it is *exactly* $K_\nu(z)$, and not some constant multiple of it.

To prove that the integral in (1.136) indeed defines a solution of the modified Bessel equation, we simply substitute it in. The easiest way to do this is to define

$$f(z,x) \equiv z^\nu\, e^{-z\,x}\,(x^2-1)^{\nu-\frac{1}{2}}\,. \tag{1.137}$$

This is the "beef" of what appears on the right-hand side of (1.136) before integration, with all the multiplicative constant factors dropped. Now substitute this into the modified Bessel equation (1.129), giving

$$z^2\, f'' + z\, f' - (z^2+\nu^2)\, f = z^{\nu+1}\, e^{-z\,x}\,(x^2-1)^{\nu-\frac{1}{2}}\,(z\,x^2 - z - (2\nu+1)\,x)\,, \tag{1.138}$$

(where a prime means a derivative with respect to z, of course). Now observe that the right-hand side here can be written as a total derivative with respect to x, and so:

$$z^2\, f'' + z\, f' - (z^2+\nu^2)\, f = \frac{d}{dx}\Big[z^{\nu+1}\, e^{-z\,x}\,(x^2-1)^{\nu+\frac{1}{2}}\Big]\,. \tag{1.139}$$

Now integrate this equation with respect to x, evaluated between the limits $x=1$ and $x=\infty$, and recall that, from (1.136), we are hoping to show that the integral of the left-hand side of (1.139) is zero. This is exactly what we find; the integral of the right-hand side of (1.136) gives

$$\Big[z^{\nu+1}\,e^{-z\,x}\,(x^2-1)^{\nu+\frac{1}{2}}\Big]_1^\infty\,, \tag{1.140}$$

and this vanishes at both limits provided that $\nu > -\frac{1}{2}$, and $\mathrm{Re}(z) > 0$. Thus it is established that (1.136) defines a function that satisfies the modified Bessel equation. It follows that it must be some linear combination of the two independent solutions $K_\nu(z)$ and $I_\nu(z)$.

Next, we want to show that there is no "contamination" from $I_\nu(z)$. This is simple, since we have seen that $K_\nu(z)$ and $I_\nu(z)$ have diametrically opposite behaviours for large positive z; $I_\nu(z)$ diverges, whilst $K_\nu(z)$ goes to zero. Now, it is manifest from (1.136) that this integral defines a function that tends to zero as z tends to positive infinity, because of the factor $e^{-z\,x}$ in the integrand. Therefore it must be that the integral is producing purely $K_\nu(z)$, with no admixture of $I_\nu(z)$. (Even a tiny admixture of the form $K_\nu(z) + \epsilon\, I_\nu(z)$, no matter how small ϵ was, would eventually have to diverge for sufficiently large z. Thus we deduce that ϵ must be rigorously zero.)

Finally, we need to check that the normalisation of the integral (1.136) is correct, so that it is producing exactly $K_\nu(z)$,, and not some multiple of it. This can be fixed by looking at a special case, since only one constant muliplication factor needs to be determined. This can be done by looking at large z, and comparing with (1.135). To do this, it is better first to make a change of integration variable in (1.136); we let $x = 1 + t/z$. This gives

$$K_\nu(z) = \sqrt{\frac{\pi}{2z}}\,\frac{e^{-z}}{\Gamma(\nu+\frac{1}{2})}\,\int_0^\infty e^{-t}\,t^{\nu-\frac{1}{2}}\,\Big(1+\frac{t}{2z}\Big)^{\nu-\frac{1}{2}}\,dt\,. \tag{1.141}$$

At large z we can neglect the $t/(2z)$ term in the integrand, since by the time t becomes large enough for $t/(2z)$ to outweigh 1, the e^{-t} factor in the integrand will have rendered the contribution from this portion of the integration range insignificant. Thus approximately we shall have

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2z}}\,\frac{e^{-z}}{\Gamma(\nu+\frac{1}{2})}\,\int_0^\infty e^{-t}\,t^{\nu-\frac{1}{2}}\,dt\,, \tag{1.142}$$

at large z. The integral now just gives $\Gamma(\nu+\frac{1}{2})$, and so we find that

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2z}}\,e^{-z}\,. \tag{1.143}$$

This is exactly the same as the normalisation in (1.135). We have thus completed the demonstration that (1.136) gives precisely the $K_\nu(z)$ modified Bessel function.

The main reason for pursuing this rather lengthy derivation is that the integral representation (1.136) for $K_\nu(z)$ provides us with a very simple way to obtain asymptotic expansions for not only $K_\nu(z)$ itself, but also $I_\nu(z)$, $J_\nu(z)$ and $Y_\nu(z)$, to arbitrary order.

More precisely, it is the integral expression (1.141) that we shall use. All we have to do is to make a binomial expansion of the factor $(1+t/(2z))^{\nu-\frac{1}{2}}$ in the integrand of (1.141), and then integrate term by term. (Recall from Part 1 of the course that one is allowed to integrate term by term in an asymptotic expansion.)

Making the binomial expanion, we find that (1.141) gives

$$\begin{aligned} K_\nu(z) &\sim \sqrt{\frac{\pi}{2z}}\,\frac{e^{-z}}{\Gamma(\nu+\frac{1}{2})}\sum_{r=0}^\infty \frac{\Gamma(\nu+\frac{1}{2})\,(2z)^{-r}}{r!\,\Gamma(\nu-r)}\int_0^\infty e^{-t}\,t^{\nu+r-\frac{1}{2}}\,dt\,, \\ &= \sqrt{\frac{\pi}{2z}}\,e^{-z}\sum_{r=0}^\infty \frac{\Gamma(\nu+r+\frac{1}{2})}{r!\,\Gamma(\nu-r+\frac{1}{2})\,(2z)^r} \end{aligned} \tag{1.144}$$

Using elementary properties of the Gamma function, one can see that this gives us the asymptotic series

$$K_\nu(z)\sim\sqrt{\frac{\pi}{2z}}\,e^{-z}\,\Big[1+\frac{(4\nu^2-1^2)}{1!\,8z}+\frac{(4\nu^2-1^2)(4\nu^2-3^2)}{2!\,(8z)^2}+\cdots\Big]\,. \tag{1.145}$$

Our derivation of this series was based on the use of the integral representation (1.136), which is convergent for $-\frac{1}{2}\pi<\arg z<\frac{1}{2}\pi$. But actually, the asymptotic expansion we have arrived at can be shown to be valid for the wider range of arguments $-\frac{3}{2}\pi<\arg z<\frac{3}{2}\pi$. (Recall that $K_\nu(z)$ has a branch point at $z=0$, as demonstrated by the z^ν factor in its power-series expansion around $z=0$. Therefore, for generic ν, the range $-\frac{3}{2}\pi<\arg z<\frac{3}{2}\pi$ sill covers a lot less than the full range of phases for z that one needs to consider, even though it is more than a complete circulit around the origina of the complex plane.)

We have arrived at the result for the complete asymptotic expansion of $K_\nu(z)$. The leading-order term is the one we found in (1.135), which came, originally, from our steepest-descent analysis of the integral represenations for $J_\nu(z)$ and the Hankel functions. In fact the asynptotic expansions for all the assorted Bessel functions can easily be given in terms of the result (1.145). First, let us write it as

$$K_\nu(z)=\sqrt{\frac{\pi}{2z}}\,e^{-z}\,\Big(P_\nu(\mathrm{i}\,z)+\mathrm{i}\,Q_\nu(\mathrm{i}\,z)\Big)\,, \tag{1.146}$$

where

$$\begin{aligned} P_\nu(z) &\sim 1-\frac{(4\nu^2-1^2)(4\nu^2-3^2)}{2!\,(8z)^2}+\frac{(4\nu^2-1^2)(4\nu^2-3^2)(4\nu^2-5^2)(4\nu^2-7^2)}{4!\,(8z)^4}+\cdots \\ Q_\nu(z) &\sim \frac{(4\nu^2-1^2)}{1!\,(8z)}-\frac{(4\nu^2-1^2)(4\nu^2-3^2)(4\nu^2-5^2)}{3!\,(8z)^3}+\cdots\,. \end{aligned} \tag{1.147}$$

From the original definition (1.131) of $K_\nu(z)$ in terms of $H_\nu^{(1)}(z)$, it then follows that

$$H_\nu^{(1)}(z)=\sqrt{\frac{2}{\pi\,z}}\,e^{\mathrm{i}\,(z-\frac{1}{2}\pi\,\nu-\frac{1}{4}\pi)}\,\Big(P_\nu(z)+\mathrm{i}\,Q_\nu(z)\Big)\,,\qquad -\pi<\arg z<2\pi\,. \tag{1.148}$$

The second Hankel function is the complex conjugate of the first, so

$$H_\nu^{(2)}(z) = \sqrt{\frac{2}{\pi\, z}}\, e^{-\mathrm{i}\,(z - \frac{1}{2}\pi\,\nu - \frac{1}{4}\pi)}\, \Big(P_\nu(z) - \mathrm{i}\, Q_\nu(z)\Big)\,, \qquad -2\pi < \arg z < \pi\,. \tag{1.149}$$

Next, since $J_\nu(z)$ is the real part of $H_\nu^{(1)}(z)$ we shall have

$$J_\nu(z) = \sqrt{\frac{2}{\pi\, z}}\, \Big(P_\nu(z)\, \cos(z - \tfrac{1}{2}\pi\,\nu - \tfrac{1}{4}\pi) - Q_\nu(z)\, \sin(z - \tfrac{1}{2}\pi\,\nu - \tfrac{1}{4}\pi)\Big)\,, \qquad -\pi < \arg z < \pi\,. \tag{1.150}$$

On the other hand $Y_\nu(z)$ is the imaginary part of $H_\nu^{(1)}(z)$, and so

$$Y_\nu(z) = \sqrt{\frac{2}{\pi\, z}}\, \Big(P_\nu(z)\, \sin(z - \tfrac{1}{2}\pi\,\nu - \tfrac{1}{4}\pi) + Q_\nu(z)\, \cos(z - \tfrac{1}{2}\pi\,\nu - \tfrac{1}{4}\pi)\Big)\,, \qquad -\pi < \arg z < \pi\,. \tag{1.151}$$

Finally, since $I_\nu(z)$ is defined in terms of $J_\nu(z)$ by (1.130), we can obtain its asymptotic expansion from (1.150), giving

$$I_\nu(z) = \frac{e^z}{\sqrt{2\pi\, z}}\, \Big(P_\nu(\mathrm{i}\, z) - \mathrm{i}\, Q_\nu(\mathrm{i}\, z)\Big)\,, \qquad -\tfrac{1}{2}\pi < \arg z < \tfrac{1}{2}\pi\,. \tag{1.152}$$

1.10 A Scattering Calculation

The special functions of mathematics, such as the Bessel functions, typically arise when solving Laplace's equation, the Schrödinger equation or the wave equation by the method of separation of variables. One class of physical problem in particular where they can arise is in the study of scattering. A typical situation is that one sits at a large distance (effectively, at infinite distance) from some particle or object, and sends in waves, which are scattered off the object. One then looks at what comes back, from one's vantage point at infinity. To calculate this scattering process, one solves the wave equation (or maybe Schrödinger equation) describing the propagation of the waves under the influence of the scattering object, and imposes appropriate boundary conditions at the scattering centre, as dictated by the physics of the problem. Essentially what one then obtains is an expression for the outgoing and ingoing waves at infinity that result from having sent in an initial wave.

Let us consider a nice example of a scattering problem where we can use some of the Bessel-function technology that we have been studying. The example is not a traditional one, but it has the merit of being simple, and maybe even a bit more interesting than the "old faithfuls." We shall consider a black hole in five spacetime dimensions. As far as the relevant equations are concerned, all that we need to know is that spin-0 fields ϕ propagating in the background geometry of this black hole satisfy the equation

$$\frac{d^2\phi}{dr^2} + \frac{3}{r}\, \frac{d\phi}{dr} + \Big[\omega^2 + \frac{\omega^2 - \ell(\ell+2)}{r^2}\Big]\, \phi = 0\,. \tag{1.153}$$

Here r is the radial coordinate, the black hole event horizon is located at $r = 0$, and we shall sit safely out at infinite distance from it, at $r = \infty$. The constant ω is the frequency of the wave, and ℓ is the angular quantum number analogous to the usual ℓ of quantum mechanics in four spacetime dimensions. (The centrifugal potential in D spacetime dimensions is of the form $\ell(\ell + D - 3)/r^2$, which explains the $\ell(\ell + 2)$ factor here. The factor of $3/r$ multiplying $d\phi/dr$ is aonther tell-tale sign that we are in $D = 5$ dimensions; it would be $(D - 2)/r$ in general.) The equation (1.153) has come from making a rather standard sort of separation of variables, writing the original scalar wavefunction Φ as

$$\Phi = \phi(r)\, Y_\ell\, e^{-\mathrm{i}\,\omega\, t}\,, \tag{1.154}$$

where the Y_ℓ represent spherical harmonics analogous to the familiar $Y_{\ell m}(\theta, \varphi)$, but now they are defined on a 3-sphere rather than a 2-sphere.

If we now let $\phi = \psi/r$, the equation (1.153) becomes

$$r^2\, \frac{d^2\psi}{dr^2} + r\, \frac{d\psi}{dr} + \Big[\omega^2\, r^2 + (\omega^2 - (\ell + 1)^2\Big]\, \phi = 0\,. \tag{1.155}$$

Introducing a new radial coordinate $z = \omega\, r$, and defining

$$\nu^2 = (\ell + 1)^2 - \omega^2\,, \tag{1.156}$$

the equation becomes precisely Bessel's equation

$$z^2\, \psi'' + z\, \psi' + (z^2 - \nu^2)\, \psi = 0\,. \tag{1.157}$$

Thus the solutions for ϕ are

$$\phi = \frac{\alpha}{r}\, J_\nu(\omega\, r) + \frac{\beta}{r}\, J_{-\nu}(\omega\, r)\,. \tag{1.158}$$

Now, we want to study what happens when we send in a wave from infinity, and to see what comes back at us from the black-hole "scatterer." We know the general solution for the waves, so now we must impose the appropriate boundary conditions. In fact the boundary conditions are very simple here. To make an analogy that will be understood by anyone who has ever had to deal with the problem of cockroaches in the kitchen, a black hole works just like the "Roach Motel" that you can buy in the stores. This useful device entices cockroaches into it, whereupon they eat an attractive-tasting poison and die. The advertising slogan for the Roach Motel is "They check in, but they don't check out!" A black hole works in just the same way. Imagine ingoing waves, represented by cockroaches walking radially inwards along the direction of decreasing r, and outgoing waves represented by cockroaches walking radially outwards, with r increasing. The black-hole boundary condition is that at the horizon ($r = 0$), there are only ingoing waves, but no outgoing waves; "they check in, but they don't check out."

How do we recognise a wave that is ingoing and one that is outgoing? Since the time dependence of the wave is of the form $e^{-\mathrm{i}\,\omega\,t}$, as in (1.154), it follows that an ingoing wave is one whose phase *increases* as r *decreases*. For example,

$$\phi \sim e^{-\mathrm{i}\,\omega\,t - \mathrm{i}\,\omega\,r} \tag{1.159}$$

is an ingoing wave, since to sit fixed on a given wavefront one has to go to smaller values of r as t gets bigger. Conversely, an example of an outgoing wave would be

$$\phi \sim e^{-\mathrm{i}\,\omega\,t + \mathrm{i}\,\omega\,r}\,. \tag{1.160}$$

Since we have to impose the boundary condition on the waves at $r = 0$, let us look at that region first. From (1.27), we know that for very small z we shall have

$$J_\nu(z) \sim \frac{1}{\Gamma(\nu+1)}\,\Big(\frac{z}{2}\Big)^\nu\,. \tag{1.161}$$

Thus from (1.158), we see that the r-dependence of the scalar waves will be of the general form $r^{\pm\nu}$, with ν given by (1.156). If ν is real, the solutions are in fact not wavelike at all. To have waves, we shall need the frequency ω to be sufficiently large that ν becomes imaginary, i.e. $\omega > \ell + 1$. Let us therefore assume that this is the case, and define $\nu = \mathrm{i}\,q$, with

$$q \equiv \sqrt{\omega^2 - (\ell+1)^2}\,, \qquad \text{with} \qquad \omega > \ell + 1\,. \tag{1.162}$$

Thus we shall have

$$\phi \approx \frac{\alpha}{r\,\Gamma(1+\mathrm{i}\,q)} e^{\mathrm{i}\,q\,\log(\omega\,r/2)} + \frac{\beta}{r\,\Gamma(1-\mathrm{i}\,q)} e^{-\mathrm{i}\,q\,\log(\omega\,r/2)} \tag{1.163}$$

near $r = 0$. (We have used that $x^y = e^{y\,\log x}$ here.)

We saw previously that an outgoing wave is one whose phase *increases* as r *increases*. This means that the first term in (1.163) is outgoing, while the second term is ingoing. The black-hole boundary condition tells us therefore that

$$\alpha = 0\,, \tag{1.164}$$

which means that the physical wave solutions (1.158) are

$$\phi = \frac{\beta}{r}\,J_{-\mathrm{i}\,q}(\omega\,r)\,. \tag{1.165}$$

Now, we look in the asymptotic region near $r=\infty$. For this, we use the asymptotic expansion (1.90), which is

$$J_\nu(z) \sim \sqrt{\frac{2}{\pi\, z}}\, \cos(z - \tfrac{1}{2}\nu\, \pi - \tfrac{1}{4}\pi)\,. \tag{1.166}$$

(This leading-order term is good enough here.) From (1.165), we therefore have

$$\begin{aligned} \phi &\sim \frac{\beta}{r}\sqrt{\frac{2}{\pi\,\omega\, r}}\, \cos(\omega\, r + \tfrac{1}{2} q\, \pi\, \mathrm{i} - \tfrac{1}{4}\pi)\,, \\ &\sim \frac{\beta}{2r}\sqrt{\frac{2}{\pi\,\omega\, r}}\, e^{\frac{1}{2}q\,\pi}\, e^{\frac{1}{4}\mathrm{i}\,\pi}\, \Big[e^{-\mathrm{i}\,\omega\, r} - \mathrm{i}\, e^{-\pi\, q}\, e^{\mathrm{i}\,\omega\, r}\Big]\,. \end{aligned} \tag{1.167}$$

We recognise the first term in the square bracket as an ingoing wave, and the second term as an outgoing wave.

The prefactor in front of the square bracket in (1.167) is unimportant for our immediate purposes, since it is a common factor in both terms. The key point is that we have found that waves out at infinity have the general structure

$$\psi \sim e^{-\mathrm{i}\,\omega\, r} + S_0\, e^{\mathrm{i}\,\omega\, r}\,, \tag{1.168}$$

with $S_0 = -\mathrm{i}\, e^{-\pi\, q}$. So sending in a wave of unit strength, we get back a wave with strength S_0. Thus S_0 tells us how much comes back, as a fraction of what is sent in. The quantity S_0 is called the *S Matrix.* We can use it to calculate the *Absorption Probability* P, which will in general be given by $P = 1 - |S_0|^2$. Thus for this black hole scattering problem, the absorption probability is given by

$$P = 1 - e^{-2\pi\, q} = 1 - e^{-2\pi\, \sqrt{\omega^2 - (\ell+1)^2}}\,, \qquad \omega > \ell + 1\,. \tag{1.169}$$

On the other hand, when $\omega \le \ell + 1$, there is no absorption at all since there is no wavelike behaviour at the horizon, and so $P = 0$. This matches on smoothly to the result in (1.169). As the frequency of the waves gets larger and larger, the scattering tends exponentially to zero, and accordingly the absorption probability tends to 1. The black hole is behaving more and more like a "sink," with everything that is sent in just disappearing behind the horizon, and no backscatter coming back to the asymptotic region near $r = \infty$.

One can consider many other physical scattering processes, and analyse them in a similar way. The general principles will always be the same, although the details, such as the boundary conditions, will depend on the physical problem one is considering. But always, the idea is to send in waves from infinity, impose appropriate boundary conditions at the scattering centre, and then look at the ratio between ingoing and outgoing wave components at infinity.

Notice that both in the solution of potential-theory problems, and in scattering calculations, an absolutely crucial point is that one needs to know how a *specific* solution of the Bessel equation behaves in different regions. For example, in the scattering calculation we needed to know the asymptotic behaviour at large z for the solution that had a given behaviour near $z = 0$. It would not be good enough simply to know that for small z the two solutions of Bessel's equation look like

$$u_1 \sim z^{\nu}\,, \qquad u_2 \sim z^{-\nu}\,, \tag{1.170}$$

(see (1.161)), and that for large z the two solutions look like

$$v_1 \sim z^{-\frac{1}{2}} \cos z\,, \qquad v_2 \sim z^{-\frac{1}{2}} \sin z\,, \tag{1.171}$$

(see (1.166)). (These asymptotic forms could, for example, be obtained directly from the Bessel equation, by taking z to be small or large respectively.) The crucial point is that one needs to know *exactly* what the relation between the small-z and large-z forms of a specific solution are; in particular, one needs to know exactly what the constants a_i and b_i are in the relation $v_1 = a_1\, u_1 + b_1\, u_2$ and $v_2 = a_2\, u_1 + b_2\, u_2$. This is precisely the sort of information that we *have* been able to obtain as a result of having integral representations for the Bessel functions.

2 Hypergeometric and Confluent Hypergeometric Functions

2.1 Hypergeometric Functions

Let us begin by considering the following power series,

$$y(z) = 1 + \frac{a\, b}{c}\, \frac{z}{1!} + \frac{a(a+1)\, b(b+1)}{c(c+1)}\, \frac{z^2}{2!} + \frac{a(a+1)(a+2)\, b(b+1)(b+2)}{c(c+1)(c+2)}\, \frac{z^3}{3!} + \cdots \tag{2.1}$$

which can be conveniently written as

$$y(z) = \sum_{n=0}^{\infty} \frac{(a)_n\, (b)_n}{(c)_n}\, \frac{z^n}{n!}\,, \tag{2.2}$$

where we define the *Pochhammer symbol* $(a)_n$ by

$$(a)_n \equiv \frac{\Gamma(a+n)}{\Gamma(a)} = a(a+1)(a+2)\cdots(a+n-1)\,. \tag{2.3}$$

(Note that $(a)_0 = 1$.) The function defined by this power series is called the *Hypergeometric Function* ${}_2F_1(a, b; c; z)$; thus

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n\, (b)_n}{(c)_n}\, \frac{z^n}{n!}\,. \tag{2.4}$$

It is, apparently, called the hypergeometric function because it is a natural generalisation of the function $1/(1-z)$ that gives the geometric series $1+z+z^2+z^3+\cdots$. The notation with the subscripts 2 and 1 on the ${}_2F_1$ signifies that the series expansion has 2 Pochhammer symbols in the numerator, and 1 in the denominator. The use of semicolons as delimiters for the c parameter is conventional too. Notice that because of the fact that $\Gamma(x)$ is infinite when $x=0$ or a negative integer, the parameter c must not be zero or a negative integer. On the other hand, if a or b is zero or a negative integer, then the series terminates and becomes just a finite polynomial. Note also that ${}_2F_1(a,b;c;z)$ is equal to ${}_2F_1(b,a;c;z)$.

It is easy to see that the hypergeometric function satisfies the *Hypergeometric Equation*

$$z(1-z)\,y''(z)+[c-(a+b+1)\,z)]\,y'(z)-a\,b\,y(z)=0\,. \tag{2.5}$$

We can check this by simply plugging (2.4) into (2.5), and shifting the summation variables in each term as necessary so as to get z-dependence z^n for each term. In other words, just check that the coefficient of each power of z vanishes. To do this, it is useful to observe that the Pochhammer symbol satisfies the relation

$$\begin{aligned}(a)_{n+1} &= \frac{\Gamma(a+n+1)}{\Gamma(a)}=(a+n)\,\frac{\Gamma(a+n)}{\Gamma(a)}\,,\\ &= (a+n)\,(a)_n\,. \end{aligned}\tag{2.6}$$

We discussed the hypergeometric equation a little in Part 1 of the course. Dividing (2.5) by $z(1-z)$, we see that the coefficient of $y'(z)$ then has first-order poles $1/z$ and $1/(1-z)$, as does the coefficient of $y(z)$ (since $z^{-1}\,(1-z)^{-1}=z^{-1}+(1-z)^{-1}$). Recalling that the differential equation

$$y''(z)+p(z)\,y'(z)+q(z)\,y(z)=0 \tag{2.7}$$

has a *regular singular point* at $z=z_0$ if $p(z)$ and/or $q(z)$ diverge there, but $(z-z_0)\,p(z)$ and $(z-z_0)^2\,q(z)$ are finite, we see that the hypergeometric equation has regular singular points at $z=0$ and $z=1$. Furthermore, if we let $z=1/w$, we find that the transformed equation is

$$(w-1)\,\frac{d^2y}{dw^2}+[2-c+(a+b-1)\,w^{-1}]\,\frac{dy}{dw}-\frac{a\,b}{w^2}\,y=0\,, \tag{2.8}$$

and therefore $w=0$, corresponding to $z=\infty$, is also a regular singular point. Thus the hypergeometric equation is non-singular everywhere except at three regular singular points, located at $z=0$, 1 and ∞. *Any* second-order linear ordinary differential equation with three regular singular points can be transformed into the canonical form of the hypergeometric equation, by making appropriate chages of variable, and so it encompasses a rather broad class of differential equations, including many that one encounters in physics.

It is a standard result in the theory of differential equations, which we discussed in Part 1, that at least one of the two solutions of a second-order ODE (ordinary differential equation) can be obtained as an expansion around a regular singular point z_0 of the equation, in the form

$$y = (z - z_0)^s \sum_{n=0}^{\infty} a_n \, (z - z_0)^n \,, \tag{2.9}$$

where s is a root of a certain second-order polynomial equation called the *indicial equation.*[5] Furthermore, in a situation where the function $q(z)$ in (2.7) actually happens not to have a second-order pole contribution at the regular singular point, one root of the indicial equation is $s = 0$. This is the case at $z = 0$ in the hypergeometric equation, and so we know that there should certainly exist one solution that is a pure analytic power series when expanded around the point $z = 0$. This is exactly what we have in (2.4); a pure analytic power-series solution to the hypergeometric equation.

Another standard result from the theory of ODEs is that the radius of convergence of this power series solution will be equal to the distance from the expansion point, $z = 0$, to the next nearest singular point of the equation. In the case of the hypergeometric equation, this will be the regular singular point at $z = 1$. Thus we learn that the power series (2.4) is convergent in the disk $|z| < 1$. This can easily be verified by applying the ratio test for convergence of a series. We take the ratio R of the $(n+1)$'th term divided by the n'th term. If the modulus of this ratio is less than 1 in the limit as n tends to infinity, then the series converges absolutely; if it is greater than 1 it diverges, and if it equals 1, a more delicate analysis is needed. In our case, from (2.4), we have

$$R = \frac{(a)_{n+1}\,(b)_{n+1}}{(c)_{n+1}\,(n+1)!}\,\frac{(c)_n\, n!}{(a)_n\,(b)_n}\, z = \frac{(n+1)\,(n+c)}{(n+a)\,(n+b)}\, z \tag{2.10}$$

in the limit when $n \longrightarrow \infty$, implying that we get $|R| = |z|$. Thus the series indeed converges for $|z| < 1$, and diverges for $|z| > 1$.

The hypergeometric equation, being of second order, must have two linearly-independent solutions. We may, in general, obtain the second solution as follows. Make the substitution $y(z) = z^{1-c}\, w(z)$ in the hypergeometric equation (2.5). After a couple of lines of simple algebra, one finds that $w(z)$ satisfies

$$z\,(1-z)\,w'' + [2 - c - (a + b - 2c + 3)\,z]\,w' - (a - c + 1)(b - c + 1)\,w = 0\,. \tag{2.11}$$

[5]Generically, if the two roots s_1 and s_2 of the indicial equation do not differ by an integer, then both solutions can be obtained in the form (2.9). But more often than not, life being what it is, it turns out that cases of particular interest correspond to the situation where $s_1 - s_2$ *is* and integer.

This can be recognised as the hypergeometric equation again, but now with the parameters $(a-c+1, b-c+1, 2-c)$ instead of (a,b,c). Thus we see that

$$y_2 = z^{1-c}\, {}_2F_1(a-c+1, b-c+1; 2-c; z) \tag{2.12}$$

is another solution of the hypergeometric equation. It is obvious that if c is not an integer, this solution is linearly independent of the original solution ${}_2F_1(a,b;c;z)$, since (2.12) is a then a power series in non-integer powers of z whereas ${}_2F_1(a,b;c;z)$ is a power series in integer powers of z. If c is an integer then one can show that (2.12) is in general the same solution as ${}_2F_1(a,b;c;z)$ (except for special values of the parameters a and b). The situation is very reminiscent of the Bessel equation, where $J_{-\nu}(z)$ provides a solution that is independent of $J_\nu(z)$, except when ν is an integer. As in that case, it turns out here that in such a "degenerate" situation, the second independent solution will include logarithm terms.

We may construct an integral representation for the hypergeometric function as follows. We begin by introducing the *Beta Function* $B(p,q)$, defined as[6]

$$B(p,q) \equiv \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}\,. \tag{2.13}$$

Clearly $B(p,q) = B(q,p)$. Now consider the following expression for $\Gamma(p)\,\Gamma(q)$, which is obtained just by taking the product of two standard integral representations for the Gamma function:

$$\Gamma(p)\,\Gamma(q) = \int_0^\infty e^{-u}\, u^{p-1}\, du \int_0^\infty e^{-v}\, v^{q-1}\, dv\,. \tag{2.14}$$

Now let $u = x^2$, $v = y^2$ and then change to polar coordinates; $x = r\,\cos\theta$, $y = r\,\sin\theta$;

$$\begin{aligned}
\Gamma(p)\,\Gamma(q) &= 4\int_0^\infty dx \int_0^\infty dy\, e^{-x^2-y^2}\, x^{2p-1}\, y^{2q-1} \\
&= 4\int_0^\infty dr \int_0^{\frac12\pi} d\theta\, e^{-r^2}\, r^{2p+2q-1}\, (\cos\theta)^{2p-1}\, (\sin\theta)^{2q-1} \\
&= 2\int_0^\infty d\rho \int_0^{\frac12\pi} d\theta\, e^{-\rho}\, \rho^{p+q-1}\, (\cos\theta)^{2p-1}\, (\sin\theta)^{2q-1} \\
&= 2\Gamma(p+q) \int_0^{\frac12\pi} d\theta\, (\cos\theta)^{2p-1}\, (\sin\theta)^{2q-1}\,,
\end{aligned} \tag{2.15}$$

where in the third line we have changed variable again, from r to $\rho = r^2$, allowing us to recognise a standard integral representation for $\Gamma(p+q)$. Finally, the further change of variable from θ to $t = \sin^2\theta$ yields the result that

$$B(p,q) = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)} = \int_0^1 (1-t)^{p-1}\, t^{q-1}\, dt\,. \tag{2.16}$$

[6] An upper-case Greek beta is written as B.

Using the Beta function, we can therefore write the ratio $(b)_n/(c)_n$ in the power series for the hypergeometric function as

$$\frac{(b)_n}{(c)_n} = \frac{B(b+n,c-b)}{B(b,c-b)} = \frac{1}{B(b,c-b)} \int_0^1 (1-t)^{c-b-1}\, t^{b+n-1}\, dt\,. \tag{2.17}$$

Thus from (2.4) we shall have

$${}_2F_1(a,b;c;z) = \frac{1}{B(b,c-b)} \sum_{n=0}^{\infty} \frac{(a)_n)}{n!}\, z^n \int_0^1 (1-t)^{c-b-1}\, t^{b+n-1}\, dt\,. \tag{2.18}$$

Interchanging the order of the integration and summation, we can sum the resulting series by noting from the binomial theorem that

$$\sum_{n=0}^{\infty} \frac{(a)_n}{n!}\, z^n\, t^n = \sum_{n=0}^{\infty} \frac{\Gamma(a+n)}{\Gamma(a)\, n!}\, (z\,t)^n = (1-z\,t)^{-a}\,. \tag{2.19}$$

Thus we arrive at the following integral representation for the hypergeometric function:

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\,\Gamma(c-b)} \int_0^1 (1-t)^{c-b-1}\, t^{b-1}\, (1-z\,t)^{-a}\, dt\,. \tag{2.20}$$

This is valid for any complex value of z provided that z is not real and larger than 1. (This restriction ensures that the $(1-z\,t)^{-a}$ factor does not give rise to a pole or branch point in the integrand at $t = 1/z$.) The branch of $(1-x\,t)^{-a}$ must be chosen so that $(1-x\,t)^{-a} \longrightarrow 1$ as t goes to zero, and the parameters b and c must be such that Re(c) >Re$(b) > 0$. Note that this represents an analytic continuation of the original power-series expression (2.4) for ${}_2F_1(a,b;c;z)$, which was convergent only for $|z| < 1$.

By playing around with this integral representation, and others, one can establish many properties and inter-relations among hypergeometric functions. We shall not go into too much further detail here, since the subject is a vast one, and is discussed at length in many books. We shall just record a few more facts here, without proof, to show the sort of relations that one can establish. Firstly, there is another integral representation for the hypergeometric function, known as the *Barnes Integral*,

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{2\pi\,\mathrm{i}\,\Gamma(a)\,\Gamma(b)} \int_{-\mathrm{i}\infty}^{\mathrm{i}\,\infty} \frac{\Gamma(a+s)\,\Gamma(b+s)\,\Gamma(-s)}{\Gamma(c+s)}\, (-z)^s\, ds\,, \tag{2.21}$$

which is proven by establishing that the term $(a)_n\,(b)_n\,z^n/((c)_n\,n!)$ in the power-series expansion (2.4) is the residue of the integrand at $s = n$. This integral gives the hypergeometric function as a function analytic in the domain defined by the inequality $|\arg z| < \pi$, and so again, it is an analytic extension of the original series definition (2.4).

One can use the Barnes representation (2.21) in order to obtain a new power series for ${}_2F_1(a,b;c;z)$ that is convergent when $|z| > 1$. After some effort, one arrives at the conclusion that

$$\frac{\Gamma(a)\,\Gamma(b)}{\Gamma(c)}\,{}_2F_1(a,b;c;z) = \frac{\Gamma(a)\,\Gamma(b-a)}{\Gamma(c-a)}\,(-z)^{-a}\,{}_2F_1(a,a-c+1;a-b+1;z^{-1}) + \frac{\Gamma(b)\,\Gamma(a-b)}{\Gamma(c-b)}\,(-z)^{-b}\,{}_2F_1(b,b-c+1;b-a+1;z^{-1})\,, \tag{2.22}$$

where $|\arg(-z)| < \pi$. Since the hypergeometric functions on the right-hand side both have $1/z$ as argument, it follows that when $|z| > 1$ the original power series (2.4) can be used in order to obtain a series expansion for the right-hand side, and hence a series expansion for ${}_2F_1(a,b;c;z)$ that is convergent for $|z| > 1$. The formula (2.22) is typical of many relations that one can obtain, relating ${}_2F_1(a,b;c;z)$ to hypergeometric functions with argument $1/z$ or $(1-z)$ or $z/(1-z)$, and so on. It can easily be shown that each term on the right-hand side of (2.22) is separately a solution of the original hypergeometric equation.

Notice that the power series in $1/z$ that we obtain by using (2.22) together with the original series (2.4) is a perfectly convergent one, rather than an asymptotic expansion. This is because $z = \infty$ is a regular singular point of the hypergeometric equation. In the next subsection we shall see what happens when we take a singular limit of the parameters in the hypergeometric equation, resulting in the regular singular point at $z = 1$ being moved out to join the one at $z = \infty$. In this limit the point at infinity becomes an irregular singular point, and correspondingly one is back to the situation where one can obtain only an asymptotic expansion, as opposed to a convergent power-series expansion, around $z = \infty$. In fact, as we shall see, this limit in which two regular singular points join together to make an irregular singular point gives an equation, called the confluent hypergeometric equation, that includes our old friend the Bessel equation as a special case.

2.2 Confluent Hypergeometric Functions

We have seen that the hypergeometric equation

$$z\,(1-z)\,y''(z) + [c - (a+b+1)\,z)]\,y'(z) - a\,b\,y(z) = 0\,. \tag{2.23}$$

has three singular points, all of them *regular singular points*, located at $z = 0$, 1 and ∞. Their precise locations can be moved around by making transformations of z, such as constant shifts and scalings. Consider in particular the following transformation, under which

$$z \longrightarrow \frac{z}{b}\,, \tag{2.24}$$

implying that the hypergeometric equation becomes

$$z\,(1 - z\,b^{-1})\,y''(z) + [c - (a+b+1)\,b^{-1}\,z)]\,y'(z) - a\,y(z) = 0\,, \tag{2.25}$$

(after dividing out by b). Evidently, at this stage the singular points of the equation have been transformed to $z = 0$, b and ∞.

Now, let us send b to infinity. We can see that this is a perfectly well-defined limit of the equation (2.25), which leads to

$$z\, y'' + (c - z)\, y' - a\, y = 0\,. \tag{2.26}$$

This is called the *Confluent Hypergeometric Equation.* The name comes from the fact that the two regular singular points $z = b$ and $z = \infty$ in (2.25) have joined together (in a confluence), at $z = \infty$. Because they are now superimposed, one finds that the singularity at $z = \infty$ is now more divergent, and in fact it is now an *irregular singular point.* (One shows this by the usual procedure of letting $z = 1/w$, and studying the structure of the singularity in the equation at $w = 0$.)

Let us see what has happened to the hypergeometric function ${}_2F_1(a, b; c; z)$ that was a solution of the hypergeometric equation, in this limiting process. We shall have

$$\lim_{b\to\infty}\ {}_2F_1(a, b; c; z/b)\,. \tag{2.27}$$

From (2.4), the b dependence of the term in z^n in the power series for ${}_2F_1(a, b; c; z/b)$ will therefore be $(b)_n/b^n$, and so we have

$$\lim_{b\to\infty} \frac{(b)_n}{b^n} = \lim_{b\to\infty} \frac{b\,(b+1)(b+2)\cdots(b+n-1)}{b^n} = 1\,. \tag{2.28}$$

Thus we have the solution

$${}_1F_1(a; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(c)_n}\, \frac{z^n}{n!} \tag{2.29}$$

to the confluent hypergeometric equation (2.26). Observe that the notation here is in accordance with the previous one, namely that the subscripts 1 and 1 on ${}_1F_1$ signify that there is 1 Pochhammer symbol in the numerator, and 1 in the denominator, in each term in the series.

Now that we have derived it, let us change the symbols of its arguments to the more conventional ones ${}_1F_1(a; b; z)$. This function is called a *Confluent Hypergeometric Function,* or a *Kummer Function.* It is often denoted by the symbol $M(a, b, z)$, and its full name is *Kummer's regular function*, so we have

$$M(a, b, z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n}\, \frac{z^n}{n!}\,, \tag{2.30}$$

satisfying the confluent hypergeometric equation

$$z\, y'' + (b - z)\, y' - a\, y = 0\,. \tag{2.31}$$

Since the singular point of the equation nearest to the regular singularity at $z = 0$ is the irregular singular point at $z = \infty$, we know that the series (2.30) will be convergent everywhere in the finite complex plane.

The same limiting process can be applied also to the second solution (2.12) of the hypergeometric equation. Doing so, we obtain the second solution for the confluent hypergeometric equation,

$$y_2 = z^{1-b}\, M(a-b+1, 2-b, z)\,. \tag{2.32}$$

As in the case of the hypergeometric equation, here this solution to the confluent hypergeometric equation is linearly-independent of $y_1 \equiv M(a,b,z)$ as long as b is not an integer.

If, on the other hand, $b = 1$ then clearly y_2 is exactly equal to y_1. If $b = N$, where N is an integer ≥ 2, then y_2 becomes singular, but can be rescaled by an appropriate constant factor before setting $b = N$ so as to render the expression finite. It then turns out to be proportional to y_1 again. For example, using the power-series expansion (2.29), the second solution given in (2.32) has the form

$$\begin{aligned} y_2 &= z^{1-b}\Big(1 + \frac{(a-b+1)\,z}{2-b} + \frac{(a-b+1)(a-b+2)\,z^2}{2!\,(2-b)(3-b)} \\ &\qquad + \frac{(a-b+1)(a-b+2)(a-b+3)\,z^3}{3!\,(2-b)(3-b)(4-b)} + \cdots\Big)\,. \end{aligned} \tag{2.33}$$

Clearly each term beyond the first diverges as b is set equal to 2, but if we first multiply by $(2-b)$, and then set $b = 2$, we get the finite result

$$y_2 = (a-1)\,\Big(1 + \tfrac12 a\, z + \tfrac1{12}\, a(a+1)\, z^2 + \tfrac1{144} a(a+1)(a+2)\, z^3 + \cdots\Big)\,. \tag{2.34}$$

This can be compared with the series expansion of $M(a,b,z)$ itself at $b = 2$, which, from (2.29), is

$$M(a,2,z) = 1 + \tfrac12 a\, z + \tfrac1{12}\, a(a+1)\, z^2 + \tfrac1{144} a(a+1)(a+2)\, z^3 + \cdots\,. \tag{2.35}$$

Thus at $b = 2$ we have that

$$\lim_{b\to 2}\, (2-b)\, y_2 = (a-1)\, y_1\,, \tag{2.36}$$

with analogous results at $b = 3$, 4, 5, *etc.*

This is exactly like the situation with the $J_\nu(z)$ and $J_{-\nu}(z)$ Bessel functions, at $\nu =$ integer. As in that case, the way to extract a second linearly-independent solution is to take the difference between the two solutions that*are* independent for non-integer parameter b, and divide out by an appropriate factor that vanishes as b approaches an integer, so as to recover a finite result analogous to $Y_n(z)$. Thus one defines the second solution here to be

$$U(a,b,z) \equiv \frac{\pi}{\sin \pi b} \Big[\frac{M(a,b,z)}{\Gamma(b)\Gamma(a-b+1)} - \frac{z^{1-b}\, M(a-b+1, 2-b, z)}{\Gamma(a)\Gamma(2-b)} \Big]\,. \tag{2.37}$$

Following similar steps to those that we used for $Y_n(z)$, one can find the series expansion for $U(a,b,z)$ around $z=0$. This involves showing first that the quantity in square brackets in (2.37) vanishes at $b = N = 2, 3, 4, \ldots$, and then carefully expanding around $b = N + \epsilon$ and picking up the terms of first order in ϵ. For example, by doing this for $b=2$ one finds that $U(a,2,z)$ becomes

$$U(a,2,z) = \frac{1}{\Gamma(a)\, z} + \frac{2\gamma + \psi(a) + \log z}{\Gamma(a-1)} + O(z, z\, \log z)\,. \tag{2.38}$$

Here γ is the Euler-Mascheroni constant and $\psi(s) = \Gamma(s)'/\Gamma(s)$ is the Digamma function. We see the familiar appearance of logarithmic terms in the series expansion. On account of this non-analyticity at $z=0$, the function $U(a,b,z)$ is called *Kummer's Irregular Function.*

In general it can be shown that at $b = n+1$, where $n \ge 0$ is an integer, the function $U(a,b,z)$ has the series expansion

$$\begin{aligned} U(a, n+1, z) \quad &= \quad \frac{(-1)^{n+1}}{n!\, \Gamma(a-n)} \Big[M(a, n+1, z)\, \log z + \\ &\qquad \sum_{r=0}^\infty \frac{(a)_r\, z^r}{(n+1)_r\, r!} \Big(\psi(a+r) - \psi(r+1) - \psi(n+r+1) \Big) \Big] \\ &\qquad + \frac{(n-1)!}{\Gamma(a)}\, z^{-n}\, M(a-n, 1-n, z)_n\,, \end{aligned} \tag{2.39}$$

where the notation $M(a-n, 1-n, z)_n$ means that just the first n terms in the series expansion for $M(a-n, 1-n, z)$ are retained.

We can also derive integral representations for the Kummer functions, by taking the appropriate limit in the original expressions for the hypergeometric functions. For example, we may begin with the integral representation (2.20) for ${}_2F_1(a,b;c;z)$. Now we actually know that this must be symmetric under the exchange of the labels a and b, even though it is not obvious, since the original series expansion for the hypergeometric function is symmetric in a and b. Thus we know from (2.20) that we must also have

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(a)\, \Gamma(c-a)} \int_0^1 (1-t)^{c-a-1}\, t^{a-1}\, (1-z\, t)^{-b}\, dt\,. \tag{2.40}$$

In this form, the process of replacing z by z/b and sending b to infinity is easily implemented, since the only b dependence comes from the factor

$$(1 - z\, t\, b^{-1})^{-b}\,. \tag{2.41}$$

Now it is a standard result[7] that the limit of $(1-x/b)^{-b}$ as b tends to infinity is just e^x, and hence we find that

$$\lim_{b\to\infty} {}_2F_1(a,b;c;z\,b^{-1}) = \frac{\Gamma(c)}{\Gamma(a)\,\Gamma(c-a)} \int_0^1 (1-t)^{c-a-1}\, t^{a-1}\, e^{z\,t}\, dt\,. \tag{2.42}$$

Finally, replacing c by b for convenience, we have the result that

$$M(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b-a)} \int_0^1 (1-t)^{b-a-1}\, t^{a-1}\, e^{z\,t}\, dt\,. \tag{2.43}$$

This has restrictions on the values of the parameters that follow directly from those for the hypergeometric integral (2.20), namely that $\mathrm{Re}(b) > \mathrm{Re}(a) > 0$. It is valid for any finite z, and so it defines $M(a,b,z)$ as a function analytic everywhere in the finite complex plane. This accords with the fact that the series expansion (2.30) is convergent for all finite z.

One can easily show from (2.43), by making the change of integration variable $t = 1-s$, that

$$M(a,b,z) = e^z\, M(b-a,b,-z)\,. \tag{2.44}$$

This is known as *Kummer's first formula.*

To close this section, here are some examples that show how special cases of the confluent hypergeometric functions correspond to other well-known functions. The Bessel functions, for example, are special cases:

$$\begin{aligned} M(\nu+\tfrac{1}{2}, 2\nu+1, 2\mathrm{i}\,z) &= \Gamma(\nu+1)\, e^{\mathrm{i}\,z} \left(\tfrac{1}{2}z\right)^{-\nu} J_\nu(z)\,, \\ U(\nu+\tfrac{1}{2}, 2\nu+1, 2\mathrm{i}\,z) &= \tfrac{1}{2}\sqrt{\pi}\, e^{-\pi\,\mathrm{i}\,(\nu+\frac{1}{2})}\, e^{\mathrm{i}\,z}\, (2z)^{-\nu}\, H_\nu^{(2)}(z)\,. \end{aligned} \tag{2.45}$$

Among many other special case are the exponential function $e^z = M(a,a,z)$, the Laguerre polynomials

$$M(-n,\alpha+1,z) = \frac{n!}{(\alpha+1)_n}\, L_n^{(\alpha)}(z)\,, \tag{2.46}$$

and the Hermite polynomials

$$M(-n,\tfrac{1}{2},\tfrac{1}{2}z^2) = \frac{(-\frac{1}{2})^{-n}\, n!}{(2n)!}\, H_{2n}(z)\,, \quad M(-n,\tfrac{3}{2},\tfrac{1}{2}z^2) = \frac{(-\frac{1}{2})^{-n}\, n!}{(2n+1)!}\, z^{-1}\, H_{2n+1}(z)\,. \tag{2.47}$$

---

[7]Which can be proven by noting that at large b we have $1 - x/b = e^{-x/b} + O(b^{-2})$, implying that $(1-x/b)^{-b} = (e^{-x/b})^{-b}\,(1 + e^{x/b}\,O(b^{-2}))^{-b} = e^x\,(1 + e^{x/b}\,O(b^{-2}))^{-b}$. Now note that $1 + e^{x/b}\,O(b^{-2})$ has the form $e^{y/b^2} + O(b^{-3})$ for some y, and hence $(1 + e^{x/b}\,O(b^{-2}))^{-b} = e^{-y/b}\,(1 + e^{-y/b^2}\,O(b^{-3}))^{-b}$. Iterating this, we see that all the factors associated with these higher terms become 1 as b is sent to infinity, leaving the result e^x

2.3 Asymptotic Expansions and the Stokes Phenomenon

Since the point $z = \infty$ in the confluent hypergeometric equation is an irregular singular point, we expect that any series expansions for its solutions expanded around $z = \infty$ will be asymptotic series rather than convergent ones. We can study this in detail for the regular Kummer function $M(a, b, z)$ by making use of the integral representation (2.43).

First, we must contrive by making an appropriate change of variables to separate out the z dependence in the exponential function from the t dependence, in such a way that we can make a series expansion of the integrand in inverse powers of z. We need the sort of transformation of integration variable that took the integral representation (1.136) for the modified Bessel function $K_\nu(z)$ into the form (1.141). However, this does not work out quite so easily in the present case, on account of the range of the integration variable t in (2.43) being $[0, 1]$ rather than $[1, \infty]$. The answer to how to handle this problem is a rather simple one, namely to write the integral $\int_0^1$ as $\int_0^1 = \int_{-\infty}^1 - \int_{-\infty}^0$. Thus we rewrite (2.43) as

$$M(a, b, z) = \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b-a)} \Big[\int_{-\infty}^1 (1-t)^{b-a-1}\, t^{a-1}\, e^{z\,t}\, dt - \int_{-\infty}^0 (1-t)^{b-a-1}\, t^{a-1}\, e^{z\,t}\, dt\Big]\,. \tag{2.48}$$

Note that this choice of lower limit $-\infty$ on both the integrals is an appropriate one when $\text{Re}(z)$ is positive.[8]

Let us consider first the case where z is taken to be real, positive and large. In the first integral, we make the change of variable from t to u defined by $t = 1 - u/z$, while in the second integral we change to w defined by $t = -w/z$. Both integrals now run from 0 to ∞ over their respective integration variables:

$$\begin{aligned} M(a, b, z) &= \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b-a)} \Big[z^{a-b}\, e^z \int_0^\infty e^{-u}\, u^{b-a-1}\, (1 - u\,z^{-1})^{a-1}\, du \\ &\qquad + (-z)^{-a} \int_0^\infty e^{-w}\, w^{a-1}\, (1 + w\,z^{-1})^{b-a-1}\, dw\Big]\,. \end{aligned} \tag{2.49}$$

We shall see below that the two integrals are approximately equal to $\Gamma(b-a)$ and $\Gamma(a)$ respectively, which are finite and non-zero for generic a and b. Since we are considering the case where z is real, large and positive it follows that the contribution from the first term will be overwhelmingly larger than that from the second term, on account of the e^z prefactor. Thus only the first term will contribute in the asymptotic expansion for large positive z.

[8] Of course one can write $\int_0^1 = \int_{t_0}^1 - \int_{t_0}^0$ for *any* choice of t_0. We shall see below that a choice other than $t_0 = -\infty$ becomes appropriate when z is to be taken large and negative.

Notice how with these changes of variable we have contrived to turn the integrands into functions that can be expanded in power series in $1/z$. Specifically, to evaluate the first term in (2.49) we use the binomial theorem to obtain

$$(1 - u\,z^{-1})^{a-1} = \sum_{r=0}^{\infty} \frac{\Gamma(a)}{r!\,\Gamma(a-r)}\,\Big(-\frac{u}{z}\Big)^r\,. \tag{2.50}$$

Substituting this into the first integral in (2.49), the term-by-term integration becomes a triviality, since all the terms are of the form $\int_0^\infty e^{-x}\,x^{c-1}\,dx$, which is just $\Gamma(c)$. Thus we obtain the asymptotic expansion for $M(a,b,z)$, valid when z is real, large and positive:

$$M(a,b,z) \sim \frac{\Gamma(b)}{\Gamma(b-a)}\,z^{a-b}\,e^z\,\sum_{r=0}^{\infty} \frac{\Gamma(b-a+r)}{r!\,\Gamma(a-r)}\,\Big(-\frac{1}{z}\Big)^r\,. \tag{2.51}$$

It should be emphasised that *every* term in this expansion is more important than even the leading-order term coming from the second integral in (2.49) that we dropped.

A brief pause for a word on terminology is appropriate here. Strictly speaking, we should not call (2.51) itself an asymptotic expansion; the exponentials factor e^z is not strictly allowed in the definition of an asymptotic series. Rigorously-speaking, an asymptotic series must involve a sum only over (inverse) powers of z, of the form $\sum_{n\ge 0} z^{c-n}$. And in fact, as we discussed in Part I, the exponential function e^z itself has the asymptotic expansion $e^z \sim 0$ when z tends to $-\infty$, and admits no asymptotic expansion at all when z tends to $+\infty$. So strictly speaking, we should really take the e^z factor in (2.51) over to the left-hand side, and say that it is $e^{-z}\,M(a,b,z)$ that has the asymptotic expansion (given by (2.51) with the e^z factor omitted). Of course we actually know perfectly well how e^z behaves at large positive and negative z and so in fact we are perfectly happy to leave it in there on the right-hand side, and in practice we usually refer to (2.51) as an asymptotic series for $M(a,b,z)$. But it is worth bearing this point in mind, to avoid possible confusion later.

Now, consider instead the situation when z is real, large and negative, so that $z = -|z|$. In this case, we should use the identity that $\int_0^1 = \int_0^\infty - \int_1^\infty$. Using this in (2.43), we now make the changes of variable $t = u/|z|$ in the first of these integrals, and $t = 1 + w/|z|$ in the second. This leads to the expression

$$\begin{aligned} M(a,b,z) \;\; &= \;\; \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b-a)}\,\Big[|z|^{-a}\int_0^\infty e^{-u}\,u^{a-1}\,(1-u\,|z|^{-1})^{b-a-1}\,du \\ &\quad -(-|z|)^{b-a-1}\,e^{-|z|}\int_0^\infty e^{-w}\,w^{b-a-1}\,(1+w\,|z|^{-1})^{a-1}\,dw\Big]\,. \end{aligned} \tag{2.52}$$

This time, it is clear that as z tends to $-\infty$ the first term overwhelmingly dominates over the second, because of the $e^{-|z|}$ prefactor in the second term. Again we perform a binomial expansion of the z-dependent factor in the integrand of the first term, this time obtaining the following asymptotic expansion, valid for z real, large and negative:

$$M(a,b,z) \sim \frac{\Gamma(b)}{\Gamma(a)}\,|z|^{-a}\,\sum_{r=0}^{\infty}\frac{\Gamma(a+r)}{r!\,\Gamma(b-a-r)}\,\Big(-\frac{1}{|z|}\Big)^r\,. \tag{2.53}$$

The nature of the asymptotic expansions for $M(a,b,z)$ for large positive z and for large negative z are totally different. To emphasise the point, let's compare the leading-order terms in the two cases:

$$M(a,b,z) \sim \begin{cases} \frac{\Gamma(b)}{\Gamma(a)}\,z^{a-b}\,e^z\,, & z\longrightarrow +\infty \\ \\ \frac{\Gamma(b)}{\Gamma(b-a)}\,|z|^{-a} & z\longrightarrow -\infty \end{cases} \tag{2.54}$$

Actually, we should not be surprised by the fact that a function can have totally different asymptotic expansions depending upon the direction in which one heads off to infinity. We already saw this in Part I of the course, in the discussion of asymptotic expansions, when we found that e^z has the asymptotic series expansion $e^z \sim 0$ for z large and negative, whilst no asymptotic expansion exists at all for z large and positive. (Recall the cautionary discussion above about the strict meaning of an asymptotic series, and interpret these observations appropriately within the spirit of those remarks!) The different asymptotic behaviours exhibited by $M(a,b,z)$ for large positive and negative z is much more interesting than the situation for the exponential function, however.

One way of seeing why the upper asymptotic expansion in (2.54) could not possibly be valid for all values of $\arg(z)$ is as follows. We know that $M(a,b,z)$ is analytic in the whole finite complex z plane, and therefore in particular, it must be a single-valued function of z. Thus if we write $z=|z|\,e^{\mathrm{i}\,\theta}$, then we know that if we allow θ to increase by an angle 2π, then the function $M(a,b,z)$ must return to its initial value.

Obviously, for generic values of the parameters a and b, the upper function in (2.54) is *not* single valued, and so if we were to allow θ to increase by 2π we would pick up a phase factor

$$e^{2\pi\,(a-b)\,\mathrm{i}} \ne 1\,, \qquad (\text{when}\quad (a-b)\ne \text{ integer})\,. \tag{2.55}$$

Thus the asymptotic expansion has a behaviour that is totally wrong, if we allow z to be swung around by a full 2π angle. Similar remarks apply to the lower formula in (2.54).

This observation is an example of what is called the *Stokes Phenomenon*, and it is in fact what almost always happens with asymptotic expansions. To see exactly what is going on, we need to do a rather more careful analysis of the asymptotic behaviour of $M(a,b,z)$ not merely for z *real* and large, but for z *complex* and large, of the form $z=|z|\,e^{\mathrm{i}\,\theta}$ with $|z|$ large and the phase θ allowed to take any value. What we shall find is that for θ in a certain range around $\theta = 0$, an appropriate generalisation of the upper asymptotic behaviour in (2.54) occurs, whislt for θ in the rest of the range, around $\theta = \pi$, an appropriate generalisation of the lower asymptotic behaviour in (2.54) occurs. There are certain crossover angles on which both types of asymptotic behaviour have roughly equal importance.

To study the Stokes phenomenon in more detail, we need to repeat the previous analysis, but now for the case where z tends to infinity with some phase θ. In other words, we take $z = e^{\mathrm{i}\,\theta}\,|z|$ and send $|z|$ to infinity, holding the angle θ fixed. We shall consider first the case of angles θ in the range $0 < \theta < \pi$; the reason for placing this restriction in this case will become apparent below. We now use the identity that

$$\int_0^1 dt = \int_0^{-\infty\, e^{-\mathrm{i}\,\theta}} dt - \int_1^{-\infty\, e^{-\mathrm{i}\,\theta}} dt\,. \tag{2.56}$$

Use this in (2.43), with the contours of integration now running with an angle θ relative to the negative real axis. In the first integral, we make the change of variable

$$t = -\frac{w\, e^{-\mathrm{i}\,\theta}}{|z|} = \frac{w\, e^{\mathrm{i}(\pi-\theta)}}{|z|}\,, \tag{2.57}$$

while in the second integral we make the change of variable

$$t = 1 - \frac{u\, e^{-\mathrm{i}\,\theta}}{|z|}\,. \tag{2.58}$$

In each case, to traverse the stated contour we shall have the new integration variable w or u running from 0 to $+\infty$. After simple algebra, we get the following:

$$\begin{aligned} M(a,b,z) \;&=\; \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b-a)}\Big[\frac{e^{\mathrm{i}(\pi-\theta)\,a}}{|z|^a}\int_0^\infty e^{-w}\, w^{a-1}\,\Big(1+\frac{w}{z}\Big)^{b-a-1}\, dw \\ &\qquad + \frac{e^z\, e^{-\mathrm{i}(b-a)\,\theta}}{|z|^{b-a}}\int_0^\infty e^{-u}\, u^{b-a-1}\,\Big(1-\frac{u}{z}\Big)^{a-1}\, du\Big]\,. \end{aligned} \tag{2.59}$$

The integration contours in the complex t-plane are depicted in Figure 11 below.

Since the integrand in (2.43) has branch points at $t = 0$ and $t = 1$, we must establish a convention about where to choose our branch cuts, and then stick with this choice in the subsequent analysis. Specifically, when we decompose the integral in (2.43) into a difference of two integrals as in (2.56), with t running off to infinity somwhere in the complex t-plane, we must establish a convention about where the branch cut running out to infinity will lie. Let us choose the negative real t axis. This means that we must restrict θ to lie in between 0 and π, so that the contours for the two t integrations don't cross over the real t axis and pass through the branch points at $t = 0$ or $t = 1$.

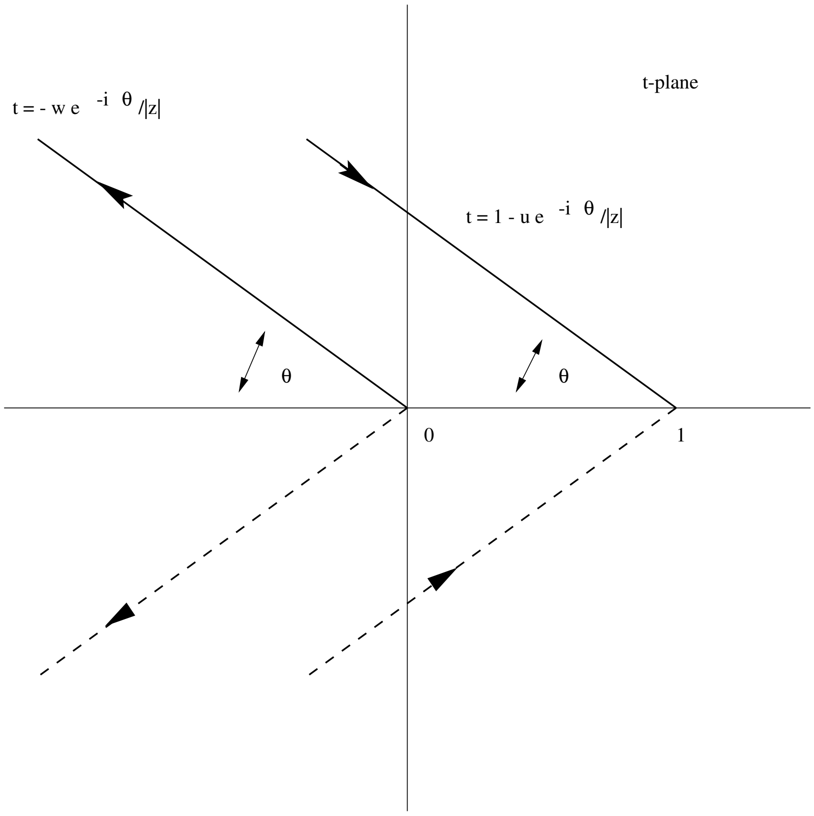

Figure 11: The contours for $0 < \theta < \pi$ (solid lines) and $-\pi < \theta < 0$ (dashed lines)

Eventually, we make binomial expansions of the quantities $(1 + w/z)^{b-a-1}$ and $(1 - u/z)^{a-1}$ in the integrands, to obtain the full asymptotic expansions. First, it is useful to focus just on the leading-order terms, where, for very large $|z|$, we approximate these factors by 1. This can be done for exactly the same reason as we discussed previously, namely that by the time w or u has become large enough that $|w/z|$ or $|u/z|$ cannot be neglected in comparison to 1, the exponential factor will have become so tiny that the error is very small. In fact in the subsequent discussion we can always focus just on the two leading-order terms, with the understanding that each is always to be supplemented by its binomial-expansion descendants.

For the leading-order terms, the integrals that remain to be evaluated then simply give $\Gamma(a)$ and $\Gamma(b-a)$ respectively, and so the leading contributions from each integral give

$$M(a,b,z) \sim \frac{\Gamma(b)}{\Gamma(b-a)}\, |z|^{-a}\, e^{\mathrm{i}(\pi-\theta)\,a} + \frac{\Gamma(b)}{\Gamma(a)}\, |z|^{a-b}\, e^{|z|\,e^{\mathrm{i}\theta}}\, e^{\mathrm{i}(a-c)\,\theta}\,, \tag{2.60}$$

where, it will be recalled, $0 < \theta < \pi$. In fact if z is real and positive we have already obtained the result (2.51), which is precisely (2.60) with $\theta = 0$, bearing in mind that the first term in (2.60) is negligable compared with the second in this case, on account of the latter's $e^{|z|}$

factor. If, on the othere hand, z is real and negative then the previously-obtained expansion (2.53) can be seen to be precisely in agreement with setting $\theta = \pi$ in (2.60), and bearing in mind that now only the first term in (2.60) contributes, on account of the $e^{-|z|}$ factor in the second term.

Suppose instead we now take θ at some intermediate angle $0 < \theta < \pi$. If we take $\theta = \frac{1}{2}\pi$, the exponential factor in the second term now just becomes $e^{\mathrm{i}\,|z|}$, which is a phase factor of unit modulus. At $\theta = \frac{1}{2}\pi$, therefore, the exponential has no damping effect, and the two terms in (2.60) have roughly equal size. Thus both terms, and their binomial-expansion descendants, will be included in the asymptotic expansion at $\theta = \frac{1}{2}\pi$. As θ ranges from 0 to π, the expression (2.60) (and its binomial descendants) therefore gives the correct asymptotic expansion, with the first term disappearing altogether at $\theta = 0$, and the second term disappearing at $\theta = \pi$.

Now let us consider what happens in the region where $-\pi < \theta < 0$, i.e. when z is in the lower-half complex plane. It can be seen from Figure 11 that if we simply allowed θ to pass through 0 and become negative in the previous integral decomposition (2.56), then the integration contours for t would now have swung down below the negative real t-axis, crossing over the branch cut running out to $-\infty$ in the complex t-plane. On the other hand, nothing untoward should happen when we switch over between $\theta = +\pi$ and $\theta = -\pi$, since this corresponds to t running out along the positive real axis, where there is no branch cut. To make sure that this works, we must now take

$$
\begin{aligned}
t &= -\frac{w\, e^{-\mathrm{i}\,\theta}}{|z|} = \frac{w\, e^{-\mathrm{i}(\pi+\theta)}}{|z|}\,, && (2.61)\\
t &= 1 - \frac{u\, e^{-\mathrm{i}\,\theta}}{|z|}\,, && (2.62)
\end{aligned}
$$

for the redefinitions in the two integrations. Note that the first redefinition here differs from the one in (2.57) that we used when $0 < \theta < \pi$. This difference precisely takes account of the need to avoid the branch cut from $t = 0$ to $t = -\infty$. Following through the analogous steps to our previous ones, we now find

$$
M(a,b,z) \sim \frac{\Gamma(b)}{\Gamma(b-a)}\, |z|^{-a}\, e^{-\mathrm{i}(\pi+\theta)\,a} + \frac{\Gamma(b)}{\Gamma(a)}\, |z|^{a-b}\, e^{|z|\, e^{\mathrm{i}\theta}}\, e^{\mathrm{i}(a-c)\,\theta}\,, \qquad (2.63)
$$

for $-\pi < \theta < 0$, replacing (2.60) that was valid for $0 < \theta < \pi$.

Notice that as θ runs from 0 to negative values, the first term here emerges from being insignificant (relative to the second term), and takes over as the dominant term by the time θ is passing through $-\frac{1}{2}\pi$. Now at $\theta = 0$ the first term in (2.63) has a factor $e^{-2\pi\, a\, \mathrm{i}}$ in comparison to the first term in (2.60) at $\theta = 0$. This makes it look as if there would be a discontinuity in the asymptotic expansion of the function $M(a,b,z)$ at $\theta=0$, but actually there isn't. The reason is precisely because the term with the apparent discontinuity is the first term in (2.60) or (2.63), and this term is absent from the asymptotic expansion at $\theta=0$ on the grounds of its insignificance in comparison to the second term. (Morse and Feschbach refer to it as being "in eclipse" at $\theta=0$, which is quite an apt description.)

On the other hand, we can see that the first term in the expansion (2.60) at $\theta=\pi$, where it dominates over the second term, is in precise agreement with the first term in (2.63) at $\theta=-\pi$. This would not have happened if we had not made the replaced the redefinition (2.57) by (2.61). Without the replacement, we would have got an answer at $\theta=-\pi$ that differed from the answer at $\theta=pi$ by a factor of $e^{2\pi\,a\,\mathrm{i}}$. This would have contradicted the fact that $M(a,b,z)$ is analytic, and should therefore not exhibit any branch-point behaviour.

The summary of this rather long and tortuous discussion is the following. The confluent hypergeometric function $M(a,b,z)$ is itself analytic in the finite complex plane, and so in particular it has no branch points. However, the presence of the branch points in the complex t-plane in the integrand of (2.43) means that one has to be careful, when deriving the asymptotic expansion of $M(a,b,z)$, to handle the choice of integration contour carefully. When this is done properly, one finds that the asymptotic expansion can be expressed as a set of results valid in different "patches," corresponding to different ranges for the phase θ of the complex variable z. In each patch the expansion naively appears to suffer from not being single-valued, but actually everything is OK because one is not allowed to let the phase angle θ stray far enough in any particular expansion expression for the lack of single-valuedness in that expression to become evident. The expressions for the asymptotic expansions in each patch join on smoothly and continuously to one another, as one swings θ around to pass from one patch to the next. This is despite the fact that certain terms in two neighbouring patches can appear to have different phase factors (like the $e^{2\pi\,a\,\mathrm{i}}$ factor we encountered above). The point is that such a term is always "in eclipse" at the value of θ where the crossover between the patches occurs, and so the two expressions merely differ by a phase factor that multiplies 0. The bottom line is that one ends up with a set of expression for the asymptotic expansions that correctly describe the large-z behaviour of the single-valued function $M(a,b,z)$.

The situation can be summarised mathematically as follows. The asymptotic expansion of the function $\Gamma(a)\,\Gamma(b-a)\,M(a,b,z)/\Gamma(b)$, with $z=e^{\mathrm{i}\,\theta}\,|z|$ and $|z$ large is given by

$$\begin{aligned}
\theta=-\pi: &\qquad \Gamma(a)\,|z|^{-a}\,,\\
-\pi<\theta<0: &\qquad \Gamma(b-a)\,|z|^{a-b}\,e^{\mathrm{i}(a-b)\,\theta}\,e^{z}+\Gamma(a)\,|z|^{-a}\,e^{-\mathrm{i}\,a\,(\pi+\theta)}\,,
\end{aligned}$$

$$
\begin{aligned}
\theta = 0: &\quad \Gamma(b-a)\,|z|^{a-b}\,e^z\,, \\
0 < \theta < \pi: &\quad \Gamma(b-a)\,|z|^{a-b}\,e^{\mathrm{i}(a-b)\,\theta}\,e^z + \Gamma(a)\,|z|^{-a}\,e^{\mathrm{i}\,a\,(\pi-\theta)}\,, \\
\theta = \pi: &\quad \Gamma(a)\,|z|^{-a}\,, \\
\pi < \theta < 2\pi: &\quad \Gamma(b-a)\,|z|^{a-b}\,e^{\mathrm{i}(a-b)\,(\theta-2\pi)}\,e^z + \Gamma(a)\,|z|^{-a}\,e^{\mathrm{i}\,a\,(\pi-\theta)}\,, \\
\theta = 2\pi: &\quad \Gamma(b-a)\,|z|^{a-b}\,e^z\,,
\end{aligned}
\tag{2.64}
$$

and so on.

3 Integral Transforms and Fourier Series

Integral transforms can provide a very useful technique for constructing the solutions of differential equations. We have in fact already encountered several examples of integral representations for solutions of differential equations, which can be derived by applying the methods of integral transforms. They are also very familiar in other contexts, such as the Fourier transform that has many applications in mathematical physics, for example in quantum mechanics and in wave theory. We shall begin with a general discussion of the use of integral transform methods for solving differential equations.

3.1 Solution of ODEs by Integral Transforms

The general idea of an integral transform is that we write a function $y(z)$ as an integral,

$$
y(z) = \int K(z,t)\,f(t)\,dt\,, \tag{3.1}
$$

where $K(z,t)$ is called the *Kernel Function.* $y(t)$ is said to be the *integral transform* of the function $f(t)$. For now, we shall leave the range of the integration over t unspecified; the choice for the integration range depends upon the details of the problem. It might sometimes be a real integral between specified limits, or it might instead be a contour integral in the complex t-plane.

Let us begin with an example, to illustrate the basic idea and utility of an integral transform. Suppose we wish to solve the second-order ODE

$$
z\,y'' + (b-z)\,y' - a\,y = 0\,. \tag{3.2}
$$

This will be recognised as the confluent hypergeometric equation, which we encountered in the previous chapter. A rather significant feature of this equation is that it is, of course, of second order in z derivatives, but the coefficients involve explicit powers of z only up to the power 1. For reasons that will emerge in a moment, this means that it is useful to write $y(z)$ as an integral transform of the form (3.1), with the kernel function $K(z,t)$ chosen to be

$$K(z,t) = e^{z\,t}\,. \tag{3.3}$$

This, of course, has the property that

$$\frac{d}{dz}\,e^{z\,t} = t\,e^{z\,t}\,, \qquad \frac{d^2}{dz^2}\,e^{z\,t} = t^2\,e^{z\,t}\,, \tag{3.4}$$

*etc.*The transformation (3.1) with a kernel of this exponential type is known as the *Laplace Transform.*

Substituting (3.1) into the differential equation (3.2), we therefore obtain

$$\int f(t)\,\Big(z\,t^2 + (b-z)\,t - a\Big)\,e^{z\,t}\,.dt = 0 \tag{3.5}$$

Now of course the kernel $e^{z\,t}$ also has the property that

$$z\,e^{z\,t} = \frac{d}{dt}\,e^{z\,t}\,, \tag{3.6}$$

which is in some sense "dual" to (3.4). Thus we can write (3.5) as

$$\int f(t)\,\Big(t^2\frac{d}{dt} + b\,t - t\,\frac{d}{dt} - a\Big)\,e^{z\,t}\,dt = 0\,, \tag{3.7}$$

and so after an integration by parts we get

$$\int \Big(t(t-1)\,\dot f(t) + (2-b)\,t\,f(t) + (a-1)\,f(t)\Big)\,e^{z\,t}\,dt = 0\,, \tag{3.8}$$

where we use a dot to denote a derivative with respect to t. We have assumed here that the boundary term from the integration by parts gives zero. This is up to us to arrange, by making a suitable choice of limits or contour for the integration.

As we shall see later, for suitable choices of kernel function $K(z,t)$, such as $e^{z\,t}$, the transform (3.1) is *invertible*, in the sense that for every admissable $y(z)$ there is a unique function $f(t)$ that produces it. In particular, the function that produces 0 must itself be 0. We may therefore conclude from (3.8) that the integrand is zero, and so in other words

$$t(t-1)\,\dot f(t) + (2-b)\,t\,f(t) + (a-1)\,f(t) = 0\,. \tag{3.9}$$

This differential equation in the transform variable t, is, luckily, much easier to solve than the original equation (3.2). In particular, it is only of first order in derivatives, unlike the original equation, which was of second order. The reason for this is precisely because of the fact that we drew attention to earlier, namely that the original equation (3.2) only involved z to the powers 0 and 1 in the coefficients of $y(z)$, $y'(z)$ and $y''(z)$. The "dual" relation between (3.4) and (3.6) for the kernel function $e^{z\,t}$ means that each derivative in the original equation becomes a multiplication by t inn the transformed equation, and *vice versa*. (Notice that (3.9) has t to the powers 0, 1 and 2 in its coefficients of $f(t)$ and $\dot{f}(t)$.)

The transformation to the first-order differential equation (3.9) has in fact given us an equation that can be solved very easily, namely

$$\frac{\dot{f}}{f} = \frac{a-1}{t} - \frac{b-a-1}{1-t}\,, \tag{3.10}$$

whose solution is

$$f = t^{a-1}\,(1-t)^{b-a-1}\,. \tag{3.11}$$

Thus we conclude that the solution of the confluent hypergeometric equation (3.2) is given by

$$y(z) = \int t^{a-1}\,(1-t)^{b-a-1}\,e^{z\,t}\,dt\,. \tag{3.12}$$

We have, essentially, reproduced the integral representation (2.43) of the previous chapter, which gave us the regular Kummer function $M(a,b,z)$. Actually, we have produced something a little more general here, since we have not yet specified any particular choice for the integration limits. In the integral representation (2.43) for $M(a,b,z)$ the integral was taken from $t=0$ to $t=1$, and indeed one can easily verify that the boundary term that we dropped in getting from (3.7) to (3.8) vanishes at these endpoints. In fact, the boundary term is

$$\Big[e^{z\,t}\,t^a\,(1-t)^{b-a}\Big]\,, \tag{3.13}$$

which indeed vanishes at $t=0$ and $t=1$, provided that $b>a>0$.

There are other ways of arranging for the boundary term (3.13) to vanish, instead of taking the integration limits to be 0 and 1. For example, we could take them to be 1 and ∞, provided that the real part of z is negative, and that $b>a$. The freedom to choose different possibilities for the contour of integration reflects the fact that the original differential equation (3.2) has two independent solutions. By making an appropriate choice, we can get the second solution $U(a,b,z)$, Kummer's irregular function. We encountered examples also in Chapter 1, where a different choice of contour gave a different and linearly-independent solution of the differential equation, in the context of the Bessel functions. Namely, we saw that the integral representation (1.29) produced the $J_\nu(z)$ Bessel function for one choice of contour, but it produced instead $H_\nu^{(1)}(z)$ or $H_\nu^{(2)}(z)$ for different choices of contour.

The integral transformation with the kernel $e^{z\,t}$ was particularly nice in the example of the confluent hypergeometric equation because of the fact that the coefficients in front of $y(z)$, $y'(z)$ and $y''(z)$ in (3.2) involve only the zero'th and first powers of z, implying that the transformed differential equation (3.9) is only a first-order equation. Sometimes, a differential equation may have higher powers of z that can be removed by making appropriate changes of the dependent and independent variables. The Bessel equation is an example of this type, as is the modified Bessel equation,

$$z^2\,y''(z) + z\,y'(z) - (\nu^2 + z^2)\,y(z) = 0\,. \tag{3.14}$$

Taken as it stands, this would give us a second-order differential equation for $f(t)$ after making the transformation (3.1) with $K(z,t) = e^{z\,t}$. However, it is easy to see that if we let

$$y(z) = z^\nu\, e^{-z}\, w(z)\,, \tag{3.15}$$

and then let $z = \frac{1}{2}\tilde{z}$, the modified Bessel equation becomes

$$\frac{d^2 w}{d\tilde{z}^2} + (2\nu + 1 - \tilde{z})\,\frac{dw}{d\tilde{z}} - (n + \tfrac{1}{2})\,w = 0\,. \tag{3.16}$$

This is just the confluent hypergeometric equation (3.2), with $a = \nu + \frac{1}{2}$ and $b = 2\nu + 1$. Indeed, this makes explicit the way in which the Bessel functions and modified Bessel functions arise as special cases of the confluent hypergeometric functions.

There are other examples, of course, where one cannot reduce the coefficients of the $y''(z)$, $y'(z)$ and $y(z)$ terms to constants and linear powers, no matter how hard one tries with changes of variable. It may well happen, therefore, that the transformed equation is "worse" then the original one. On the other hand, it may be that by making a different choice for the kernel function $K(z,t)$, the situation might like better. In fact the kernel $K(z,t) = e^{z\,t}$ is the suitable one when dealing with an equation with one regular singular point and one irregular singular point of a certain particular kind. Specifically, this kernel works well in the case of the confluent hypergeometric equation, which has an irregular singular point that comes from the confluence of two regular singular points. In fact, we obtained the equation by taking a limit of the hypergeometric equation, in which its regular singular points at $z = 1$ and $z = \infty$ fused together.

To transform the hypergeometric equation

$$z(1-z)\,y''(z) + [c - (a + b + 1)\,z]\,y'(z) - a\,b\,y(z) = 0 \tag{3.17}$$

into a nice form, a different kernel, namely $K(z,t) = (z - t)^\mu$, is appropriate, where μ is a constant that we shall choose for convenience. An integral transform using a kernel of this type is known as an *Euler Transform.* Thus if we transform $y(z)$ according to

$$y(z)=\int(z-t)^{\mu}\,f(t)\,dt, \tag{3.18}$$

then substituting into (3.17) we get, after collecting powers of z,

$$\int(z-t)^{\mu-2}\,\Big[(\mu+a)(\mu+b)\,z^2-[\mu(\mu+c-1)+(2ab+\mu(a+b+1))\,t]\,z+(\mu\,c+a\,b\,t)\,t\Big]\,f(t)\,dt=0\,. \tag{3.19}$$

Now recall that we are free to choose the constant μ at will. By choosing $\mu=-a$ or $\mu=-b$, the term in z^2 in the large square brackets in (3.19) will disappear. The two choices are equivalent, so let us, w.o.l.o.g., choose $\mu=-a$. The integral (3.19) now becomes

$$\int\Big[(z-t)^{-a-1}\,[c-b\,t+(a+1)(t-1)]+(a+1)\,t\,(t-1)(z-t)^{-a-2}\Big]\,f(t)\,dt=0\,. \tag{3.20}$$

Observe that we can write the last factor in the large square brackets as

$$(a+1)\,t\,(t-1)(z-t)^{-a-2}=t\,(t-1)\,\frac{d}{dt}\,(1-z\,t)^{-a-1}\,, \tag{3.21}$$

giving us

$$\int\Big[(z-t)^{-a-1}\,[c-b\,t+(a+1)(t-1)]+t\,(t-1)\,\frac{d}{dt}\,(z-t)^{-a-1}\Big]\,f(t)\,dt=0\,. \tag{3.22}$$

Integrating by parts, and invoking the expected uniquness of transform, we then deduce that $f(t)$ must satify the first-order differential equation

$$t\,(t-1)\,\dot f(t)-[c-a+(a-b-1)\,t]\,f(t)=0\,. \tag{3.23}$$

It is easy to solve this, to obtain $f(t)=t^{a-c}\,(t-1)^{c-b-1}$, and hence we learn that the solution of the hypergeometric equation is given by

$$y(z)=\int(t-1)^{c-b-1}\,t^{a-c}\,(z-t)^{-a}\,dt\,. \tag{3.24}$$

This is very like the integral representation for ${}_2F_1(a,b;c;z)$ that we encountered in the previous chapter, in equation (2.20); in fact if we send t to $1/t$ in (3.24), then up to an unimportant constant factor we recover the integral representation in (2.20). As usual, we must choose the contour of integration such that the boundary terms arising from the integration by parts give zero. From (3.22), and the solution for $f(t)$, this means that

$$\Big[t^{a-c-1}\,(t-1)^{c-b}\,(z-t)^{-a-1}\Big] \tag{3.25}$$

should vanish when evaluated between the integration limits. One possible choice, provided that $\mathrm{Re}(c)>\mathrm{Re}(b)>0$, is to take t to run from $t=1$ to $t=\infty$. This is precisely equivalent to the integration range used in (2.20), bearing in mind the inversion $t \longrightarrow 1/t$ between (2.20) and (3.24).

We have now seen two examples of integral transforms, one using the kernel $K(z,t) = e^{z\,t}$, for solving the conluent hypergeometric equation, and the other using the kernel $K(z,t) = (z-t)^\mu$, for solving the hypergeometric equation. In each case the kernel has nice "reciprocal" properties, in that derivatives with respect to z and with respect to t bear some nice relation to one another. To complete this part of the discussion, let us consider the procedure in a more general setting, leaving the choice of kernel in the integral transform (3.1) unspecified.

Suppose we wish to solve the second-order ODE (ordinary differential equation)

$$\mathcal{L}_z[y(z)] \equiv p_0(z)\, y''(z) + p_1(z)\, y'(z) + p_2(z)\, y(z) = 0\,. \tag{3.26}$$

The subscript z on the differential operator $\mathcal{L}_z$ defined by this equation indicates that the derivatives are with respect to z:

$$\mathcal{L}_z = p_0(z)\, \frac{d^2}{dz^2} + p_1(z)\, \frac{d}{dz} + p_2(z)\,. \tag{3.27}$$

Acting with this operator on the integral transform (3.1), we can take the differential operator inside the integration, provided that the integral is suitably convergent, to give then gives

$$\mathcal{L}_z[y(z)] = \int \mathcal{L}_z[K(z,t)]\, f(t)\, dt\,. \tag{3.28}$$

If the kernel $K(z,t)$ has been chosen appropriately, the quantity $\mathcal{L}_z[K(z,t)]$ can be re-expressed as a different differential operator $\mathcal{M}_t$ acting on some other function $\widetilde{K}(z,t)$, this time with the derivatives being with respect to t instead of z:

$$\mathcal{L}_z[K(z,t)] = \mathcal{M}_t[\widetilde{K}(z,t)]\,. \tag{3.29}$$

Sometimes it may be the case that $\widetilde{K}(z,t)$ is actually the same function as $K(z,t)$ itself.

As an example, recall our integral transform of the hypergeometric equation, where we used $K(z,t) = (z-t)^{-a}$. From (3.17) and (3.22), it will be seen that $\widetilde{K}(z,t) = (z-t)^{-a-1}$, with

$$\begin{aligned} \mathcal{L}_z &= z(1-z)\, \frac{d^2}{dz^2} + [c - (a+b+1)\, z]\, \frac{d}{dz} - a\, b\,, \\ \mathcal{M}_t &= t(t-1)\, \frac{d}{dt} + c - b\, t + (a+1)(t-1)\,. \end{aligned} \tag{3.30}$$

On the other hand, in the example of the confluent hypergeometric equation, where the kernel was $K(z,t) = e^{z\,t}$, we see from (3.2) and (3.7) that in this case we have $\widetilde{K}(z,t) =$

$e^{z\,t} = K(z,t)$, and

$$\mathcal{L}_z = z\,\frac{d^2}{dz^2} + (b-z)\,\frac{d}{dz} - a\,,$$
$$\mathcal{M}_t = t(t-1)\,\frac{d}{dt} + b\,t - a\,. \tag{3.31}$$

More generally, let us suppose that with a choice of kernel function $K(z,t)$ that is appropriately "matched" to the differential operator (3.27) for the specific functions $p_0(z)$, $p_1(z)$ and $p_2(z)$ in question, there is some differential operator $\mathcal{M}_t$ such that (3.29) is satisfied, where $\mathcal{M}_t$ has the form[9]

$$\mathcal{M}_t = \alpha_0(t)\,\frac{d^2}{dt^2} + \alpha_1(t)\,\frac{d}{dt} + \alpha_2(t)\,. \tag{3.32}$$

The idea now is that after acting on (3.1) with the differential operator $\mathcal{L}_z$, we use (3.29) and then integrate by parts to move the t derivatives off $\widetilde{K}(z,t)$ and onto $f(t)$:

$$\begin{aligned}
\mathcal{L}_z[y(z)] &= \int \mathcal{L}_z[K(z,t)]\,f(t)\,dt \\
&= \int \mathcal{M}_t[\widetilde{K}(z,t)]\,f(t)\,dt \\
&= \int \Big(\alpha_0(t)\,f(t)\,\frac{d^2\widetilde{K}(z,t)}{dt^2} + \alpha_1(t)\,f(t)\,\frac{d\widetilde{K}(z,t)}{dt} + \alpha_2(t)\,f(t)\,\widetilde{K}(z,t)\Big)\,dt \\
&= \int \Big(- \frac{d(\alpha_0(t)\,f(t))}{dt}\,\frac{d\widetilde{K}(z,t)}{dt} - \frac{d(\alpha_1(t)\,f(t))}{dt}\,\widetilde{K}(z,t) + \alpha_2(t)\,f(t)\,\widetilde{K}(z,t) \\
&\quad + \frac{d}{dt}\Big[\alpha_0(t)\,f(t)\,\frac{d\widetilde{K}(z,t)}{dt} + \alpha_1(t)\,f(t)\,\widetilde{K}(z,t)\Big]\Big)\,dt \\
&= \int \Big(\Big[\frac{d^2(\alpha_0(t)\,f(t))}{dt^2} - (\frac{d(\alpha_1(t)\,f(t))}{dt} + \alpha_2(t)\,f(t)\Big]\,\widetilde{K}(z,t) \\
&\quad + \frac{d}{dt}\Big[\alpha_0(t)\,f(t)\,\frac{d\widetilde{K}(z,t)}{dt} - \widetilde{K}(z,t)\,\frac{d(\alpha_0\,f(t)}{dt} + \alpha_1(t)\,f(t)\,\widetilde{K}(z,t)\Big]\Big)\,dt\,.
\end{aligned} \tag{3.33}$$

We may write this as

$$\begin{aligned}
\mathcal{L}_z[y(z)] &= \int \Big(\widetilde{K}(z,t)\,\overline{\mathcal{M}}_t[f(t)] + \frac{dP(f,\widetilde{K})}{dt}\Big)\,dt\,, \\
&= \int \widetilde{K}(z,t)\,\overline{\mathcal{M}}_t[f(t)]\,dt + \Big[P(f,\widetilde{K})\Big]\,,
\end{aligned} \tag{3.34}$$

[9]We are assuming here that the operator $\mathcal{M}_t$ is of at most second order in derivatives. This, of course, is not guaranteed; it all depends on the details of the original differential operator $\mathcal{L}_z$, and on one's choice of kernel function $K(z,t)$. In practice, it is unlikely that we would want to use this method for solving the differential equation if the transformed equation turned out to be of higher order in derivatives than the original one. Since we are assuming that we start with a second-order differential operator $\mathcal{L}_z$, then we may restrict our discussion to those cases where $\mathcal{M}_t$ involves no higher than second derivatives also. The extension to higher-order operators is totally straightforward.

where $\overline{\mathcal{M}}_t$ is the *adjoint* of the operator $\mathcal{M}_t$, and $P(f,\widetilde{K})$ is the the *bilinear concomitant* of $f(t)$ and $\widetilde{K}(z,t)$:

$$\begin{aligned}
\overline{\mathcal{M}}_t[f(t)] &\equiv \frac{d^2}{dt^2}\,(\alpha_0(t)\,f(t)) - \frac{d}{dt}\,(\alpha_1(t)\,f(t)) + \alpha_2(t)\,f(t)\,, &(3.35)\\
P(f,\widetilde{K}) &\equiv \alpha_0(t)\,f(t)\,\frac{d\widetilde{K}(z,t)}{dt} - \widetilde{K}(z,t)\,\frac{d(\alpha_0\,f(t)}{dt} + \alpha_1(t)\,f(t)\,\widetilde{K}(z,t)\,. &(3.36)
\end{aligned}$$

The square brackets enclosing $P(f,\widetilde{K})$ in the second line indicate that it is to be evaluated at the endpoints of the integration.

Now, we make the usual kind of argument that we shall choose a contour for the integration in (3.1) such that the bilinear concomitant $P(f,\widetilde{K})$ returns to its initial value at the end of the contour, so that the boundary term $[P(f,\widetilde{K})]$ in (3.34) is zero, and so we simply have

$$\mathcal{L}_z[y(z)] = \int \widetilde{K}(z,t)\,\overline{\mathcal{M}}_t[f(t)]\,dt\,. \tag{3.37}$$

Thus we conclude that $y(z)$ defined by (3.1) satisfies the original differential equation $\mathcal{L}_z[y(z)] = 0$ if the function $f(t)$ satisfies the differential equation $\overline{\mathcal{M}}_t[f(t)] = 0$. Of course the hope is that we have made a fortunate choice for $K(z,t)$ so that the transformed equation is easier to solve than the original one.

In our example of the hypergeometric equation, we see from (3.22), (3.35) and (3.36) that in this case we shall have

$$\begin{aligned}
\overline{\mathcal{M}}_t[f(t)] &= -\frac{d}{dt}\Big(t(t-1)\,f(t)\Big) + \Big(c - b\,t + (a+1)(t-1)\Big)\,f(t)\,,\\
P(f,\widetilde{K}) &= t(t-1)\,f(t)\,(z-t)^{-a-1}\,. &(3.38)
\end{aligned}$$

On the other hand, for the example of the confluent hypergeometric equation, it follows from (3.7), (3.35) and (3.36) that in this case

$$\begin{aligned}
\overline{\mathcal{M}}_t[f(t)] &= -\frac{d}{dt}\Big(t(t-1)\,f(t)\Big) + (b\,t - a)\,f(t)\,,\\
P(f,\widetilde{K}) &= t(t-1)\,f(t)\,e^{z\,t}\,. &(3.39)
\end{aligned}$$

Both these examples are rather simpler than the general discussion, because the differential operator $\mathcal{M}_t$ is only of first order in derivatives, and so $\alpha_0(t) = 0$.

3.2 The Fourier Transform

We concluded the previous subsection by considering the general case of an integral transform (3.1) where the kernel function $K(z,t)$ is unspecified. We also looked at specific examples, for which we had $K(z,t) = e^{z\,t}$ and $K(z,t) = (z-t)^\mu$. The integral transform is called the Laplace transform when $K(z,t) = e^{z\,t}$, and the Euler transform when $K(z,t) = (z-t)^\mu$.

In practice, there is a rather small number of different kernels that turn out to be useful, and most of these are closely related to the Fourier transform. The Fourier transform is the name given to the case where one uses $K(z,t) = e^{\mathrm{i}\,z\,t}$ as the kernel function. Its relation to the Laplace transform $K(z,t) = e^{z\,t}$ is obvious. We shall now proceed with a more detailed study of the Fourier transform, since it is one that is used extensively in mathematical physics.

First, let us establish some notation. We shall define the Fourier transform $F(k)$ of a function $f(x)$ as follows:

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\mathrm{i}\,k\,x}\, f(x)\, dx\,. \tag{3.40}$$

The need for 2π factors somewhere in the discussion is inevitable, and stems from the inconvenient fact that a unit circle has circumference 2π rather than 1. Putting in a $\sqrt{2\pi}$ in the definition of the Fourier transform gives the symmetrical result that the inverse Fourier transform is

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\mathrm{i}\,k\,x}\, F(k)\, dk\,. \tag{3.41}$$

The fact that this is the inverse of the Fourier transform (3.40) is a non-trivial result, known as *Fourier's Theorem.* We can prove it by viewing the Fourier transform as the limit of a Fourier series. Before doing this, note that be substituting (3.40) into (3.41), we have an equivalent statement of Fourier's theorem, namely that

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \int_{-\infty}^{\infty} dy\, e^{\mathrm{i}\,k\,(y-x)}\, f(y)\,. \tag{3.42}$$

Yet another way of expressing this is that since this is true for any (reasonable) function $f(x)$, it must be that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{\mathrm{i}\,k\,(y-x)} = \delta(y-x)\,, \tag{3.43}$$

where $\delta(y-x)$ is the Dirac delta function, with the property that

$$f(x) = \int_{-\infty}^{\infty} f(y)\, \delta(y-x)\, dy\,, \tag{3.44}$$

for any (reasonable) function $f(x)$. We shall postpone for now the issue of defining exactly what constitutes a "reasonable" function. We shall return to this later, when we discuss the topic of *Generalised Functions*, of which the Dirac delta function is an example.[10] Note that by replacing the integration variable k by $-k$ in (3.43), we immediately see that the Dirac delta function is symmetrical:

$$\delta(y-x) = \delta(x-y)\,. \tag{3.45}$$

Now for the proof of Fourier's theorem. First, consider the Fourier series for functions $f(x)$ defined on the interval $-\frac{1}{2}b \le x \le \frac{1}{2}b$. It is much simpler to work with the Fourier series using complex exponentials, rather than dealing separately with sines and cosines, so we shall consider the following expansion:

$$f(x) = \sum_{n=-\infty}^{\infty} a_n\, e^{2\pi\, \mathrm{i}\, n\, x/b}\,. \tag{3.46}$$

Note that all the functions $e^{2\pi\, \mathrm{i}\, n\, x/b}$ used in this expansion indeed have the property of returning to their original values after x is advanced through a distance b, since every term in the series has this property. The Fourier coefficients a_n can be determined by multiplying (3.46) by $e^{-2\pi\, \mathrm{i}\, m\, x/b}$, and integrating over the interval $-b/2 \le x \le b/2$. Since we have

$$\int_{-b/2}^{b/2} e^{2\pi\, \mathrm{i}\,(n-m)\, x/b}\, dx = \Big[\frac{b\, e^{2\pi\, \mathrm{i}\,(n-m)\, x/b}}{2\pi\, \mathrm{i}\,(n-m)}\Big]_{-b/2}^{b/2} = 0 \tag{3.47}$$

when $m \ne m$, while it gives

$$\int_{-b/2}^{b/2} dx = b \tag{3.48}$$

when $m = n$, this implies that

$$\int_{-b/2}^{b/2} f(x)\, e^{-2\pi\, \mathrm{i}\, m\, x/b}\, dx = b\, a_m\,. \tag{3.49}$$

Substituting back into (3.46) then gives[11]

$$f(x) = \frac{1}{b} \sum_{n=-\infty}^{\infty} \int_{-b/2}^{b/2} f(y)\, e^{2\pi\, \mathrm{i}\, n\,(x-y)/b}\, dy\,. \tag{3.50}$$

We want to consider the limit where the interval b is sent to infinity. To do this, we introduce a continuous variable k which at discrete points k_n takes the values $k_n = 2\pi\, n/b$.

[10] Mathematicians grumbled at first when Dirac introduced the delta function, maintaining that it wasn't well-defined. Later, they introduced the notion of generalised functions, and made it respectable. So instead of the mathematicians' eyes glazing over when the physicists make dubious manipulations with ill-defined functions, now the physicists' eyes glaze over when the mathematicians make them rigorous in excrutiating detail.

[11] There are some interesting subtleties in the theory of Fourier series, associated with what is known as the *Gibbs Phenomenon*. We shall return to look at this later.

The difference between adjacent points is $\Delta k \equiv k_{n+1} - k_n = 2\pi/b$. We can rewrite (3.50) as

$$f(x) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \Delta k \int_{-b/2}^{b/2} f(y)\, e^{\mathrm{i}\, k_n\,(x-y)}\, dy\,. \tag{3.51}$$

Now, as we take $b \longrightarrow \infty$, the interval Δk between adjacent values of k_n goes to zero, and the sum is replaced by an integral:

$$\sum_{n=-\infty}^{\infty} \Delta k \longrightarrow \int_{-\infty}^{\infty} dk\,. \tag{3.52}$$

Thus (3.51) becomes

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \int_{-\infty}^{\infty} f(y)\, e^{\mathrm{i}\, k\,(x-y)}\, dy\,. \tag{3.53}$$

This is precisely equivalent to (3.42) (send k to $-k$ to get exactly (3.42)), and so Fourier's theorem is proven.

One can easily prove some general properties of the Fourier transform. Trivially obvious ones are that the Fourier transform is a linear operator acting on f to give F. Let us denote the operation of taking the Fourier transform by $\mathcal{L}_{\mathbf{F}}$ (where the subscript F here stands for Fourier), so that we have $\mathcal{L}_{\mathbf{F}}[f] = F$, $\mathcal{L}_{\mathbf{F}}[g] = G$, *etc.* Then the linearity implies

$$\begin{aligned} \mathcal{L}_{\mathbf{F}}[f+g] &= \mathcal{L}_{\mathbf{F}}[f] + \mathcal{L}_{\mathbf{F}}[g]\,, \\ \mathcal{L}_{\mathbf{F}}[a\,f] &= a\,\mathcal{L}_{\mathbf{F}}[f]\,, \end{aligned} \tag{3.54}$$

where in the second line the quantity a is an arbitrary constant. Another general property is that the Fourier transform of the derivative of a function is equal to $-\mathrm{i}\,k$ times the Fourier transform of the function itself:

$$\mathcal{L}_{\mathbf{F}}[f'(x)] = -\mathrm{i}\,k\,\mathcal{L}_{\mathbf{F}}[f(x)] = -\mathrm{i}\,k\,F(k)\,. \tag{3.55}$$

This is easily proved by writing down the Fourier transform of $f(x)$ and then integrating by parts to push the derivative onto the exponential $e^{\mathrm{i}\,k\,x}$. The assumption that the function $f(x)$ is a "reasonable" one justifies the neglect of the boundary terms at $x = \pm\infty$ that arise from the integration by parts.

Parseval's Theorem:

A useful result that can be proven from the definition (3.40) of the Fourier transform is the following, known as *Parseval's Theorem*:

$$\int_{-\infty}^{\infty} |F(k)|^2\, dk = \int_{-\infty}^{\infty} |f(x)|^2\, dx\,. \tag{3.56}$$

To show this, we substitute from (3.40) into the left-hand side, interchange the orders of integration, and then use the expression (3.43) for the Dirac delta function:

$$\begin{aligned}
\int_{-\infty}^{\infty} |F(k)|^2\, dk &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk \int_{-\infty}^{\infty} dx\, e^{\mathrm{i}\,k\,x}\, f(x) \int_{-\infty}^{\infty} dy\, e^{-\mathrm{i}\,k\,y}\, \overline{f(y)}\,,\\
&= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x)\, \overline{f(y)}\, \Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{\mathrm{i}\,k\,(x-y)}\Big)\,,\\
&= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x)\, \overline{f(y)}\, \delta(x-y)\\
&= \int_{-\infty}^{\infty} f(x)\, \overline{f(x)}\, dx\\
&= \int_{-\infty}^{\infty} |f(x)|^2\, dx\,.
\end{aligned} \tag{3.57}$$

(As usual, a more careful discussion could be given in which the circumstances where the interchange of the orders of integration are determined. In practice, it is valid for all "reasonable" functions $f(x)$.)

A small generalisation of Parseval's theorem can be obtained by replacing the function $f(x)$ by $f(x)+g(x)$. Of course since the Fourier transform (3.40) is a *linear* operation on $f(x)$, it trivially follows that the Fourier transform of $f(x)+g(x)$ is $F(k)+G(k)$, where $F(k)$ and $G(k)$ are the Fourier transforms of $f(x)$ and $g(x)$ respectively. Thus we immediately have from Parseval's theorem (3.56) that

$$\int_{-\infty}^{\infty} |F(k)+G(k)|^2\, dk = \int_{-\infty}^{\infty} |f(x)+g(x)|^2\, dx\,. \tag{3.58}$$

Expanding this out, we get

$$\begin{aligned}
&\int_{-\infty}^{\infty} \Big(|F(k)|^2 + |G(k)|^2 + F(k)\,\overline{G(k)} + \overline{F(k)}\,G(k)\Big)\, dk\\
&= \int_{-\infty}^{\infty} \Big(|f(x)|^2 + |g(x)|^2 + f(x)\,\overline{g(x)} + \overline{f(x)}\,g(x)\Big)\, dx\,.
\end{aligned} \tag{3.59}$$

Using the original statement (3.56) of Parseval's theorem, we see that the first terms on each side are equal, as are the second terms on each side, and so

$$\int_{-\infty}^{\infty} \Big(F(k)\,\overline{G(k)} + \overline{F(k)}\,G(k)\Big)\, dk = \int_{-\infty}^{\infty} \Big(f(x)\,\overline{g(x)} + \overline{f(x)}\,g(x)\Big)\, dx\,. \tag{3.60}$$

If instead we were to replace $f(x)$ by $f(x)+\mathrm{i}\,g(x)$ in (3.56), we would, by a similar argument, have that

$$\int_{-\infty}^{\infty} \Big(F(k)\,\overline{G(k)} - \overline{F(k)}\,G(k)\Big)\, dk = \int_{-\infty}^{\infty} \Big(f(x)\,\overline{g(x)} - \overline{f(x)}\,g(x)\Big)\, dx\,. \tag{3.61}$$

Combining these two results, we arrive at the conclusion that

$$\int_{-\infty}^{\infty} F(k)\,\overline{G(k)}\, dk = \int_{-\infty}^{\infty} f(x)\,\overline{g(x)}\, dx\,. \tag{3.62}$$

The Convolution Integral:

Another useful property of the Fourier transform involves the following integral:

$$h(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, f(y)\, g(x-y)\,, \tag{3.63}$$

which is called the *convolution* of f and g. It is also sometimes known as the *Faltung* of f and g, from the German for "folding." (It is a kind of shifted overlap between $f(x)$ and $g(-x)$.) If the functions $f(x)$, $g(x)$ and $h(x)$ have Fourier transforms $F(k)$, $G(k)$ and $H(k)$ respectively, then we can show that

$$H(k) = F(k)\, G(k)\,. \tag{3.64}$$

This is easily proven, by multiplying (3.63) by $1/(\sqrt{2\pi})\, e^{\mathrm{i}\, k\, x}$ and integrating over all x. This gives

$$H(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dy\, f(y) \int_{-\infty}^{\infty} dx\, g(x-y)\, e^{\mathrm{i}\, k\, x}\,. \tag{3.65}$$

Now change integration variable from x to $z = x - y$ in the second integral here, giving

$$H(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dy\, f(y)\, e^{\mathrm{i}\, k\, y} \int_{-\infty}^{\infty} dz\, g(z)\, e^{\mathrm{i}\, k\, z}\,, \tag{3.66}$$

and hence (3.64).

Note that the expression (3.63) is actually symmetrical between f and g, as may be seen by changing the integration variable from y to $z = x - y$. Of course this symmetry is even more obvious in the Fourier-transformed version (3.64).

Fourier Transforms and Quantum Mechanics:

The Fourier transform can be viewed as a mapping between position space and momentum space representations in quantum mechanics. Consider first wavefunction ψ_p in one spatial dimension that is an eigenstate of the momentum operator, with eigenvalue p: $\psi_p(x) = 1/(\sqrt{2\pi})\, e^{\mathrm{i}\, p\, x/\hbar}$. Defining the wave-vector $k = p/\hbar$, this is

$$\psi_k(x) = \frac{1}{\sqrt{2\pi}}\, e^{\mathrm{i}\, k\, x}\,. \tag{3.67}$$

We shall refer to k simply as the momentum, since up to an irrelevant constant factor, that's what it is.[12] To map into momentum space, we take the inverse Fourier transform of

[12]In high-energy physics one usually takes the bull by the horns and chooses units where $\hbar = 1$, which saves a lot of tedious writing. The same is done for the speed of light, and for Newton's constant, so that one works in dimensionless units where $\hbar = c = G = 1$. For mysterious reasons, people in other disciplines apparently prefer to carry around the redundant baggage of superfluous dimensionful constants. There is no physics contained in these; it is merely a reflection of one's decision to measure, for example, distance in metres, while time is measured in seconds, rather than "the time taken for light to travel a certain number of metres."

$\psi_{k_0}(x)$, obtaining

$$\begin{aligned}\Psi(k) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\psi_{k_0}(x)\,e^{-\mathrm{i}\,k\,x}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}\,(k_0-k)\,x}\,dx\\ &= \delta(k-k_0)\,,\end{aligned}\tag{3.68}$$

where in the final step we have used the definition (3.43) of the Dirac delta function. Note that the rôles of k and x are reversed here, relative to our definition of the Fourier transform (3.40) and the inverse transform (3.41). (This is a minor inconvenience in the notation, resulting from the fact that we conventionally give a positive-frequency wave a time dependence $e^{-\mathrm{i}\,\omega\,t}$, which implies that a positive-momentum wave has x dependence $e^{\mathrm{i}\,k\,x}$. This does not mesh ideally with the conventional choice of $e^{\mathrm{i}\,k\,x}$ as the kernel in the Fourier transform (3.40). C'est la vie!) There should be no confusion on this point, but just to clarify our conventions, let us emphasise that we shall always refer to an integral of the form $1/(\sqrt{2\pi})\int\gamma(\xi)\,e^{\mathrm{i}\,\xi\,\zeta}\,d\xi$ as a Fourier transform, and an integral of the form $1/(\sqrt{2\pi})\int\gamma(\xi)\,e^{-\mathrm{i}\,\xi\,\zeta}\,d\xi$ as an inverse Fourier transform, regardless of the names that we happen to be using for the variables.

More generally, if a wave function $\psi(x)$ in position space is a superposition of momentum eigenstates, then it has an equivalent representation $\Psi(k)$ in momentum space, given by

$$\Psi(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\psi(x)\,e^{-\mathrm{i}\,k\,x}\,dx\,.\tag{3.69}$$

The inverse of this, by Fourier's theorem, is

$$\psi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Psi(k)\,e^{\mathrm{i}\,k\,x}\,dk\,.\tag{3.70}$$

One can view this as the continuous limit of a sum over momentum eigenstates, and the function $\Psi(k)$ has the imterpretation of being the "amplitude" of the momentum eigenstate $e^{\mathrm{i}\,k\,x}$ in the sum. The derivative operator d/dx in position space therefore becomes simply a multiplication by $\mathrm{i}\,k$ in momentum space:

$$\frac{d\psi(x)}{dx} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mathrm{i}\,k)\,\Psi(k)\,e^{\mathrm{i}\,k\,x}\,dk\,.\tag{3.71}$$

If we substitute (3.70), with $\tilde k$ as the integration variable, into the Schrödinger equation

$$-\frac{d^2\psi(x)}{dx^2} + V(x)\,\psi(x) = E\,\psi(x)\,,\tag{3.72}$$

we therefore get

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}d\tilde k\,\Big(\tilde k^2\,\Psi(\tilde k) + V(x)\,\Psi(\tilde k) - E\,\Psi(\tilde k)\Big)\,e^{\mathrm{i}\,\tilde k\,x} = 0\,.\tag{3.73}$$

Muliplying this by $1/(\sqrt{2\pi})\, e^{-\mathrm{i}\,k\,x}$ and integrating over x, this gives

$$k^2\,\Psi(k) + \int_{-\infty}^{\infty} d\tilde{k}\,\Psi(\tilde{k})\,\Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} dx\,V(x)\,e^{\mathrm{i}\,(\tilde{k}-k)\,x}\Big) - E\,\Psi(k) = 0\,, \tag{3.74}$$

since the x integrations in the first and last terms simply give Dirac delta functions. The x integration in the potential term gives $1/(\sqrt{2\pi})\,\mathcal{V}(k-\tilde{k})$, where $\mathcal{V}$ is the inverse Fourier transform of the potential V, and so the Schrödinger equation in momentum space has become

$$k^2\,\Psi(k) + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathcal{V}(k-\tilde{k})\,\Psi(\tilde{k})\,d\tilde{k} = E\,\Psi(k)\,. \tag{3.75}$$

The term involving the potential here is precisely of the form of the convolution integral (3.63), and in fact we effectively re-derived the relation (3.64) here.

In quantum mechanics $|\psi(x)|^2\,dx$ is the probability that the particle lies in the interval $[x, x+dx]$ in position space. In terms of the momentum-space representation, $|\Psi(k)|^2\,dk$ is the probability that the momentum lies in the interval $[k, k+dk]$. This can be established by showing that the expectation value of the momentum, and all higher powers of the momentum, are the same whether calculated in the position-space or momentum-space representation. Parseval's theorem (3.56) tells us that the total probability for the particle to be somewhere ($=1$) is equal to the total probability for its momentum to be something. More generally, from (3.62), we can learn that an overlap integral between two wavefunctions $\psi_1(x)$ and $\psi_2(x)$ in position space is equal to the overlap integral evaluated in momentum space using their inverse Fourier transforms $\Psi_1(k)$ and $\Psi_2(k)$.

Poisson Summation Formula:

This can be expressed as follows. If $F(k)$ is the Fourier transform of $f(x)$, then

$$\sum_{n=-\infty}^{\infty} f(n\,z) = \frac{\sqrt{2\pi}}{z}\sum_{n=-\infty}^{\infty} F(2\pi\,n/z)\,. \tag{3.76}$$

To prove this, we simply use the definition of the inverse Fourier transform (3.41), together with the usual assumption of the interchangeability of the orders of integration and summation:

$$\begin{aligned}\sum_{n=-\infty}^{\infty} f(n\,z) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dk \sum_{n=-\infty}^{\infty} e^{-\mathrm{i}\,k\,n\,z}\,F(k)\,,\\ &= \sqrt{2\pi}\int_{-\infty}^{\infty} dk \sum_{n=-\infty}^{\infty} \delta(k\,z-2\pi\,n)\,F(k)\,,\\ &= \frac{\sqrt{2\pi}}{z}\int_{-\infty}^{\infty} dk' \sum_{n=-\infty}^{\infty} \delta(k'-2\pi\,n)\,F(k'/z)\,,\end{aligned}$$

$$= \frac{\sqrt{2\pi}}{z} \sum_{n=-\infty}^{\infty} F(2\pi\, n/z)\,, \tag{3.77}$$

where in the step from line 2 to line 3 we changed integration variable from k to $k' = k\,z$.

In the step from line 1 to line 2, we used the fact that

$$\sum_{n=-\infty}^{\infty} e^{\mathrm{i}\,n\,x} = 2\pi \sum_{n=-\infty}^{\infty} \delta(x - 2\pi\, n)\,. \tag{3.78}$$

Essentially, this is the statement that the functions $e^{\mathrm{i}\,n\,x}$ form a complete set on the unit circle: Taking our discussion at the begining of the section, and setting $b = 2\pi$ in (3.50), we see that for x restricted to a single covering of the unit circle, such as $-\pi \le x \le \pi$, we must have

$$\sum_{n=-\infty}^{\infty} e^{\mathrm{i}\,n\,x} = 2\pi\, \delta(x)\,. \tag{3.79}$$

Since obviously $e^{\mathrm{i}\,n\,x}$ is periodic in x, with period 2π, it must be that when x is allowed to range over the entire real line the function (3.79) must get repeated at intervals of 2π, giving rise to the "comb" of delta functions, as in (3.78).

An example of the use of the Poisson summation formula is to evaluate certain specific infinite sums. Consider, for example, the function $f(x) = 1/(1+x^2)$. Its Fourier transform is given by

$$F(k) = \frac{1}{\sqrt{2\pi\,\mathrm{i}}} \int_{-\infty}^{\infty} dx\, \frac{e^{\mathrm{i}\,k\,x}}{1+x^2} = \sqrt{\frac{\pi}{2}}\, e^{-|k|}\,. \tag{3.80}$$

(This is easily proven using the calculus of residues: If $k > 0$, the integration contour can be closed off with a large semicircle in the upper-half x plane, and so the integral is given by the residue of the pole at $x = \mathrm{i}$. On the other hand if $k < 0$, the contour can instead be closed off with a semicircle in the lower-half plane, and now one picks up the residue at $x = -\mathrm{i}$.) Applying the Poisson summation formula (3.76), we therefore get

$$\begin{aligned} \sum_{n=-\infty}^{\infty} f(n\,z) = \sum_{n=-\infty}^{\infty} \frac{1}{1+n^2\,z^2} &= \frac{\pi}{z} \sum_{n=-\infty}^{\infty} e^{-2\pi\,|n/z|}\,, \\ &= \frac{\pi}{z} \sum_{n=-\infty}^{-1} e^{2\pi\,n/z} + \frac{\pi}{z} \sum_{n=0}^{\infty} e^{-2\pi\,n/z}\,, \\ &= \frac{\pi}{z} \Big[\frac{e^{-2\pi/z}}{1-e^{-2\pi/z}} + \frac{1}{1-e^{-2\pi/z}}\Big]\,, \end{aligned} \tag{3.81}$$

and hence

$$\sum_{n=-\infty}^{\infty} \frac{1}{1+n^2\,z^2} = \frac{\pi}{z} \coth\Big(\frac{\pi}{z}\Big)\,. \tag{3.82}$$

Another application of the Poisson summation formula is the following. In the study of differential operators such as the Laplace operator ∇^2, it is sometimes necessary to study the distribution of its eigenvalues λ_n, defined by $-\nabla^2 u_n = \lambda_n u_n$, where u_n are the corresponding eigenfunctions. This can be done by studying the so-called *heat kernel*

$$\theta(t) \equiv \sum_n d_n\, e^{-\pi\, t\, \lambda_n}\,, \tag{3.83}$$

where d_n is the degeneracy of the eigenvalue λ_n. Clearly, if $\theta(t)$ is known for all t, then this encodes a lot of information about the values, and degeneracies, of the eigenvalues. Of particular importance is to know how $\theta(t)$ behaves for very small values of t, since this gives information about the limiting distribution of the eigenvalues for large λ_n.

Consider the following simple example, where we look at the 1-dimensional Laplacian $\nabla^2 = d^2/dx^2$ on the unit circle. The eigenfunctions are $e^{\mathrm{i}\, n\, x}$, with eigenvalues $\lambda_n = n^2$, and so

$$\theta(t) = \sum_{n=-\infty}^{\infty} e^{-\pi\, t\, n^2}\,. \tag{3.84}$$

If we let $f(x) = e^{-x^2/2}$, then $\theta(t)$ is of the form $\sum_n f(n\, z)$ as in (3.76), with $z = \sqrt{2\pi\, t}$. But the Fourier transform of $e^{-x^2/2}$ is just $e^{-k^2/2}$, since

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, e^{-x^2/2}\, e^{\mathrm{i}\, k\, x} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, e^{-(x-\mathrm{i}\, k)^2/2}\, e^{-k^2/2} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, e^{-y^2/2}\, e^{-k^2/2} = e^{-k^2/2}\,, \end{aligned} \tag{3.85}$$

where we have changed integration variable from x to $y = x - \mathrm{i}\, k$. Thus from (3.76) we find that

$$\sum_{n=-\infty}^{\infty} e^{-\pi\, t\, n^2} = \frac{1}{\sqrt{t}} \sum_{n=-\infty}^{\infty} e^{-\pi\, n^2/t}\,, \tag{3.86}$$

which when re-expressed in terms of $\theta(t)$, is nothing but

$$\theta(t) = \frac{1}{\sqrt{t}}\, \theta\Big(\frac{1}{t}\Big)\,. \tag{3.87}$$

Thus we have a remarkable relation between the large-t and small-t behaviour of the heat kernel for the Laplacian on the circle. In particular, since it is obvious from (3.84) that at large t have $\theta \sim 1$, we see that at small t we have

$$\theta(t) \sim \frac{1}{\sqrt{t}}\,. \tag{3.88}$$

3.3 The Laplace Transform

The Laplace transform is closely related to the Fourier transform. In the Fourier transform (3.40), it is evident that the function $f(x)$ should obey some suitable fall-off conditions at $x = \pm\infty$, in order that the integral be well-defined. Essentially, we sould require that $f(x) \longrightarrow 0$ as x tends to $\pm\infty$. Actually, since we have adopted the principle that delta-functions are acceptable "functions" we can be a little more tolerant. For example, we would say that the constant function $f(x) = 1$ has a valid Fourier integral (3.40), giving $F(k) = \sqrt{2\pi}\,\delta(k)$. More generally, $f(x)$ can be a sine or cosine or complex exponential. For example, if $f(x) = \cos x$, we shall have, from (3.40)

$$F(k) = \sqrt{\frac{\pi}{2}}\,\Big(\delta(k-1) + \delta(k+1)\Big)\,. \tag{3.89}$$

As it stands, we cannot, however, allow the function $f(x)$ to have any divergent behaviour at large $|x|$. The Laplace transform is effectively a modification of the concept of the Fourier transform that does allow such kinds of divergent behaviour for $f(x)$. The Laplace transform $F_L(p)$ of $f(x)$ is defined by

$$F_L(p) = \int_0^\infty e^{-p\,x}\, f(x)\, dx\,. \tag{3.90}$$

It is evident that this will be well-defined for $p > 0$, even if $f(x)$ has a power-law divergence $f(x) \sim x^m$ as x tends to infinity, for any arbitrarily large constant m. Even if $f(x)$ diverges exponentially, $f(x) \sim e^{a\,x}$, the integral will still be well-defined provided that $p > a$.

Obviously there is a rather close connection between the Laplace and the Fourier transforms. In fact, if we define $f_+(x)$ by

$$f_+(x) = \begin{cases} f(x) & x > 0 \\ 0 & x < 0 \end{cases}\,, \tag{3.91}$$

then the Fourier transform of $f_+(x)$ will be $F_+(k)$ given by

$$F_+(k) = \frac{1}{\sqrt{2\pi}} \int_0^\infty f(x)\, e^{\mathrm{i}\,k\,x}\, dx\,, \tag{3.92}$$

and so evidently we shall have

$$F_L(p) = \sqrt{2\pi}\, F_+(\mathrm{i}\,p)\,. \tag{3.93}$$

We now need to find the inverse of the Laplace transform. Again, this can be done by using what we already know about Fourier transforms.

Suppose that we are considering a function $f(x)$ that has an exponential divergence of the form $e^{a\,x}$ as x tends to infinity, where a is a constant with a positive real part. We may then introduce the function $g(x)$, which tends to zero as x tends to infinity, where

$$f(x) = e^{\gamma\,x}\, g(x)\,, \tag{3.94}$$

and γ is a real positive number such that $\gamma > \mathrm{Re}(a)$. The Fourier transform $G_+(k)$ of the function $g_+(x)$ given by

$$g_+(x) = \begin{cases} g(x) & x > 0 \\ 0 & x < 0 \end{cases} \tag{3.95}$$

is therefore well-defined, and so by Fourier's theorem we can then take the inverse Fourier transform of $G_+(k)$ to get back to $g_+(x)$. Hence we have

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt\, e^{\mathrm{i}\,x\,t} \int_0^{\infty} dy\, e^{-\mathrm{i}\,t\,y}\, g(y)\,. \tag{3.96}$$

From (3.94) this means that

$$f(x) = \frac{1}{2\pi}\, e^{\gamma\,x} \int_{-\infty}^{\infty} dt\, e^{\mathrm{i}x\,t} \int_0^{\infty} dy\, e^{-\mathrm{i}\,t\,y}\, e^{-\gamma\,y}\, f(y)\,. \tag{3.97}$$

Now change integration variable from t to $s = \gamma + \mathrm{i}\,t$. This gives

$$f(x) = \frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty} ds\, e^{s\,x} \int_0^{\infty} dy\, e^{-s\,y}\, f(y)\,. \tag{3.98}$$

The y integral here can be recognised as giving precisely the Laplace transform $F_L(s)$ of $f(y)$, and so (3.98) allows us to read off the inverse of the Laplace transform:

$$f(x) = \frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty} ds\, e^{s\,x}\, F_L(s)\,. \tag{3.99}$$

This is called the *Bromwich Integral*. The integration contour runs vertically in the complex s plane, along a line whose real part is γ. The real constant γ can be chosen arbitrarily, subject only to the requirement that the contour should run to the right of any singularities of $F_L(s)$. Any choice of γ that achieves this will do, and the answer does not depend on which such value for γ we choose.

Let us consider an example. Suppose we are given the function

$$F_L(s) = \frac{1}{s-a}\,, \tag{3.100}$$

where a is a real constant, and we are required to calculate its inverse Laplace transform. The function $F_L(s)$ has a pole at $s = a$, so we should take a contour in (3.99) with $\gamma > a$. The integral (3.99) will be

$$\frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty} ds\, \frac{e^{s\,x}}{s-a}\,. \tag{3.101}$$

This can be evaluated by means of the calculus of residues, by closing off the contour with a large semicircle swinging out and around to the west. This is justifiable for $x > 0$, since the function $e^{s\,x}$ will then become exponentially small on the semicircle as the radius goes to infinity. (See Part I of the course for a discussion of such integrals.) The closed contour encloses the simple pole at $s = a$, meaning that by the calculus of residues the integral just evaluates to give

$$\frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty} ds\, \frac{e^{s\,x}}{s-a} = e^{a\,x}\,, \qquad \text{for} \qquad x > 0\,. \tag{3.102}$$

Thus we have derived that the inverse Laplace transform of the function $1/(s-a)$ is $e^{a\,x}$.

This result is easily verified, by simply checking what the Laplace transform of $e^{a\,x}$ is. From (3.90), this will be

$$\begin{aligned} F_L(p) &= \int_0^\infty e^{a\,x}\, e^{-p\,x}\, dx = \int_0^\infty e^{-(p-a)\,x}\, dx \\ &= \Big[-\frac{e^{-(p-a)\,x}}{p-a} \Big]_{x=0}^{x=\infty} = \frac{1}{p-a}\,, \qquad (p > a)\,, \end{aligned} \tag{3.103}$$

which is indeed back to where we started. Observe how the function $e^{a\,x}$, whose Laplace transform is $1/(s-a)$, does diverge at large x (assuming a is positive), and, accordingly, the argument s of the Laplace transform $F_L(s) = 1/(s-a)$ is restricted to have $s > a$.[13]

The Laplace transform obeys general properties that are closely analogous to those for the Fourier transform that we discussed prevsiouly. If we denote by $\mathcal{L}_{\mathbf{L}}$ the operation of taking the Laplace transform, then we obviously have the linearity properties

$$\begin{aligned} \mathcal{L}_{\mathbf{L}}[f+g] &= \mathcal{L}_{\mathbf{L}}[f] + \mathcal{L}_{\mathbf{L}}[g]\,, \\ \mathcal{L}_{\mathbf{L}}[a\,f] &= a\,\mathcal{L}_{\mathbf{L}}[f]\,, \end{aligned} \tag{3.104}$$

where a is any constant. The analogue of the Fourier result (3.55) is a little more involved here, owing to the fact that the integration range in the Laplace transform is only semi-infinite. Thus if $F_L(p) = \mathcal{L}_{\mathbf{L}}[f(x)]$ is the Laplace transform of $f(x)$, then taking the Laplace transform of $f'(x)$ we get

$$\begin{aligned} \mathcal{L}_{\mathbf{L}}[f'(x)] &= \int_0^\infty dx\, e^{-p\,x}\, f'(x) = p\,F_L(p) + \Big[e^{-p\,x}\, f(x) \Big]_{x=0}^{x=\infty} \\ &= p\,F_L(p) - f(0)\,. \end{aligned} \tag{3.105}$$

[13]It might seem surprising that although the Laplace transform $F_L(s)$ is valid only for $s > a$, in our evaluation of the inverse transform in (3.99) we precisely place ourselves in the region $\mathrm{Re}(s) < a$ in the complex s-plane. This is just a manifestation of analytic continuation: The Laplace transform $F_L(s)$ was constructed under the requirement $s > a$, but having obtained it, it can actually be analytically extended to the entire complex s-plane, where it defines the meromorphic function $1/(s-a)$. It is this analytically extended function that is used in (3.99) to evaluate the inverse Laplace transform.

The Laplace transforms of higher derivatives of $f(x)$ can be calculated similarly. One finds, for example, that

$$\mathcal{L}_{\mathbf{L}}[f''(x)] = p^2\, F_L(p) - p\, f(0) + f'(0)\,. \tag{3.106}$$

<u>*Some Simple Laplace Transforms, and Their Uses:*</u>

First, let's take the Laplace transform of a few simple functions, to see what we get. The simplest of all is $f(x) = 1$, for which the Laplace transform will be

$$\mathcal{L}_{\mathbf{L}}[1] = \int_0^\infty dx\, e^{-p\,x} = \frac{1}{p}\,. \tag{3.107}$$

Of course we should note that this is true for $p > 0$. If $p \le 0$ the Laplace transform of $f(x) = 1$ does not exist.

Slightly less trivially, take $f(x) = x^{\nu-1}$. In order to have convergence of the integral at the lower limit, we must require $\mathrm{Re}(\nu) > 0$. However, it doesn't matter how big the real part of ν gets, because the exponential $e^{-p\,x}$ in (3.90) will ensure convergence at $x = \infty$, provided that p is positive. Then we shall have

$$\mathcal{L}_{\mathbf{L}}[x^{\nu-1}] = \int_0^\infty dx\, e^{-p\,x}\, x^{\nu-1} = p^{-\nu} \int_0^\infty dy\, e^{-y}\, y^{\nu-1} = \Gamma(\nu)\, p^{-\nu}\,. \tag{3.108}$$

Finally, consider taking $f(x) = e^{\mathrm{i}\,a\,x}$, which is closely related to a case we looked at previously. This gives

$$\begin{aligned} \mathcal{L}_{\mathbf{L}}[e^{\mathrm{i}\,a\,x}] &= \int_0^\infty dx\, e^{-x\,(p-\mathrm{i}\,a)} = \frac{1}{p - \mathrm{i}\,a}\,, \\ &= \frac{p + \mathrm{i}\,a}{p^2 + a^2}\,, \end{aligned} \tag{3.109}$$

again valid only for $p > 0$. Taking real and imaginary parts, we thus learn that the Laplace transforms of the cosine and sine functions are given by

$$\begin{aligned} \mathcal{L}_{\mathbf{L}}[\cos a\,x] &= \frac{p}{p^2 + a^2}\,, \\ \mathcal{L}_{\mathbf{L}}[\sin a\,x] &= \frac{a}{p^2 + a^2}\,. \end{aligned} \tag{3.110}$$

We saw earlier that one of the applications of integral transforms is for solving differential equations, by transforming them into a (hopefully!) simpler form. In fact we have studied some fairly complicated examples. For a little light relief, let's take a differential equation from kindergarten, and solve that using the Laplace transform. Suppose we have a harmonic oscillator, satisfying the familiar old equation

$$f''(x) + f(x) = 0\,, \tag{3.111}$$

subject, let's say, to the boundary conditions $y(0) = 1$, $y'(0) = 0$. Taking the Laplace transform of (3.111), and making use of the results (3.105) and (3.106) above, we obtain in general

$$p^2 F_L(p) + F_L(p) - p\, f(0) - f'(0) = 0\,. \tag{3.112}$$

This can then be solved algebraically for $F_L(p)$, in terms of the boundary conditions on $f(x)$ and $f'(x)$ at $x = 0$. In our example, we have $f(0) = 1$ and $f'(0) = 0$, and so

$$F_L(p) = \frac{p}{p^2+1}\,. \tag{3.113}$$

As it happens, we saw just a few paragraphs previously what function has this as its Laplace transform, namely $\cos x$ (see (3.110)), and so from (3.113) we conclude that the solution to the differential equation (3.111), subject to the given boundary conditions, is

$$f(x) = \cos x\,. \tag{3.114}$$

More generally, if $f(0)$ and $f'(0)$ were both non-vanishing, we would solve (3.113) to get

$$F_L(p) = f(0)\,\frac{p}{p^2+1} + f'(0)\,\frac{1}{p^2+1}\,. \tag{3.115}$$

Again, by good chance, we already know what function has this second term as its Laplace transform (see (3.110) again), and so here we conclude that the original differential equation (3.111) has the general solution

$$f(x) = f(0)\,\cos x + f'(0)\,\sin x\,. \tag{3.116}$$

Of course if we had not been fortunate enough to know the functions whose Laplace transforms give the two terms in (3.115) we could easily have derived them using the Bromwich integral (3.99) for the inverse Laplace transform, much as we did earlier in equation (3.101). One might begin to wonder, though, whether in this example one were using a sledgehammer to crack a nut![14] However, it is perhaps useful to have looked at the details of how one solves a differential equation by Laplace transform methods in a trivially simple example, since essentially the same techniques are used in more complicated cases too.

Convolution Theorem for the Laplace Transform:

There is a convolution theorem for the Laplace transform that is closely analogous to the one for the Fourier transform that we met previously. Recalling that we first obtained the

[14] There is a Latin phrase *ignotum per ignotius*, which is perhaps applicable here.

Laplace transform from the Fourier transform by considering functions of the form $f_+(x)$ defined in (3.91), which vanish for $x < 0$ and equal $f(x)$ for $x > 0$, we should now use such functions in the type of convolution integral (3.63) that we studied before. Thus we may define

$$h(x) = \int_{-\infty}^{\infty} f_+(y)\, g_+(x-y)\, dy = \int_0^x f(y)\, g(x-y)\, dy\,. \tag{3.117}$$

(We do not include a $1/\sqrt{2\pi}$ factor here because the overall 2π that comes from taking a transform followed by its inverse is, by convention, treated asymmetrically in the case of the Laplace transform.) The substantial point to notice is that the convolution integral for two functions, in the context of a Laplace transform, is defined with integration limits running from 0 to x:

$$h(x) \equiv \int_0^x f(y)\, g(x-y)\, dy\,. \tag{3.118}$$

This has happened, obviously, because of the vanishing of $f_+(x)$ and $g_+(x)$ when x is negative.

The most direct way to derive the convolution theorem here is to take a Laplace transform of (3.118). Thus we get

$$\begin{aligned} H_L(p) &= \int_0^\infty dx\, e^{-p\,x}\, h(x) = \int_0^\infty dx \int_0^x dy\, e^{-px}\, f(y)\, g(x-y) \\ &= \int_0^\infty dy \int_y^\infty dx\, e^{-px}\, f(y)\, g(x-y) \\ &= \int_0^\infty dy \int_0^\infty dz\, e^{-p\,(y+z)}\, f(y)\, g(z) = \int_0^\infty dy\, e^{-p\,y}\, f(y) \int_0^\infty dz\, e^{-p\,z}\, g(z) \\ &= F_L(p)\, G_L(p)\,. \end{aligned} \tag{3.119}$$

In getting to the second line, we have used the fact that the original region of integration in the (x, y) plane is only the lower-triangular half of the positive (x, y) quadrant, i.e. the triangular area between the positive x-axis and the line $y = x$. In the integration on line 1, it is covered by vertical strips, $0 < y < x$, with x then running up to infinity. It can instead be covered by horizontal strips, $y < x < \infty$, with y running from 0 to infinity, and this is what is done in line 2. To get to line 3, we then make a shift of the x integration variable, to $z = x - y$, implying that now the second integral runs from $z = 0$ to $z = \infty$. The two integrals now fall apart into a product of two independent ones, giving the product of the Laplace transforms of $f(x)$ and $g(x)$. Thus we have concluded that if $F_L(p)$, $G_L(p)$ and $H_L(p)$ are the Laplace transforms of $f(x)$, $g(x)$ and $h(x)$ respectively, and if $h(x)$ is the convolution of $f(x)$ and $g(x)$ defined in (3.118), then

$$H_L(p) = F_L(p)\, G_L(p)\,. \tag{3.120}$$

Notice, by the way, that the convolution (or *Faltung*) defined in (3.118) has the same symmetry property as the one defined in (3.63) for the Fourier transform. Namely, if we change integration variable in (3.118) from y to $z = x - y$, then we find that

$$h(x) = \int_0^x f(y)\, g(x-y)\, dy = \int_0^x g(z)\, f(x-z)\, dz\,. \tag{3.121}$$

Again, the symmetry between f and g is even more manifest in the Laplace-transformed expression (3.120).

Here is a simple example of the use of the convolution theorem in solving a differential equation. Like our previous example, we'll take the simple-harmonic equation, but this time with a source term:

$$f''(x) + f(x) = g(x)\,. \tag{3.122}$$

For simplicity, suppose that $f(0) = f'(0) = 0$ here. Thus from (3.105) and (3.106), we find that the Laplace transform of the equation is

$$p^2\, F_L(p) + F_L(p) = G_L(p)\,, \tag{3.123}$$

where $G_L(p)$ is the Laplace transform of the source term $g(x)$. Solving for $F_L(p)$ we get

$$F_L(p) = G_L(p)\, \frac{1}{p^2+1}\,. \tag{3.124}$$

Since we can recognise the factor $1/(p^2+1)$ as the Laplace transform of $\sin x$ (see (3.110)), we can invoke the convolution theorem to give us

$$f(x) = \int_0^x g(x-y)\, \sin y\, dy\,. \tag{3.125}$$

This result is, of course, easily derivable by other methods too, but again it serves to illustrate a method that has rather general applicability.

3.4 The Gibbs Phenomenon

In our proof of Fourier's theorem earlier, we invoked the easily-proven results for the discrete analogue of the Fourier transform, namely the Fourier series. We remarked at that time that there was an interesting subtlety in the Fourier expansion, known as the *Gibbs Phenomenon.* Although it is slightly off the mainstream of our present discussion, it is perhaps interesting to look at it here, since it may not come up again later.

The Gibbs phenomenon is seen when one considers the Fourier series expansion for a function with a discontinuity. This happens quite often in a Fourier series, since it describes a periodic function which can, for example, have a sudden "jump" when the end of the period is reached. Let us consider a concrete example, of a square-wave with period 2π, which can therefore be expanded in terms of the complex exponential functions $e^{\mathrm{i}\,n\,x}$, as

$$f(x) = \sum_{n=-\infty}^{\infty} a_n\, e^{\mathrm{i}\,n\,x}\,. \tag{3.126}$$

Let us take $f(x)$ to be

$$f(x) = \begin{cases} +1 & 0 < x < \pi \\ -1 & \pi < x < 2\pi \end{cases}\,. \tag{3.127}$$

As in (3.49), the Fourier coefficients will then be given by

$$\begin{aligned} a_n &= \frac{1}{2\pi}\int_0^{2\pi} dy\, e^{-\mathrm{i}\,n\,y}\, f(y) \\ &= \frac{1}{2\pi}\int_0^{\pi} dy\, e^{-\mathrm{i}\,n\,y} - \int_\pi^{2\pi} dy\, e^{-\mathrm{i}\,n\,y} \\ &= \frac{1}{\mathrm{i}\,\pi\,n}\Big(1-(-1)^n\Big)\,, \end{aligned} \tag{3.128}$$

and they are non-zero only when n is odd. Noting that in the sum (3.126) we can then replace n by $-n$ as the summation variable when n is negative, we conclude that the square-wave (3.127) has the Fourier series expansion

$$f(x) = \frac{4}{\pi}\sum_{r=0}^{\infty}\frac{1}{(2r+1)}\,\sin[(2r+1)\,x] = \frac{4}{\pi}\Big(\sin x + \tfrac{1}{3}\,\sin 3x + \tfrac{1}{5}\,\sin 5x + \cdots\Big)\,. \tag{3.129}$$

Obviously the terms are getting smaller in magnitude as r increases, and so we can expect that if we consider a partial sum from $r=0$ only as far as $r=M$, we should get a better and better approximation to the square wave as M increases. And essentially, this expectation is correct, except that there is one small subtlety that one might not have foreseen. This can be best illustrated first by looking at a few plots of the partial sums in (3.129) where only the first few terms are inlcuded. Below, in Figures 12-16, we give the plots for the first term alone (a sine wave); the first two terms; the first three; the first ten, and finally the first twenty.

As can be seen from the various plots, it is indeed broadly-speaking true that as we include more and more terms in the sum, we get a closer and closer approximation to the square wave (3.127). However, it also becomes apparent that no matter how many terms we include, there always seems to be an "overshoot" every time there is a discontinuity in the

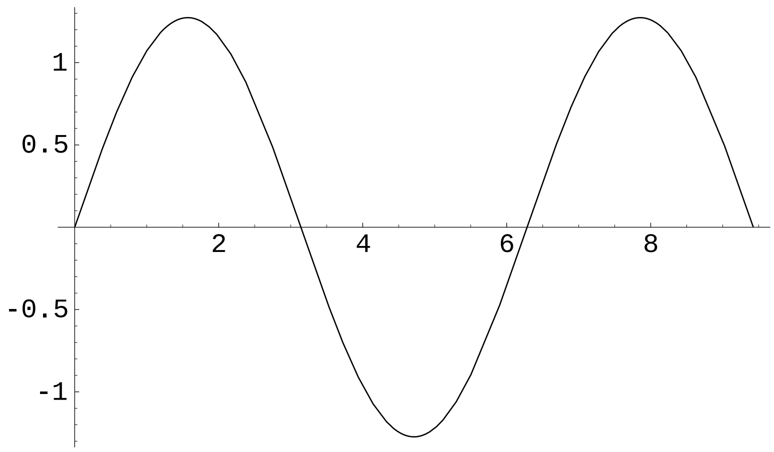

Figure 12: The first term in the Fourier series for the square wave

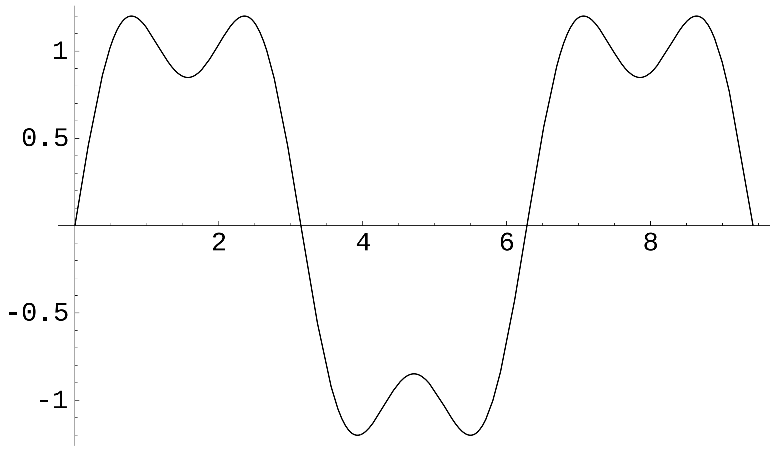

Figure 13: The first 2 terms in the Fourier series for the square wave

square-wave. As we include more terms in the sum, the width of the overshoot gets less, but its height seems to be staying roughly the same. This overshoot is the Gibbs phenomenon. We can show relatively easily that it will *always* be there, no matter how many terms we include in the sum. And indeed, it always leads to something like an 18% overshoot of the true value of the function, at the discontinuity. Actually, we should remark that there is more than just a single overshoot; as can be seen rather clearly in Figure 16 there is a sort of "ringing" phenomenon which occurs after the overshoot, which takes a while to settle down.

To study the Gibbs phenomenon, we go back to the second line in (3.128), and leaving the integrals unevaluated, substitute the expressions for the coefficients a_n back into (3.126). However, we shall now restrict the summation to run only over the finite range $-N \le n \le N$.

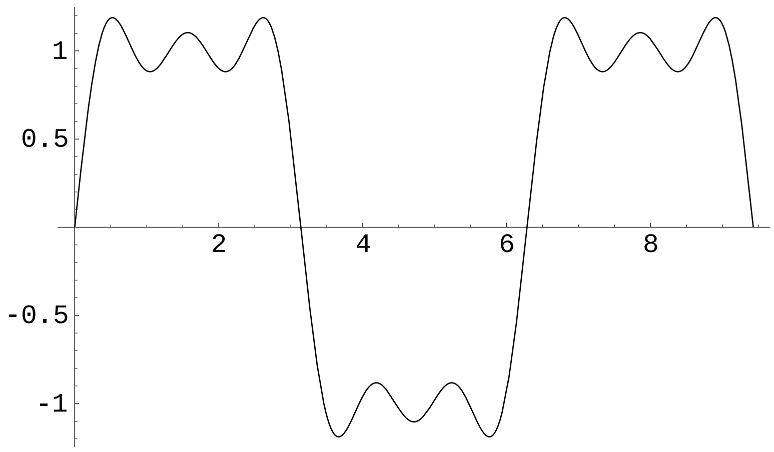

Figure 14: The first 3 terms in the Fourier series for the square wave

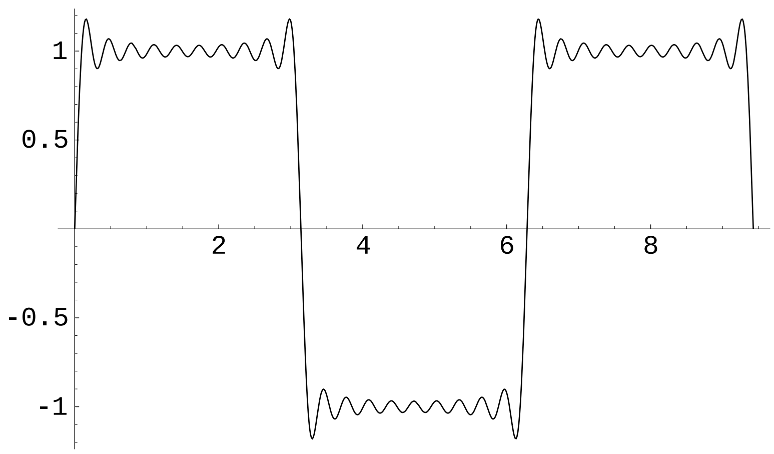

Figure 15: The first 10 terms in the Fourier series for the square wave

At the same time interchanging the orders of the integration and the summation, this gives

$$S_N(x) = \frac{1}{2\pi} \int_0^\pi dy \sum_{n=-N}^{N} e^{\mathrm{i}\, n\, (x-y)} - \frac{1}{2\pi} \int_\pi^{2\pi} dy \sum_{n=-N}^{N} e^{\mathrm{i}\, n\, (x-y)} \,. \tag{3.130}$$

We can explicitly evaluate the sum here, since it is just a geometrical series:

$$\begin{aligned} \sum_{n=-N}^{N} e^{\mathrm{i}\, n\, (x-y)} &= e^{-N\,(x-y)} \sum_{n=0}^{2N} e^{\mathrm{i}\, n\, (x-y)} = e^{-N\,(x-y)} \left[\frac{1 - e^{\mathrm{i}\,(2N+1)\,(x-y)}}{1 - e^{\mathrm{i}\,(x-y)}} \right] , \\ &= \frac{\sin[(N+\frac{1}{2})(x-y)}{\sin[\frac{1}{2}(x-y)]} \,. \end{aligned} \tag{3.131}$$

Plugging (3.131) into (3.130), and changing integration variable from y to $\theta = y - x$ in the first integral, and $\theta = 2\pi - (y - x)$ in the second, we get

$$S_N(x) = \frac{1}{2\pi} \int_{-x}^{\pi - x} d\theta\, \frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta} - \frac{1}{2\pi} \int_x^{\pi + x} d\theta\, \frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta} \,. \tag{3.132}$$

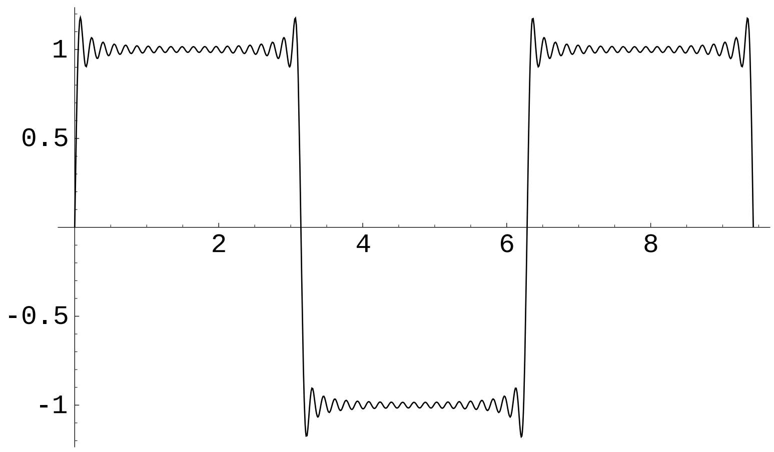

Figure 16: The first 20 terms in the Fourier series for the square wave

Juggling the integration limits around, by using

$$\int_{-x}^{\pi-x} - \int_{x}^{\pi+x} = \int_{c}^{\pi-x} - \int_{c}^{-x} - \int_{c}^{\pi+x} + \int_{c}^{x} = \int_{-x}^{x} - \int_{\pi-x}^{\pi+x}, \tag{3.133}$$

this can be rewritten as

$$S_N(x) = \frac{1}{2\pi} \int_{-x}^{x} d\theta \, \frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta} - \frac{1}{2\pi} \int_{\pi-x}^{\pi+x} d\theta \, \frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta}. \tag{3.134}$$

Now let $u = (N+\frac{1}{2})\,\theta$, leading to

$$\begin{aligned} S_N(x) &= \frac{1}{\pi} \int_{-(N+\frac{1}{2})x}^{(N+\frac{1}{2})x} du \, \frac{\sin u}{(2N+1)\, \sin[u/(2N+1)]} \\ &\quad - \frac{1}{\pi} \int_{(N+\frac{1}{2})(\pi-x)}^{(N+\frac{1}{2})(\pi+x)} du \, \frac{\sin u}{(2N+1)\, \sin[u/(2N+1)]}. \end{aligned} \tag{3.135}$$

Suppose now that we look in the region $0 < x < \pi$, with x significantly smaller than π. The first integral in (3.135) will be much larger than the second one, when N is large. To see this, note that the argument of the sine function in the denominator of the integrand, $u/(2N+1)$ is ranging over the values

$$-\tfrac{1}{2}\,x \le u/(2N+1) \le \tfrac{1}{2}\,x \tag{3.136}$$

in the first integral, while in the second integral it is ranging over the values

$$\tfrac{1}{2}\,(\pi - x) \le u/(2N+1) \le \tfrac{1}{2}\,(\pi + x)\,. \tag{3.137}$$

Thus the denominator of the integrand never goes to zero in the second integral, and this integral tends to zero as N tends to infinity. On the other hand, the denominator of the integrand *does* go to zero within the integration range in the first integral. At large N, this gives, to a good approximation

$$S_N(x) \approx \frac{1}{\pi} \int_{-\infty}^{\infty} du\, \frac{\sin u}{u}\,, \qquad \text{for} \qquad 0 < x < \pi\,, \tag{3.138}$$

when N gets very large. The integral here is a standard one (we evaluated it in Part I of the course, using Cauchy's principal-value integral, for example), implying that

$$S_N(x) \approx 1\,, \qquad \text{for} \qquad 0 < x < \pi\,, \tag{3.139}$$

exactly as we would hope.

In the above, we assumed that x was greater than zero, but less than π, and that it is held fixed as N was sent to infinity. We showed that $S_N(x)$ then converges to 1 as N is sent to infinity. Suppose instead we now arrange to sit on the peak of the Gibbs overshoot, and see what happens there as N is sent to infinity. This peak will occur when $S_N'(x)$ has its first zero as x increases from 0, and clearly it will be at a very small value of x when N is large. Let it occur at $x = \delta$. Again the second integral in (3.135) will be negligible compared with the first when N gets large, and so for small positive x we know that $S_N(x)$ is given approximately by

$$S_N(x) \approx \frac{1}{\pi} \int_{-(N+\frac{1}{2})\,x}^{(N+\frac{1}{2})\,x} du\, \frac{\sin u}{u}\,, \tag{3.140}$$

since the argument $u/(N+\frac{1}{2})$ in the sine function in the denominator is so small that we can approximate $\sin[u/(N+\frac{1}{2})]$ by $u/(N+\frac{1}{2})$. This integral is expressible in terms of the *Sine Integral*

$$Si(x) \equiv \int_0^x du\, \frac{\sin u}{u}\,. \tag{3.141}$$

First, however, we need to differentiate (3.140) with respect to x, to find the first zero of $S_N'(x)$ as x increases from 0. This is easy, since it just gives

$$S_N'(x) \approx \frac{2}{\pi\, x}\, \sin[(N+\tfrac{1}{2})\, x]\,. \tag{3.142}$$

The first zero therefore occurs at

$$x = \delta = \frac{2\pi}{2N+1}\,. \tag{3.143}$$

Plugging into the expression (3.140 for $S_N(x)$, we find that

$$\lim_{N\to\infty} S_N(\pi/(2N+1)) = \frac{1}{\pi} \int_{-\pi}^{\pi} du\, \frac{\sin u}{u} = \frac{2}{\pi}\, Si(\pi) = 1.1798\ldots\,. \tag{3.144}$$

Thus we see that the first peak exceeds the true value $f(x) = 1$ by about 18%, even as N is sent to infinity.[15] As can be seen from (3.143), the width of the overshoot spike gets smaller and smaller as N increases, becoming vanishingly small in the limit.

It may be recalled, for example from Part I of the course, that the expressions in the top line of (3.128) for the Fourier expansion coefficients a_n can be shown to optimise the accuracy of the expansion for the function $f(x)$. Furthermore, these expressions for the a_n are optimal not only for the entire infinite series expansion, but also if one takes only a partial sum, as we have been doing. How does this square up with what we have been seeing with the Gibbs phenomenon? After all, 18% is a pretty serious error! The resolution, of course, is that as we have seen, the width of the overshoot-spike gets less and less as the number of terms included in the partial sum is increased. And when one says that the choice (3.128) for the a_n coefficients in the Fourier series is the one that gives the "best fit" to the function $f(x)$, it should be recalled that the measure of success here is defined to be a least-squares average. Namely, the choice for the coefficients a_n in (3.128) minimises the quantity

$$Q_N \equiv \int_0^{2\pi} \left| f(x) - \sum_{n=-N}^{N} a_n \, e^{\mathrm{i}\, n\, x} \right|^2 dx \,, \tag{3.146}$$

making it vanish in the limit where N goes to infinity. It is evident that the overshoot-spikes associated with the Gibbs phenomenon will give no contribution in the limit when N goes to infinity, since their height is finite (about 9% of the discontinuity; in our example the function jumps from -1 to $+1$ at $x = 0$), while their width goes to zero.

We can also examine the details of the "ringing" that is clearly visible in Figure 16, by looking at the values of the function $S_N(x)$ at its first few extrema. As before, the locations of these points are easily determined from the expression (3.142) for $S'_N(x)$. Thus the m'th zero of $S'_N(x)$ is at

$$x = \delta_m = \frac{2\pi\, m}{2N+1} \,. \tag{3.147}$$

[15] Note that Morse and Feshbach spoil an otherwise nice derivation of this result (at least in the edition I have) by miscalculating the location of the peak in the final stage of the computation. They obtain the expression (3.144) with limits $\pm\pi/2$ in the integral, and then make the false claim that

$$\frac{1}{\pi} \int_{-\pi/2}^{\pi/2} du \, \frac{\sin u}{u} = 1.1798 \ldots \tag{3.145}$$

although the actual value of their integral is $0.8726\ldots$. Their mis-identification of the location of the peak has actually set them at a point where $S_N(x)$ is *smaller* than 1. Even Homer nods, occasionally!

In the limit when N becomes large, the value of $S_N(\delta_m)$ is then given by

$$S(\delta_m) = \frac{1}{\pi} \int_{-m\,\pi}^{m\,\pi} du \, \frac{\sin u}{u} \,. \tag{3.148}$$

Taking $m = 1$ gives us back the results (3.144) for the value at the first peak. As we take $m = 3, 5, 7, \ldots$ we will get the values at the later peaks, while taking $m = 2, 4, 6, \ldots$ will give the values at the successive troughs in between the peaks. The results for the first few peaks and troughs are given below:

$m =$	1	3	5	7	9
$S(\delta_m) =$	1.17898	1.06619	1.04021	1.02883	1.02246

The values of the first five peaks

$m =$	2	4	6	8	10
$S(\delta_m) =$	0.90282	0.94994	0.96641	0.97475	0.97978

The values of the first five troughs

Finally, we may remark that although we focussed on the example of a square-wave function expressed as a Fourier series, the Gibbs phenomenon is a very general one. Any time that one makes a series expansion of a function with discontinuities, as a sum over some complete set of eigenfunctions of a Sturm-Liouville operator, the same phenomenon of overshoot-spikes and ringing will occur.

4 Integral Equations

4.1 Introduction

The idea of formulating physical laws in terms of differential equations is a very familiar and fundamental one. Indeed, all the fundamental laws of physics fall into this category; for example the Maxwell equations, the Einstein equations of general relativity, and the equations governing the fundamental particle interactions of the strong and weak interactions. There are times, however, when it turns out that a system can be more conveniently described in terms of integral equations, and in some cases where one is dealing with an effective macroscopic theory rather than a fundamental one, a description in terms of integal equations becomes a necessity.

Let us begin by introducing the most common types of integral equation that one encounters. We shall discuss four types, which are as follows:

Fredholm Equation of the First Kind:

$$f(x) = \int_a^b K(x,t)\,\phi(t)\,dt\,, \tag{4.1}$$

Fredholm Equation of the Second Kind:

$$\phi(x) = f(x) + \lambda \int_a^b K(x,t)\,\phi(t)\,dt\,, \tag{4.2}$$

Volterra Equation of the First Kind:

$$f(x) = \int_a^x K(x,t)\,\phi(t)\,dt\,, \tag{4.3}$$

Volterra Equation of the Second Kind:

$$\phi(x) = f(x) + \lambda \int_a^x K(x,t)\,\phi(t)\,dt\,, \tag{4.4}$$

In all four cases, $\phi(t)$ is the unknown function that must be solved for. The kernel $K(x,t)$ is given, as is the function $f(x)$ in the two equations of the second kind. If the function $f(x)$ is zero, the equation is said to be *homogeneous*, since it then scales uniformly under a constant scaling of $\phi(t)$. The quantity λ the integral equations of the second kind is a constant.

First, let's establish a mnemonic for remembering which equation is which. The difference between the Fredholm and the Volterra equations is that the **F**redholm equations have **F**ixed limits of integration, while the **V**olterra equations have **V**ariable limits of integration. Integral equations of the **S**econd kind have a **S**econd term as well as the integral, while the equations of the **F**irst kind have **F**ewer terms. So that is easy!

Notice that the Fredholm equation of the first kind looks very like the sort of equation we have encountered already in our discussion of integral transforms. Essentially, the equation can be viewed as taking the transform of $\phi(t)$ using the kernel $K(x,t)$. In order to solve for $\phi(t)$, we therefore need to find the inverse transform. This would be very easy, for example, if the given kernel function was $K(x,t) = e^{\mathrm{i}\,x\,t}$, since then we would simply have to take the inverse Fourier transform of the given function $f(x)$ in order to obtain our solution $\phi(t)$.

Another example of an integral equation that we have already encountered is the Schrödinger equation re-expressed in momentum space, which we obtained in equation (3.75):

$$(E - k^2)\,\Psi(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathcal{V}(k-\tilde{k})\,\Psi(\tilde{k})\,d\tilde{k}\,, \tag{4.5}$$

where $\mathcal{V}$ is the inverse Fourier transform of the potential $V(x)$. This is a homogeneous Fredholm equation of the second kind. We already have a clue about how one might solve it, from the fact that we obtained it from an ordinary differential equation by taking a Fourier transform.

We can, however, imagine a more general situation in this quantum-mechanical example, for which an integral equation beomces unavoidable. Let us go back to the original x-space Schrödinger equation,

$$-\frac{d^2\psi(x)}{dx^2} + V(x)\,\psi(x) = E\,\psi(x)\,, \tag{4.6}$$

and re-write it as

$$\frac{d^2\psi(x)}{dx^2} + E\,\psi(x) = \int_{-\infty}^{\infty} V(x,x')\,\psi(x')\,dx'\,. \tag{4.7}$$

This becomes identical to (4.6) if $V(x,x')$ is given by

$$V(x,x') = V(x)\,\delta(x-x')\,. \tag{4.8}$$

When (4.8) holds the interaction is an ordinary local one; the wavefunction at the point x senses the potential at the same point x. More generally, one could consider situations with non-local interactions, in which the wavefunction at x senses the effects from other positions too, and this is what is described by (4.7). Such interactions would not be desirable in a theory at the fundamental level (imagine the possible implications for acausal faster-than-light transfer of information, for example!).[16] However, they could arise at some effective level. The non-local equation (4.7) is an integro-differential equation, with $\psi(x)$ appearing both *via* its derivatives, and within an integral.

[16]In any case the Schrödinger equation itself is clearly not "fundamental" since it is not even relativistic.

One can Fourier-transform the non-local equation (4.7), much as we did earlier for the usual local equation, to obtain

$$(E - k^2)\,\Psi(k) = \int_{-\infty}^{\infty} \mathcal{V}(k,\tilde{k})\,\Psi(\tilde{k})\,d\tilde{k}\,, \tag{4.9}$$

where

$$\mathcal{V}(k,\tilde{k}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, V(x,y)\, e^{-\mathrm{i}\,(k\,x - \tilde{k}\,y)}\,. \tag{4.10}$$

The previous local condition (4.8) can easily be seen to imply

$$\mathcal{V}(k,\tilde{k}) = \frac{1}{\sqrt{2\pi}}\,\mathcal{V}(k-\tilde{k})\,, \tag{4.11}$$

and then (4.9) reduces to the previous result (4.9). The general result (4.9) is itself of the form of a homogeneous Fredholm equation of the second kind.

In this example, once we have generalised to the non-local interaction, it is most natural to write the equation for $\psi(x)$ in the form of an integro-differential equation, and indeed there is really no way to write a pure differential equation. This is inevitable, in view of the non-local nature of the interaction that is being described. We improve things, in some sense, by transforming to momentum space, since now the equation becomes purely an integral equation.

In other examples one has a choice as to whether to work with an equation in integral or differential form. One might think that in such cases it is better to stick with the more familiar differential form. There are, however, certain advantages to having an equation expressed in integral form, most notably assoicated with the issue of boundary conditions. In a differential equation one has to supply information about the boundary conditions as supplementary data. In an integral equation, on the other hand, the information about the boundary conditions is effectively already encoded in the equation itself. This can be useful, for example, if one is wanting to study the asymptotic properties of the solution, subject to specific boundary conditions, in a case where approximate methods must be used.

An Integral Equation from a Differential Equation:

The point about the boundary conditions can be illustrated by constructing an example, somewhat artificially. Consider the second-order ordinary differential equation

$$y''(x) + p(x)\,y'(x) + q(x)\,y(x) = g(x)\,, \tag{4.12}$$

with specified boundary conditions

$$y(a) = y_0\,, \qquad y'(a) = y_0'\,. \tag{4.13}$$

This can be turned into an integral equation by the following procedure. First, we integrate (4.12):

$$y'(x) = -\int_a^x p(t)\, y'(t)\, dt - \int_a^x q(t)\, y(t)\, dt + \int_a^x g(t)\, dt + y_0' \,. \tag{4.14}$$

Notice that we have specified the lower limit of the integration, and thus we have been able to incorporate the boundary condition on $y'(a)$ from (4.13). Now integrate the first term on the right-hand side by parts, to get

$$y'(x) = -p(x)\, y(x) + \int_a^x \Big(p'(t) - q(t)\Big)\, y(t)\, dt + \int_a^x g(t)\, dt + p(a)\, y_0 + y_0' \,. \tag{4.15}$$

Next, we integrate this equation again:

$$\begin{aligned} y(x) \;\;&=\;\; -\int_a^x p(t)\, y(t)\, dt + \int_a^x ds \int_a^s dt \Big(p'(t) - q(t)\Big)\, y(t) + \int_a^x ds \int_a^s dt\, g(t) \\ &\quad\; + \Big(p(a)\, y_0 + y_0'\Big)\, (x-a) + y_0 \,. \end{aligned} \tag{4.16}$$

At this stage we note that by integrating by parts, we can show that for any function $f(t)$ we shall have[17]

$$\int_a^x ds \int_a^s dt\, f(t) = -\int_a^x ds\, s\, f(s) + \Big[s \int_a^s dt\, f(t)\Big]_{s=a}^{s=x} = \int_a^x dt\, (x-t)\, f(t)\, dt \,. \tag{4.17}$$

Using this, we can re-express (4.16) as

$$\begin{aligned} y(x) \;\;&=\;\; -\int_a^x dt\, p(t)\, y(t) + \int_a^x dt\, (x-t)\, \Big(p'(t) - q(t)\Big)\, y(t) + \int_a^x dt (x-t)\, g(t) \\ &\quad\; + \Big(p(a)\, y_0 + y_0'\Big)\, (x-a) + y_0 \,. \end{aligned} \tag{4.18}$$

Finally, we introduce functions $K(x,t)$ and $f(x)$ defined as follows:

$$\begin{aligned} K(x,t) \;&\equiv\; (x-t)\, \Big(p'(t) - q(t)\Big) - p(t)\,, \\ f(x) \;&\equiv\; \int_a^x dt\, (x-t)\, g(t) + \Big(p(a)\, y_0 + y_0'\Big)\, (x-a) + y_0 \,. \end{aligned} \tag{4.19}$$

(Note that these are constructed purely from the original quantities given in the differential equation and the boundary conditions.) We can now write the equation (4.18) in the final form

$$y(x) = f(x) + \int_a^x K(x,t)\, y(t)\, dt \,. \tag{4.20}$$

This can be recognised as a Volterra equation of the second kind. Notice that all information about the boundary conditions is already encoded in the formulation of the equation. For

[17]If you look at this discussion in Arfken, he makes a real dog's breakfast of it, by confusing the dummy integration variable s and the integration limit x.

example, if we set $x = a$ in (4.20) we learn that $y(a) = f(a)$, and from the definition of $f(x)$ in (4.19), this tells us that $y(a) = y_0$.

Consider a simple example, where $p(x) = 0$ and $q(x) = 1$, and $g(x) = 0$, so that the original differential equation (4.12) is just the simple harmonic oscillator,

$$y''(x) + y(x) = 0\,. \tag{4.21}$$

Suppose also that we choose our boundary conditions so that $y_0 = 0$, $y_0' = 1$. From (4.19) and (4.20) we therefore get the integral equation

$$y(x) = x + \int_0^x (t - x)\, y(t)\, dt\,. \tag{4.22}$$

One can easily verify that this is satisfied by $y(x) = \sin x$. Of course this is not a "derivation" of the solution, more a verification that what we already know actually works. We shall discuss later how one goes about solving such equations.

<u>An Example with Two End-point Boundary Conditions:</u>

The derivation above was tailored specifically to the case where the boundary conditions were as stated in (4.13). Clearly we could adjust the derivation slightly to accomodate different types of boundary condition. Since our principle objective at this stage is not simply to turn familar differential equations into unfamiliar integral equations, we shall not pursue this point in great detail here. Let us take one specific example, with different boundary conditions, in order to illustrate the point. Consider again the harmonic oscillator equation (4.21), but now with the boundary conditions

$$y(0) = 0\,, \qquad y(a) = 0\,. \tag{4.23}$$

Integrating (4.21) once gives

$$y'(x) = -\int_0^x y(t)\, dt + y'(0)\,. \tag{4.24}$$

We don't know yet what to substitute for $y'(0)$, since this is not one of the given boundary conditions any more. So we proceed by integrating again, to get

$$y(x) = -\int_0^x (x - t)\, y(t)\, dt + y'(0)\, x\,, \tag{4.25}$$

after using (4.17). Now we can set $x = a$, and thereby obtain an expression for $y'(0)$:

$$y'(0) = \frac{y(0)}{a} + \frac{1}{a} \int_0^a (a - t)\, y(t)\, dt\,, \tag{4.26}$$

which can be plugged back into (4.25) to give

$$y(x) = -\int_0^x (x-t)\, y(t)\, dt + \frac{x}{a} \int_0^a (a-t)\, y(t)\, dt\,. \tag{4.27}$$

Using the identity that $-(x-t) = t\,(a-x)/a - x\,(a-t)/a$, we therefore get

$$y(x) = \int_0^a \frac{t}{a}\,(a-x)\, y(t)\, dt + \int_x^a \frac{x}{a}\,(a-t)\, dt\,. \tag{4.28}$$

Now define the kernel $K(x,t)$ by

$$K(x,t) = \begin{cases} \frac{t}{a}\,(a-x)\,, & t < x \\ \\ \frac{x}{a}\,(a-t)\,, & x < t \end{cases}\,, \tag{4.29}$$

in terms of which (4.28) can be written as

$$y(x) = \int_0^a K(x,t)\, y(t)\, dt\,. \tag{4.30}$$

This is a homogeneous Fredholm equation of the second kind. The kernel $K(x,t)$ here is in fact the Green function for the equation (4.21), subject to the boundary conditions $y(0) = y(a) = 0$. It is symmetric in x and t. If plotted as a function of t, it consists of a straight-line segment starting at the origin, and increasing with positive gradient $1 - x/a$ until the point $t = x$ is reached. For $t > x$ it is a straight-line segment with negative gradient $-x/a$, which reaches the t axis at $t = a$. The kernel is continuous at $t = x$, but with a discontinuity of -1 in its gradient there.

Solutions Using Fourier and Laplace Transforms:

We have already remarked that if one were presented with the following Fredholm equation of the first kind,

$$f(x) = \int_{-\infty}^{\infty} e^{\mathrm{i}\,x\,t}\, \phi(t)\, dt\,, \tag{4.31}$$

then solving for $\phi(t)$ would be easy, since we just recognise this as a Fourier transform. Thus we can invoke Fourier's theorem and immediately write down the solution, namely

$$\phi(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}\,x\,t}\, f(x)\, dx\,. \tag{4.32}$$

Of course when we say that we have *solved* the equation here, what we mean is that we have "reduced it to quadratures." Whether or not an explicit closed-form solution can be presented depends on whether the given function $f(x)$ allows us to perform the integral explicitly.

Similarly, there are other Fredholm equations of the first kind that could be recognised as Laplace transforms, or certain other related transforms such as the Mellin or Hankel transforms. In all such cases, a procedure for solving the equation by inverting the transformation exists.

There are somewhat more general types of integral equation that can also be solved by Fourier transform techniques, or by analogous procedures related to the other classified integral transforms. Suppose we have the following Fredholm equation of the first kind:

$$f(x) = \int_{-\infty}^{\infty} k(x-t)\,\phi(t)\,dt\,, \tag{4.33}$$

where $k(x-t)$ is the given kernel, and we wish to solve for $\phi(t)$. Note that the kernel is rather special here, being a function of just the single variable combination $(x-t)$. We can recognise (4.33) as being nothing but a convolution integral of the functions k and ϕ. As we saw in our discussion of Fourier transforms, the Fourier transform of the convolution of two functions is proportional to the product of the Fourier transforms of the two convolved functions. The precise statements, with all 2π factors, are given in (3.63) and (3.64). Comparing with (4.33), we see that the solution to (4.33) will be given by

$$\phi(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathrm{i}\,x\,t}\,\frac{F(t)}{K(t)}\,dt\,, \tag{4.34}$$

where $F(t)$ and $K(t)$ are the Fourier transforms of $f(x)$ and $k(x)$:

$$F(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\mathrm{i}\,x\,t}\,f(x)\,dx\,,\qquad K(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\mathrm{i}\,x\,t}\,k(x)\,dx\,. \tag{4.35}$$

So provided that the necessary integrals can be evaluated, the solution for $\phi(x)$ can be obtained.

It is clear that a straightforward extension of this procedure allows us to solve the Fredholm equation of the second kind, again in the special case where the kernel is $k(x-t)$, and where the limits of the integration are $\pm\infty$. Fourier transforming the integral equation

$$\phi(x) = f(x) + \lambda\int_{-\infty}^{\infty} k(x-t)\,\phi(t)\,dt \tag{4.36}$$

and using the convolution theorem gives

$$\Phi(t) = F(t) + \lambda\,\sqrt{2\pi}\,K(t)\,\Phi(t)\,, \tag{4.37}$$

which can be solved for $\Phi(t)$ to give:

$$\Phi(t) = \frac{F(t)}{1-\lambda\,\sqrt{2\pi}\,K(t)}\,. \tag{4.38}$$

Finally, we take the inverse Fourier transform to get the solution as

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{F(t)}{1 - \lambda\sqrt{2\pi}\,K(t)}\, e^{-\mathrm{i}\,x\,t}\, dt\,. \tag{4.39}$$

A similar technique can be used to solve the Volterra equation of the second kind, in the special case where the kernel is of the form $k(x-t)$, and the lower limit of the integration is 0:

$$\phi(x) = f(x) + \lambda \int_0^x k(x-t)\,\phi(t)\, dt \tag{4.40}$$

The integral here can be recognised as the convolution integral (3.118) of the Laplace transform. Thus using (3.120) we now conclude that the solution for $\phi(x)$ is

$$\phi(x) = \frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\,\infty}^{\gamma+\mathrm{i}\,\infty} \frac{F(s)}{1-\lambda\,K(s)}\, e^{x\,s}\, ds\,, \tag{4.41}$$

where $F(s)$ and $K(s)$ are the Laplace transforms of $f(x)$ and $k(x)$. The integral in (4.41) is the Bromwich integral for the inverse Laplace transform, which we discussed in section 3.3. Recall that the real constant γ should be chosen so that the vertical contour of integration lies to the right of any singularities of the integrand. The solution for the Volterra equation of the first kind is easily derivable by this method too. Or, one can obtain it from (4.41) by noting from the original Volterra equations (4.3) and (4.4) that if we replace $f(x)$ by $-\lambda\, f(x)$ in (4.4), and then send $\lambda \longrightarrow \infty$, we obtain (4.3). Thus the solution to the Volterra equation of the first kind, for the kernel $k(x-t)$, will be

$$\phi(x) = \frac{1}{2\pi\,\mathrm{i}} \int_{\gamma-\mathrm{i}\,\infty}^{\gamma+\mathrm{i}\,\infty} \frac{F(s)}{K(s)}\, e^{x\,s}\, ds\,, \tag{4.42}$$

4.2 Degenerate Kernels

One might think from this title that we were about to stray off the topic of integral equations and undertake an investigation of improper goings-on in the Officers' Mess, but actually this will be a perfectly respectable analysis of a rather general technique for solving integral equations with a particular type of kernel function $K(x,t)$. In fact a less sensational-sounding and more descriptive terminology is *Separable Kernels*.

The idea is the following. Suppose the kernel function $K(x,t)$ in an integral equation is *separable*, in the sense that it can be written as a *finite* sum of N factorised terms:

$$K(x,t) = \sum_{j=1}^{N} M_j(x)\, N_j(t)\,. \tag{4.43}$$

A kernel $K(x,t)$ that was of the form of any polynomial in x and t would thus be of this degenerate type. So also would the kernel $\cos(x-t)$, since

$$\cos(x-t) = \cos x\, \cos t + \sin x\, \sin t\,. \tag{4.44}$$

Suppose we wish to solve a Fredholm equation of the second kind, for a degenerate kernel of the form (4.43). Substituting into (4.2) we obtain

$$\phi(x) = f(x) + \lambda \sum_{j=1}^{N} M_j(x) \int_a^b dt\, N_j(t)\, \phi(t)\,. \tag{4.45}$$

The integrals appearing here are just constants, say

$$c_j = \int_a^b dt\, N_j(t)\, \phi(t)\,, \tag{4.46}$$

and if we knew what they were we would have the solution for $\phi(x)$, since (4.45) gives

$$\phi(x) = f(x) + \lambda \sum_{j=1}^{N} c_j\, M_j(x)\,. \tag{4.47}$$

Of course we dont yet know what the constants c_i are, since they are given by the integrals (4.46) which themselves involve the unknown function $\phi(x)$. However, if we multiply (4.47) by $N_i(x)$ and integrate, we get

$$c_i = b_i + \lambda \sum_{j=1}^{N} A_{ij}\, c_j\,, \tag{4.48}$$

where we have also defined constants b_i and A_{ij} by

$$\begin{aligned} b_i &= \int_a^b dx\, N_i(x)\, f(x)\,, \\ A_{ij} &= \int_a^b dx\, N_i(x)\, M_j(x)\,. \end{aligned} \tag{4.49}$$

Now, since the constants b_i and A_{ij} are simply calculated as integrals of given functions, it follows that we can view (4.48) as a system of N simultaneous equations for the N unknowns c_i. In matrix notation, these equations are

$$\vec{c} = \vec{b} + \lambda\, \mathbf{A}\, \vec{c}, \tag{4.50}$$

or in other words

$$(\mathbb{1} - \lambda\, \mathbf{A})\, \vec{c} = \vec{b}\,. \tag{4.51}$$

This can be solved for $\vec{c}$ by inverting the matrix, to give

$$\vec{c} = (\mathbb{1} - \lambda\, \mathbf{A})^{-1}\, \vec{b}, \tag{4.52}$$

and so the problem is solved.

If the Fredholm equation is homogeneous, meaning $f(x) = 0$ and hence $\vec{b} = 0$, then (4.51) becomes

$$(\mathbb{1} - \lambda\, \mathbf{A})\, \vec{c} = 0\,, \tag{4.53}$$

which does not in general admit any non-zero solution for $\vec{c}$. The only way it can admit a solution is if the determinant of $(\mathbb{1} - \lambda\,\mathbf{A})$ should happen to vanish. This is because having a solution of (4.53) would imply that $\vec{c}$ was an eigenvector of $(\mathbb{1} - \lambda\,\mathbf{A})$ with zero eigenvalue. But the determinant of a matrix is equal to the product of its eigenvalues, and hence a zero eigenvalue means a zero determinant. Thus for a homogeneous Fredholm equation with a degenerate kernel to have a non-zero solution, it would have to be that

$$\det(\mathbb{1} - \lambda\,\mathbf{A}) = 0\,. \tag{4.54}$$

This is a standard eigenvalue equation, giving an N'th-order polynomial equation for the eigenvalues $1/\lambda$ of the matrix $\mathbf{A}$.

Let us consider an example. Suppose we wish to solve the homogeneous Fredholm equation

$$\phi(x) = \lambda \int_{-1}^{1} (x+t)\,\phi(t)\,dt\,. \tag{4.55}$$

The kernel is degenerate, with

$$M_1(x) = 1\,, \qquad M_2(x) = x\,, \qquad N_1(t) = t\,, \qquad N_2(t) = 1\,. \tag{4.56}$$

Simple integration gives $A_{11} = A_{22} = 0$, $A_{12} = 2/3$ and $A_{21} = 2$, or in other words

$$\mathbf{A} = \begin{pmatrix} 0 & \frac{2}{3} \\ 2 & 0 \end{pmatrix}\,. \tag{4.57}$$

The condition (4.54) for the vanishing of the determinant then implies

$$\begin{vmatrix} 1 & -\frac{2}{3}\lambda \\ -2\lambda & 1 \end{vmatrix} = 0\,. \tag{4.58}$$

One easily finds that this gives $1 - 4\lambda^2/3 = 0$, with solutions $\lambda_1 = \sqrt{3}/2$ and $\lambda_2 = -\sqrt{3}/2$, with the corresponding eigenvectors

$$\vec{c}_1 = \mu_1 \begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix}\,, \qquad \vec{c}_2 = \mu_2 \begin{pmatrix} 1 \\ -\sqrt{3} \end{pmatrix}\,, \tag{4.59}$$

where μ_1 and μ_2 are arbitrary constants. (One cannot expect these to be determined when solving a homogeneous equation.) Plugging these results back into (4.47), we get the solutions

$$\begin{aligned} \lambda = \tfrac{\sqrt{3}}{2}: &\qquad \phi(x) = \tfrac{1}{2}\sqrt{3}\,\mu_1\,(1 + \sqrt{3}\,x)\,, \\ \lambda = -\tfrac{\sqrt{3}}{2}: &\qquad \phi(x) = -\tfrac{1}{2}\sqrt{3}\,\mu_2\,(1 - \sqrt{3}\,x)\,. \end{aligned} \tag{4.60}$$

4.3 Neumann Series Solution of Integral Equations

Another method that can sometimes be useful for solving integral equations is the Neumann series expansion method. This can, in particular, be useful as a way of getting an approximate solution, up to the first few orders in an expansion parameter. The idea can be illustrated by considering an inhomogeneous Fredholm equation of the second kind:

$$\phi(x) = f(x) + \lambda \int_a^b dt\, K(x,t)\, \phi(t)\,. \tag{4.61}$$

The simplest way to describe the idea of the method is as follows. Let us suppose that λ can be thought of as a "small parameter." We may therefore say that as a leading-order approximation, the integral equation (4.61) is simply $\phi(x) \approx f(x)$. Let us write this leading-order result as

$$\phi_0 = f(x)\,. \tag{4.62}$$

Since λ is assumed small, we can then make a next-order approximation in which we use ϕ_0 in place of ϕ in the integral in (4.61), and get the next approximation to the true solution:

$$\phi_1(x) = f(x) + \lambda \int_a^b dt\, K(x,t)\, \phi_0(t)\,. \tag{4.63}$$

Since already have our expression for ϕ_0 as the known function $f(x)$, this means that everything on the right-hand-side of (4.63) is in principle calculable. The process can then be repeated again and again, and at each stage one uses the just-obtained approximation ϕ_n in the integral in (4.61) in order to get the next approximation ϕ_{n+1}:

$$\phi_{n+1}(x) = f(x) + \int_a^b dt\, K(x,t)\, \phi_n(t)\,. \tag{4.64}$$

It is helpful to express this in a slightly different way, as follows. Viewing λ as a parameter for keeping track of the order in the expansion, we may write

$$\phi_n(x) = \sum_{k=0}^{n} \lambda^k\, u_k(x)\,. \tag{4.65}$$

Substituting this into the original integral equation (4.61), and then equating order-by-order in λ we clearly obtain

$$\begin{aligned}
u_0(x) &= f(x)\,,\\
u_1(x) &= \int_a^b dt_1\, K(x,t_1)\, f(t_1)\,,\\
u_2(x) &= \int_a^b dt_2 \int_a^b dt_1\, K(x,t_1)\, K(t_1,t_2)\, f(t_2)\,,\\
&\cdots\\
u_n(x) &= \int_a^b dt_n \int_a^b dt_{n-1} \cdots \int_a^b dt_1\, K(x,t_1)\, K(t_1,t_2) \cdots K(t_{n-1},t_n)\, f(t_n)\,.
\end{aligned} \tag{4.66}$$

If we are lucky, the procedure described above will be a convergent one, and the solution to the original integral equation (4.61) will be given by

$$\phi(x) = \lim_{n\to\infty} \phi_n(x) = \sum_{k=0}^{\infty} \lambda^k\, u_k(x)\,. \qquad (4.67)$$

Of course in practice it might be that explictly performing the integrals (4.66) might get too difficult to do once n gets very big, and so we might well just stop after a few terms and view that as an approximate solution to the problem. But still, we should like to know that the series would in principle be convergent.

Testing for convergence is, of course, not going to be easy if we can't evaluate the integrals, but we can achieve something, at least, by making the traditional sort of "worst-case" estimates. Thus we may observe from (4.66) that we shall have

$$|\lambda^n\, u_n(x)| \le |\lambda^n|\; |f|_{\rm max}\; |K|^n_{\rm max}\; |b-a|^n\,. \qquad (4.68)$$

Here, $|f|_{\rm max}$ means the maximum value of $|f(x)|$ in the interval $a \le x \le b$, and $|K|_{\rm max}$ means the maximum value of $|K(x,t)|$ that it achieves anywhere in the ranges taken by x and t. By Cauchy's ratio test we can certainly therefore be sure of convergence if

$$|\lambda|\; |K|_{\rm max}\; |b-a| < 1\,. \qquad (4.69)$$

One can view this as a condition on the smallness of the parameter λ that is needed for convergence. Of course if this condition is not satisfied it may still be that the series is convergent, since we made some pretty drastic worst-case assumptions in getting to (4.68).

Let us look at an example. Consider the following inhomogeneous Fredholm equation of the second kind:

$$\phi(x) = x + \lambda \int_{-1}^{1} dt\, (t-x)\, \phi(t)\,. \qquad (4.70)$$

For the leading approximation we have $\phi_0(x) = x$, and plugging this into the integral in (4.70) we then get

$$\phi_1(x) = x + \lambda \int_{-1}^{1} dt\, (t-x)\, t = x + \tfrac{2}{3}\,\lambda\,. \qquad (4.71)$$

Using this to calculate $\phi_2(x)$, and then this for $\phi_3(x)$ gives

$$\begin{aligned} \phi_2(x) &= \tfrac{2}{3}\lambda + (1 - \tfrac{4}{3}\lambda^2)\, x\,, \\ \phi_3(x) &= \tfrac{2}{3}\lambda\, (1 - \tfrac{4}{3}\lambda^2) + (1 - \tfrac{4}{3}\lambda^2)\, x\,. \end{aligned} \qquad (4.72)$$

Clearly we only ever generate x to the powers 0 and 1 in each iteration, so we can usefully simply the discussion by making the definition

$$\phi_n(x) = a_n + b_n\, x\,, \qquad (4.73)$$

where a_n and b_n are constants. Substituting this into

$$\phi_n(x) = x + \lambda \int_{-1}^{1} dt\,(t-x)\,\phi_{n-1}(t)\,, \tag{4.74}$$

we easily get

$$a_n = \tfrac{2}{3}\lambda\, b_{n-1}\,, \qquad b_n = 1 - 2\lambda\, a_n\,. \tag{4.75}$$

From this we can see that

$$a_n = \tfrac{2}{3}\lambda\,(1 - 2\lambda\, a_{n-2})\,, \qquad b_n = 1 - \tfrac{4}{3}\,\lambda^2\, b_{n-2}\,. \tag{4.76}$$

It is actually nicer at this point to define a new eigenvalue μ instead of λ, related by

$$\lambda = \frac{\sqrt{3}}{2}\,\mu\,, \tag{4.77}$$

so that we have

$$a_n = \tfrac{\mu}{\sqrt{3}} - \mu^2\, a_{n-2}\,, \qquad b_n = 1 - \mu^2\, b_{n-2}\,. \tag{4.78}$$

It is then easy to show by induction that

$$\begin{aligned} a_{2p} &= a_{2p-1} = \tfrac{1}{\sqrt{3}}\Big(1 - \mu^2 + \mu^4 - \mu^6 + \cdots - (-1)^p\,\mu^{2(p-1)}\Big)\,, \qquad p \ge 1\,,\\ b_{2p-2} &= b_{2p-1} = 1 - \mu^2 + \mu^4 - \mu^6 + \cdots - (-1)^p\,\mu^{2(p-1)}\,, \qquad p \ge 1\,, \end{aligned} \tag{4.79}$$

with $a_0 = 0$. The first few examples are

$$\begin{aligned} &a_0 = 0\,, \quad a_1 = a_2 = \tfrac{\mu}{\sqrt{3}}\,, \quad a_3 = a_4 = \tfrac{\mu}{\sqrt{3}}(1-\mu^2)\,, \quad a_5 = a_6 = \tfrac{\mu}{\sqrt{3}}(1 - \mu^2 + \mu^4)\,,\\ &b_0 = b_1 = 1\,, \quad b_2 = b_3 = 1 - \mu^2\,, \quad b_4 = b_5 = 1 - \mu^2 + \mu^4\,, \end{aligned} \tag{4.80}$$

and so on.

The final solution $\phi(x)$ to our equation (4.70) is obtained by taking the limit where n goes to infinity, so that $\phi(x) = a + b\,x$ where

$$a = \lim_{n\to\infty} a_n = \tfrac{\mu}{\sqrt{3}} \sum_{m=0}^{\infty} (-1)^m\,\mu^{2m}\,, \qquad b = \lim_{n\to\infty} b_n = \sum_{m=0}^{\infty} (-1)^m\,\mu^{2m}\,. \tag{4.81}$$

Clearly these sums converge if $\mu^2 < 1$, and they diverge if $\mu^2 > 1$, so in this case the Neumann series solution is convergent for

$$|\lambda| < \frac{\sqrt{3}}{2}\,. \tag{4.82}$$

Actually, we can do rather better here, since the infinite series in (4.81) is geometric, and therefore explicitly summable:

$$\sum_{m=0}^{\infty} (-1)^m\,\mu^{2m} = \frac{1}{1+\mu^2}\,. \tag{4.83}$$

This gives us the final solution

$$\phi(x) = \frac{\mu}{\sqrt{3}\,(1+\mu^2)} + \frac{x}{1+\mu^2}\,. \tag{4.84}$$

After rewriting in terms of λ again, this is

$$\phi(x) = \frac{2\lambda}{3+4\lambda^2} + \frac{3x}{3+4\lambda^2}\,. \tag{4.85}$$

In fact we have been lucky here, since now as a result of summing the infinite series, we have achieved an analytic continuation of the Neumann series solution, which is now valid for all λ except $\lambda = \pm \mathrm{i}$. It is easy to verify, by direct substitution, that (4.85) solves[18] the original integral equation (4.70) for all values of λ.

The same general idea of solving by the Neumann series methods can also be applied to integral equations the Volterra type. To illustrate this, let us take an integral equation that looks very like our previous example (4.70), except that now we take the integration limit to involve x:

$$\phi(x) = x + \lambda \int_0^x dt\,(t-x)\,\phi(t)\,. \tag{4.87}$$

Again, we think of λ as an order parameter, and thus we have the leading-order solution $\phi_0 = x$. Substituting this into the integral on the right-hand side gives us the next approximation

$$\phi_1(x) = x + \lambda \int_0^x dt\,(t-x)\,t = x - \lambda\,\frac{x^3}{6}\,. \tag{4.88}$$

Substituting this again, we get

$$\phi_2(x) = x + \lambda \int_0^x dt\,(t-x)\left(t - \lambda\,\frac{t^3}{6}\right) = x - \lambda\,\frac{x^3}{6} + \lambda^2\,\frac{x^5}{120}\,. \tag{4.89}$$

One further step yields

$$\phi_3(x) = x - \lambda\,\frac{x^3}{6} + \lambda^2\,\frac{x^5}{120} - \lambda^3\,\frac{x^7}{5040}\,. \tag{4.90}$$

It is pretty clear where this is leading:

$$\phi_n(x) = \lambda^{-1/2} \sum_{r=0}^{n} (-1)^r\,\frac{(\lambda^{1/2}\,x)^{2r+1}}{(2r+1)!}\,. \tag{4.91}$$

[18] Actually, of course, we could have solved this even more simply without ever using a series solution. At the stage where we observed that $\phi_n(x)$ was of the form (4.73) we could have seen that this would continue to be true in the limit where n tends to infinity. Thus we could simply have substituted the trial solution $\phi(x) = a + b\,x$ into (4.70), and solved the two algebraic equations result from separately equating the terms of orde 0 and 1 in x, namely

$$a = \tfrac{2}{3}\lambda\,b\,, \qquad b = 1 - 2\lambda\,a\,. \tag{4.86}$$

This directly gives the same result as (4.85). Bear in mind, therefore, that (4.70) is really a rather trivial toy example that we are considering just to illustrate a few of the general methods that have been discussed.

In the limit as n tends to infinity we get the complete solution

$$\begin{aligned}\phi(x) &= \lim_{n\to\infty} \phi_n(x) = \lambda^{-1/2} \sum_{r=0}^{\infty} (-1)^r \, \frac{(\lambda^{1/2}\, x)^{2r+1}}{(2r+1)!} \\ &= \lambda^{-1/2} \sin(\lambda^{1/2}\, x)\,. \end{aligned} \tag{4.92}$$

We could, of course, quite easily set up an iterative scheme to *derive* this rigorously, rather than simply observing the trend from the first few terms in the series. If we did so, there would be no surprises or subtleties, and we would rather quickly get the result in a deductive way. Alternatively, we can just substitute (4.92) back into the integral equation (4.87), and verify that it is indeed a solution. Since it is obvious from the Neumann series approach that at each stage in the iteration we get a *specific* and unique result for ϕ_n, there can only be one possible final answer and so if we find that our proposed solution indeed solves the integral equation then we know that it is the unique answer.

Notice, by the way, that (4.87) with $\lambda = 1$ is precisely the integral equation that we produced a while backi in (4.22), by integrating the simple harmonic oscillator equation $y'' + y = 0$, subject to the boundary conditions $y(0) = 0$ and $y'(0) = 1$. It is worth emphasising again that when we solved the integral equations (4.70) and (4.87) above we got *unique* answers in each case. This illustrates the point made earlier, about how the boundary conditions are built into the integral equation. Notice also that these two examples show us that the solution is radically different for a Volterra equation, as compared with a Fredholm equation with a very similar structure.

5 Conformal Mappings

5.1 Introduction

At this stage in the course we revert to a topic that is concerned directly with complex analysis. Recall that if we have an analytic function

$$w(z) = u(x,y) + \mathrm{i}\, v(x,y)\,, \tag{5.1}$$

where $z = x + \mathrm{i}\, y$ is a complex variable, then the real and imaginary parts $u(x,y)$ and $v(x,y)$ satisfy the Cauchy-Riemann equations,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\,, \qquad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}\,. \tag{5.2}$$

An equivalent, but more elegant, statement of the same thing is

$$\frac{\partial w}{\partial \bar z} = 0\,, \tag{5.3}$$

where we are treating $z = x + \mathrm{i}\, y$ and $\bar{z} = x - \mathrm{i}\, y$ as independent variables here[19]

$$\frac{\partial}{\partial z} = \frac{1}{2}\frac{\partial}{\partial x} + \frac{1}{2\mathrm{i}}\frac{\partial}{\partial y}\,, \qquad \frac{\partial}{\partial \bar{z}} = \frac{1}{2}\frac{\partial}{\partial x} - \frac{1}{2\mathrm{i}}\frac{\partial}{\partial y}\,. \tag{5.5}$$

Thus if $w(z)$ is analytic in some region, then it depends only on z but not on $\bar{z}$ in that region.

We can view the function $w(z)$ as a mapping from the complex z-plane into the complex w-plane. This mapping has some very important properties. The first of these is that it preserves angles. To see what is mean by this, we need to consider a pair of lines in the z-plane, which intersect each other at some point, at a certain angle. As we trace along the path of one of these lines in the z-plane, we shall find that an image of this path is traced out in the w-plane. If we look at the images of the two intersecting paths in the z-plane, we get two intersecting paths in the w-plane. The statement about the preservation of angles is that the angle between the intersecting paths in the z-plane is equal to the angle between the intersecting paths in the w-plane.

To show this, let us suppose that the two lines in the z-plane intersect at $z = a$. Let us refer to these two lines as Path 1 and Path 2. Points on Path 1 near to $z = a$ must clearly lie approximately on a straight line (any well-behaved path looks straight if a short enough segment is examined), and so we can say that points on Path 1 near to $z = a$ are characterised by

$$dz_1 = |dz_1|\, e^{\mathrm{i}\,\theta_1}\,, \tag{5.6}$$

where θ_1 measures the angle that Path 1 makes with the real axis. Likewise, near to $z = a$ points on Path 2 will be such that

$$dz_2 = |dz_1|\, e^{\mathrm{i}\,\theta_2}\,. \tag{5.7}$$

[19]One might feel uneasy about this, since we know that $\bar{z}$ is *not* independent of z! The best way to clarify what is going on is to think initially of writing $x - \mathrm{i}\, y$ as $\tilde{z}$, and not yet to assume that x and y are real. It is now clear that the equations $z = x + \mathrm{i}\, y$, $\tilde{z} = x - \mathrm{i}\, y$ give a perfectly legitimate mapping from the complex variables (x, y) to the complex variables $(z, \tilde{z})$, and so the equations

$$\frac{\partial}{\partial z} = \frac{1}{2}\frac{\partial}{\partial x} + \frac{1}{2\mathrm{i}}\frac{\partial}{\partial y}\,, \qquad \frac{\partial}{\partial \tilde{z}} = \frac{1}{2}\frac{\partial}{\partial x} - \frac{1}{2\mathrm{i}}\frac{\partial}{\partial y}\,. \tag{5.4}$$

make perfect sense. Then, at the end of the day in any calculation, we finally replace $\tilde{z}$ by $\bar{z}$ (the complex conjugate of z), which amounts to choosing the "real section" where x and y are real. Having been through this argument we can then see that in fact we can be impatient and not bother to wait untile the end of the day before setting $\tilde{z} = \bar{z}$; we can just use $\bar{z}$ right from the beginning, and keep at the back of our minds what it is that it really means. (If you weren't confused about this point before reading this footnote, it would probably have been better if you hadn't read it!)

Tha angle between the two paths is clearly $\theta_2 - \theta_1$.

Now, we consider the mapping into the complex w-plane. We shall have

$$dw = \frac{dw}{dz}\, dz\,, \tag{5.8}$$

Now a crucial property of the derivative dw/dz of an analytic function is that at a given point z it is *independent* of the direction of dz. (This is a standard result, which was proved in Part 1 of the course.) Therefore if we write $dw/dz = |dw/dz|\, e^{\mathrm{i}\,\alpha}$ at $z = a$, we shall have

$$dw = |dw/dz|\, e^{\mathrm{i}\,\alpha}\, dz \tag{5.9}$$

at $z = a$, *independent of the angle of* dz. Thus the images of our two paths in the w-plane, which intersect at $w(a)$, will be characterised at nearby points by

$$dw_1 = |dw/dz|\, |dz_1|\, e^{\mathrm{i}\,(\alpha+\theta_1)}\,, \qquad dw_2 = |dw/dz|\, |dz_2|\, e^{\mathrm{i}\,(\alpha+\theta_2)}\,. \tag{5.10}$$

Thus the angle between the two image paths in the w-plane is clearly therefore $(\alpha + \theta_2) - (\alpha + \theta_1) = \theta_2 - \theta_1$. This is the same as the angle between the original paths in the z-plane, and so the result is established.

Another important point is that not only the angles but also the *shapes* of infinitesimal figures in the z-plane are mapped into the same angles and shapes in the w-plane. To understand this, we have to think about how to measure infinitesimal separations in the complex plane. In the z-plane, Pythagoras' Theorem tells us that the distance ds between to infinitesimally separated points (x, y) and $(x + dx, y + dy)$ is given by

$$ds^2 = dx^2 + dy^2\,, \tag{5.11}$$

which can be written also as

$$ds^2 = dz\, d\bar{z} = |dz|^2\,. \tag{5.12}$$

The quantity ds^2 is called the *metric* on the complex z-plane. Similarly, in the complex w-plane we have a metric $d\hat{s}^2$, given by

$$d\hat{s}^2 = du^2 + dv^2 = dw\, d\bar{w} = |dw|^2\,. \tag{5.13}$$

In view of the fact that $dw = (dw/dz)\, dz$, and that if $w(z)$ is analytic at z then dw/dz has an unambiguous meaning independent of the direction of dz, we see that there is a simple relation between the metrics in the w-plane and the z-plane:

$$d\hat{s}^2 = \Big|\frac{dw}{dz}\Big|^2\, ds^2\,. \tag{5.14}$$

This equation in fact summarises all the properties of the mapping between the z-plane and the image in the w-plane. There is an overall scale factor $|dw/dz|$, but aside from that, infinitesimal distances all map over in the same way. So we have established that an infinitesimal figure in the z-plane is mapped into a *similar* figure in the w-plane, with all relative angles, and ratios of lengths, preserved. An infinitesimal object in the z-plane maps into one that looks exactly the same in the w-plane, up to some overall rotation and scaling. This is what is meant by a *conformal mapping*, or *conformal transformation.*

5.2 Two-dimensional Laplace Equation

An important application of conformal mappings is for solving Laplace's equation in two dimensions. Situations where this problem arises include solving for electrostatic potentials in two dimensions, and solving hydrodynamical equations in two dimensions. Of course such problems might not only arise by considering two dimensions in its own right; they can also arise if one has a three-dimensional configuration that has a translational invariance along one axis (for example, and infinite cylinder lying along the z-axis). It turns out that the methods of conformal mapping can be an extremely powerful tool.

To understand this, consider a potential $\psi(x,y)$ that satisfies Laplace's equation in two dimensions:

$$\nabla^2\,\psi \equiv \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial x^2} = 0\,. \tag{5.15}$$

Note that from (5.5) we have

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial z} + \frac{\partial}{\partial \bar z}\,, \qquad \frac{\partial}{\partial y} = \mathrm{i}\Big(\frac{\partial}{\partial z} - \frac{\partial}{\partial \bar z}\Big)\,, \tag{5.16}$$

and so we can also write the Laplacian as

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x^2} = 4\frac{\partial^2}{\partial z\,\partial \bar z}\,. \tag{5.17}$$

Now let us see what happens if we map into the complex w-plane In the w-plane we may consider a function $\Psi(u,v)$ which is simply the image of the function $\psi(x,y)$ in the z-plane:

$$\Psi(u,v) = \Psi(u(x,y), v(x,y)) = \psi(x,y)\,. \tag{5.18}$$

What we shall now show is that if $\psi(x,y)$ satisfies Laplace's equation in the z-plane, then $\Psi(u,v)$ satisfies Laplace's equation in the w-plane. To see this, we note that

$$\frac{\partial}{\partial z} = \frac{\partial w}{\partial z}\,\frac{\partial}{\partial w}\,. \tag{5.19}$$

Notice that there is no term $(\partial\bar{w}/\partial z)\,\partial/\partial\bar{w}$ here because we are assuming that $w(z)$ is analytic. By the same token, we shall have

$$\frac{\partial}{\partial\bar{z}} = \frac{\partial\bar{w}}{\partial\bar{z}}\,\frac{\partial}{\partial\bar{w}}\,. \tag{5.20}$$

Furthermore, we also have

$$\frac{\partial^2}{\partial z\,\partial\bar{z}} = \Big|\frac{\partial w}{\partial z}\Big|^2\,\frac{\partial^2}{\partial w\,\partial\bar{w}}\,. \tag{5.21}$$

The crucial point here is that for the same reason of analyticity of $w(z)$, we don't pick up any "extra" term where the $\partial/\partial z$ derivative lands on the $(\partial\bar{w}/\partial\bar{z})$ factor in (5.20). So we see that the Laplacians ∇^2 and $\hat{\nabla}^2$ in the z-plane and w-plane respectively, which are given by

$$\nabla^2 = 4\frac{\partial^2}{\partial z\,\partial\bar{z}}\,,\qquad \hat{\nabla}^2 = 4\frac{\partial^2}{\partial w\,\partial\bar{w}}\,, \tag{5.22}$$

are related by

$$\nabla^2 = \Big|\frac{\partial w}{\partial z}\Big|^2\,\hat{\nabla}^2\,. \tag{5.23}$$

In particular, if $\psi(x,y)$ satisfies $\nabla^2\,\psi = 0$ in the z-pane, then the $\Psi(u,v)$, the image of $\psi(x,y)$ in the w-plane as in (5.18), satisfies $\hat{\nabla}^2\,\Psi = 0$.

The upshot of this discussion is that we now have a nice way of solving two-dimensional potential-theory problems at our disposal. Namely, if we can solve Laplace's equation subject to certain boundary conditions in one particular "conformal frame," (say the z-plane), then we immediately know that after making a conformal mapping to the $w(z)$ plane, the same potential will be a solution of Laplace's equation in the w-plane. Clearly the original boundary conditions on $\psi(x,y)$ will map over into "image" boundary conditions on $\Psi(u,v) = \psi(x,y)$. For example, if $\psi(x,y)$ vanishes on a certain curve in the z-plane, then $\Psi(u,v)$ will vanish on the image curve in the w-plane. Of course the idea is that we choose our conformal mapping judiciously, to transform a difficult problem into an easier one.

Let us consider an example. Suppose we wish to solve for the two-dimensional electrostatic potential for the following situation. There is a conductor lying along the entire y axis, at $x = 0$, and circular conductor of radius R, centred on $(x,y) = (d,0)$. The infinite line is held at zero potential, and the circle is held potential ψ_0. The problem is to find the potential everywhere in the region $x \ge 0$, outside the circular conductor, by using the conformal mapping technique.

The whole art of solving problems like this is to spot the right conformal transformation that maps the original problem into a simpler one. In this case, fortunately, an artist has been here before us, and so we are invited to contemplate the following transformation:

$$z = a \tanh \frac{\mathrm{i}w}{2}, \tag{5.24}$$

where a is a constant. Of course it would actually be the inverse of this transformation that gave us w as a function of z. Writing $w = u + \mathrm{i}\, v$, some simple t(h)rigonometric manipulations lead us to

$$x = -\frac{a \sinh v}{\cosh v + \cos u}, \qquad y = \frac{a \sin u}{\cosh v + \cos u}. \tag{5.25}$$

Thus if we look at the y-axis, $x = 0$, we see that it corresponds to taking $v = 0$, with u ranging from $-\pi$ to π as y ranges from $-\infty$ to ∞. So we have found the image of the infinite line conductor.

Now, consider what happens if we eliminate u from the equations (5.25). We do this by first noting that we have

$$\begin{aligned} \cos u &= -\Big(\frac{a}{x} \sinh v + \cosh v\Big), \\ \sin u &= \frac{y}{a} (\cosh v + \cos u) = -\frac{y}{x} \sinh v. \end{aligned} \tag{5.26}$$

Using $\cos^2 u + \sin^2 u = 1$, we therefore get

$$(\frac{a}{x} \sinh v + \cosh v\Big)^2 + \frac{y^2}{x^2} \sinh^2 v = 1, \tag{5.27}$$

which then can be rearranged as

$$(x + a \coth v)^2 + y^2 = \frac{a^2}{\sinh^2 v}. \tag{5.28}$$

Thus we see that at fixed v we have a circle of radius $|a/\sinh v|$, centred on the point $(x, y) = (-a \coth v, 0)$ in the z-plane. This is exactly what we want, if we choose a, and the fixed value v_0 for v, such that

$$d = -a \coth v_0, \qquad R = -\frac{a}{\sinh v_0}. \tag{5.29}$$

It is easy to see that as u ranges from $-\pi$ to π at this fixed value $v = v_0$, the image in the z-plane traces out the points on the circle of radius R, centred on $(x, y) = (d, 0)$ in the z-plane. This is shown in the figure below.

We have succeeded in mapping the geometry of the original problem into a considerably simpler one; the original infinite line and circular conductors have become the two line

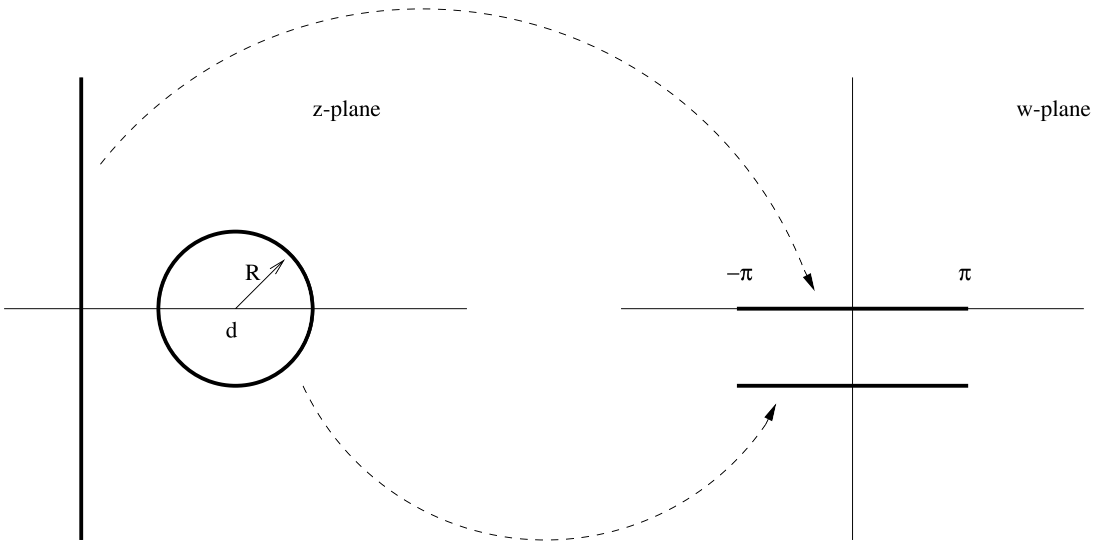

Figure 17: The line and circle in the z-plane are mapped to two parallel line segments in the w-plane.

segments $v = 0$ and $v = v_0$, with u in the range $-\pi \le u \le \pi$ to cover each conductor. Furthermore, it is easy to check that the region between these two line segments in the w-plane maps into the region between the two conductors in the z-plane.

In fact luckily, we can think of extending the line segments to the entire range $-\infty \le u \le \infty$ in the w-plane, since x and y are periodic in u and so as u traverses the entire real line we just get multiple coverings of the two conductors. This is an important point, because it now means that we merely have to solve Laplace's equation between the two infinitely-long parallel "plates" at $v = 0$ and $v = v_0$ in the w-plane. Since our boundary conditions are that $\Psi(u, v) = 0$ on the conductor at $v = 0$, and $\Psi(u, v) = \psi_0$ on the conductor at $v = v_0$, it follows that the solution everywhere between the parallel plates in the w-plane is

$$\Psi(u, v) = \frac{v}{v_0}\, \psi_0\,. \tag{5.30}$$

It only remains to express the potential (5.30) back in terms of the (x, y) coordinates, in order to obtain the required solution for the potential in the z-plane. From (5.24) we have

$$w = -2\,\mathrm{i}\, \mathrm{arctanh}\Big(\frac{z}{a}\Big)\,, \tag{5.31}$$

and so v is given by taking the imaginary part of this. Thus we arrive at the solution for the potential in terms of x and y:

$$\psi(x, y) = -\frac{2\psi_0}{v_0}\, \mathrm{Re}\Big[\mathrm{arctanh}\Big(\frac{z}{a}\Big)\Big]\,. \tag{5.32}$$

Finally, we may note that since the equipotentials in the w-plane are clearly simply given by $v =$ constant, it follows that in the original z-plane the equipotentials are the circles defined at fixed v by equation (5.28). (The "circle" corresponding to $v = 0$ has in fact blown up to become the y-axis.)

5.3 Schwarz-Christoffel Transformation

It should be clear from the previous discussion that solving a potential theory problem in two dimensions can become rather simple, *if* one is able to find a conformal transformation that maps the geometry of the original problem into a nicer one, where Laplace's equation can be easily solved. Of course the key word in the last sentence is "if." It is not easy to give general prescriptions for how to find the required transformation, and at times the procedure can seem more like an art than a science. There is one class of geometries, however, for which a general prescription *can* be given. Namely, we can construct general formulae for mapping an N-sided polygon in the z-plane onto the real axis of the w-plane.

An alarm-bell might perhaps start ringing at this point. At the beginning of our discussion of conformal transformations much was made of the fact that they are angle-preserving. Now, we are proposing to "unwrap" a polygon and lay it out flat along the real axis; what is going on? There is, in fact, no paradox here. The crucial property that guaranteed the angle-preserving nature of the conformal transformation was that the mapping $w(z)$ was assumed to be *analytic*. Clearly, therefor, if we are to map a polygon into a line, the function $w(z)$ that does the job must have singularities at the vertices of the polygon. We shall now proceed to see how to construct this function, known as the *Schwarz-Christoffel* transformation.

Consider first what happens if we have a function $w(z)$ such that

$$\frac{dz}{dw} = A\,(w - w_0)^{-\sigma_0}\,, \tag{5.33}$$

where A is a complex constant, σ_0 is a real constant, and w_0 is a real constant specifying a point on the real axis in the w-plane. Let us investigate what happens as w is allowed to range along the real axis in the w-plane. Since σ_0 is not in general an integer, we must make a definition about where to place the branch cut. When $w > w_0$, we define the phase, or argument, of $(w - w_0)^{-\sigma_0}$ to be 0.

When w becomes less than w_0, we imagine that it detours in a little semi-circle around w_0 that takes it *above* the real axis, which implies that the argument of $(w - w_0)$ will be

$-\pi\,\sigma_0$ when $w < w_0$. Thus we have

$$\arg\frac{dz}{dw} = \begin{cases} \arg A - \pi\,\sigma_0\,, & w < w_0 \\ \arg A\,, & w > w_0 \end{cases} \tag{5.34}$$

Now, let us consider what happens as w increases along the real axis. At all points, if w advances by an infinitesimal amount dw, we shall have $\arg dw = 0$, since dw is a real quantity, and so from (5.33) and (5.34) it follows that we must have

$$\arg dz = \begin{cases} \arg A - \pi\,\sigma_0\,, & w < w_0 \\ \arg A\,, & w > w_0 \end{cases} \tag{5.35}$$

Thus we see that as w approaches w_0 from the left, a straight-line path in the z-plane is traced out, at an angle given by $\arg A - \pi\,\sigma_0$. After w has advanced to the right past w_0, a straight-line path is again being traced out in the z-plane, bu now at an angle given by $\arg A$. In other words, the total path in the z-plane consists of a straight-line segment, then a sharp turn *to the left* by an angle $\pi\,\sigma_0$, and then another straight-line segment going off at this new angle.

We now generalise the above construction, by choosing $w(z)$ to be such that

$$\frac{dz}{dw} = A\,(w-w_0)^{-\sigma_0}\,(w-w_1)^{-\sigma_1}\cdots(w-w_n)^{-\sigma_n}\,. \tag{5.36}$$

This will map the real axis of the w-plane into a sequence of straight-line segments L_i in the z-plane, each successive line segment swinging round to the left by an angle $\pi\,\sigma_i$ relative to the previous one. If we choose the exponents σ_i to be such that

$$\sum_{i=0}^{n}\sigma_i = 2\,, \tag{5.37}$$

then the sum total of all the left-turning angle changes will be 2π, and so provided we choose the starting and finishing values of w appropriately, will shall have nicely constructed a closed polygon,[20] since the sum of the interior angles will be 2π. (See figure below.) All that remains is to integrate (5.36), and to choose the various constants in the construction appropriately, so as to describe the desired polygon in the complex z-plane.[21] Notice that since the corners in the polygon twist round to the left as we move along the real w axis in the direction of increasing w, the interior of the polygon is corresponds to the region *above* the real axis in the complex w-plane.

[20]Note that we are not obliged to construct a closed polygon. In fact, it is quite common that one uses a Schwarz-Christoffel transformation to construct an open geometry with angles, such as a U-shaped channel.

[21]Of course there is also the little matter of inverting the resulting expression for $z(w)$ that one obtains by this means, in order to express w as a function of z. Recall from our example in the previous section that we eventually need to know $w(z)$, since the potential is easily solved for in the w-plane, and must now be re-expressed in terms of the z variable.

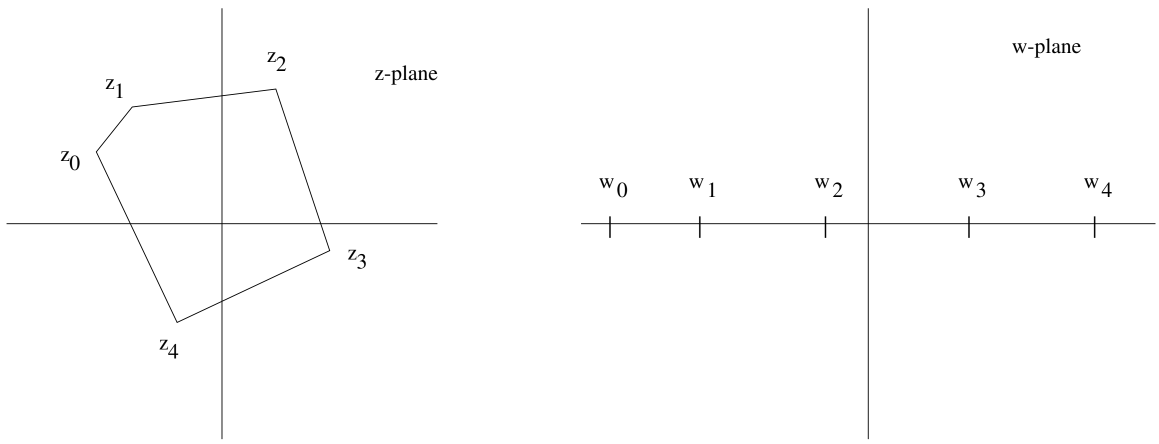

Figure 18: The Schwarz-Christoffel transformation.

To see how the choice of constants will work, let us perform a counting of parameters. We specify our N-sided polygon in the z-plane by specifying the location of its N vertices z_i (so we have $n = N-1$, in terms of the integer n appearing in (5.36)). Each of these is a complex number, so there are $2N$ real parameters needed here. After integrating (5.36) we shall have

$$z(w) = z_0 + A \int^w dt\,(t-w_0)^{-\sigma_0}\,(t-w_1)^{-\sigma_1}\cdots(t-w_n)^{-\sigma_n}\,, \tag{5.38}$$

where z_0 is the (complex) constant of integration. Thus we have at our disposal N real parameters w_i, a further $(N-1)$ real parameters from σ_i (recalling that we have the single real constraint (5.37)), and 2 real parameters each from A and z_0. In total, therefore, we have $2N+3$ real parameters available, and we need only $2N$ in order to match up with our required polgygon in the z-plane. This means that three of the locations w_i can in fact be chosen arbitrarily, and then the rest of the parameters will be uniquely determined. Usually, one chooses three of the w_i so as to make life as simple as possible, from the point of view of making the evaluation of the integral (5.38) as straightforward as possible.

Commonly, one of the transformed points w_i is chosen to be at infinity. Let us therefore take $w_0 = \infty$. If we send w_0 to infinity, after first rescaling the constant A by the factor $(-w_0)^{\sigma_0}$, then clearly (5.38) becomes

$$z(w) = z_0 + A \int^w dt\,(t-w_1)^{-\sigma_1}\,(t-w_2)^{-\sigma_2}\cdots(t-w_n)^{-\sigma_n}\,. \tag{5.39}$$

Let us consider some examples. Actually, there are not really that many examples one can easily consider explicitly, because if there are too many factors in the integrand in (5.38) or (5.39) the integral becomes difficult or impossible to evaluate. For example, already if we take (5.39) with two generic factors only, we have quite a complicated result:

$$\begin{aligned} z(w) &= z_0 + A \int^w dt\,(t-w_1)^{-\sigma_1}\,(t-w_2)^{-\sigma_2}\,, \\ &= z_0 + A'\,(w-w_2)^{1-\sigma_2}\,{}_2F_1\Big(1-\sigma_2,\sigma_1;2-\sigma_2;\frac{w-w_2}{w_1-w_2}\Big)\,. \end{aligned} \tag{5.40}$$

The cases that lead to elementary functions are degenerate triangles and rectangles.

Consider first the example of an infinite U-shaped channel, formed by the lines $x=0$ to $x=\infty$ at $y=0$ and at $y=h$, together with the line $y=0$ to $y=h$ at $x=0$. Suppose that we are interested in solving Laplace's equation inside this channel, and thus we should like to map the geometry into a simpler one. The idea here will be to "unwrap" the U-shaped channel, so that it ends up flattened out along the real axis in the w-plane.

If you imagine coming in along the semi-infinite line at $y=h$, from $x=\infty$ down to $x=0$, the channel then makes a 90-degree left turn at $(x,y)=(0,h)$. It then makes another 90-degree left turn at $(x,y)=(0,0)$, before heading out to the east again along the real axis. Thus we have $\sigma_1=\frac{1}{2}$ and $\sigma_2=\frac{1}{2}$, and from (5.39) the required transformation is

$$z(w) = z_0 + A\int^w dt\,(t-w_1)^{-\frac{1}{2}}\,(t-w_2)^{-\frac{1}{2}}\,. \tag{5.41}$$

Effectively, we are taking a degenerate triangle, with an exterior angle of π at the third vertex located at $z=\infty$.

It is convenient to make a symmetrical choice $w_1=-1$, $w_2=1$ here, and so the integral becomes

$$z(w) = z_0 + A\int^w \frac{dt}{\sqrt{t^2-1}} = z_0 + A\,\mathrm{arcosh}\,w\,. \tag{5.42}$$

We shall want the vertex at $z=0$ to correspond to $w=1$, so

$$0 = z_0 + A\,\mathrm{arcosh}\,1 = z_0\,, \tag{5.43}$$

while the vertex at $z=\mathrm{i}\,h$ must be at $w=-1$, and so

$$\mathrm{i}\,h = A\,\mathrm{arcosh}\,(-1) = A\,\mathrm{i}\,\pi\,. \tag{5.44}$$

Thus the conformal mapping for this problem is

$$z = \frac{h}{\pi}\,\mathrm{arcosh}\,w\,, \tag{5.45}$$

which, luckily, is easily inverted to give

$$w = \cosh\Big(\frac{\pi\,z}{h}\Big)\,. \tag{5.46}$$

It is easy to check that the real axis in the w-plane has indeed been mapped onto the U-shaped channel in the z-plane. The mapping is as follows:

$$\begin{array}{rll} -\infty \le w \le -1 & \text{maps to} & z = \infty + \mathrm{i}\, h \longrightarrow z = \mathrm{i} h\,, \\ -1 \le w \le 1 & \text{maps to} & z = \mathrm{i}\, h \longrightarrow z = 0 \\ 1 \le w \le \infty & \text{maps to} & z = 0 \longrightarrow z = \infty\,. \end{array} \tag{5.47}$$

This is depicted in the figure below. Furthermore, it is also easy to see that points in the upper-half w-plane map into the interior region of the channel in the z-plane. If we take

$$z = x + \frac{\mathrm{i}\, h\, \theta}{\pi}\,, \tag{5.48}$$

then (5.46) gives

$$w = \cosh\Big(\frac{x\,\pi}{h} + \mathrm{i}\,\theta\Big) = \cosh\Big(\frac{x\,\pi}{h}\Big)\, \cos\theta + \mathrm{i}\, \sinh\Big(\frac{x\,\pi}{h}\Big)\, \sin\theta\,. \tag{5.49}$$

The phase ϕ of w is therefore given by

$$\tan\phi = \tanh\Big(\frac{x\,\pi}{h}\Big)\, \tan\theta\,, \tag{5.50}$$

implying that as θ goes from 0 to π (corresponding to increasing the y value inside the channel), the phase in the w plane increases from 0 to π. For example at $\theta = \frac{1}{2}\pi$, corresponding to sitting on the line at $y = \frac{1}{2}h$ along the middle of the channel, we find $\phi = \frac{1}{2}\pi$. Thus the positive imaginary axis of the w plane maps onto the line running up the middle of the channel.

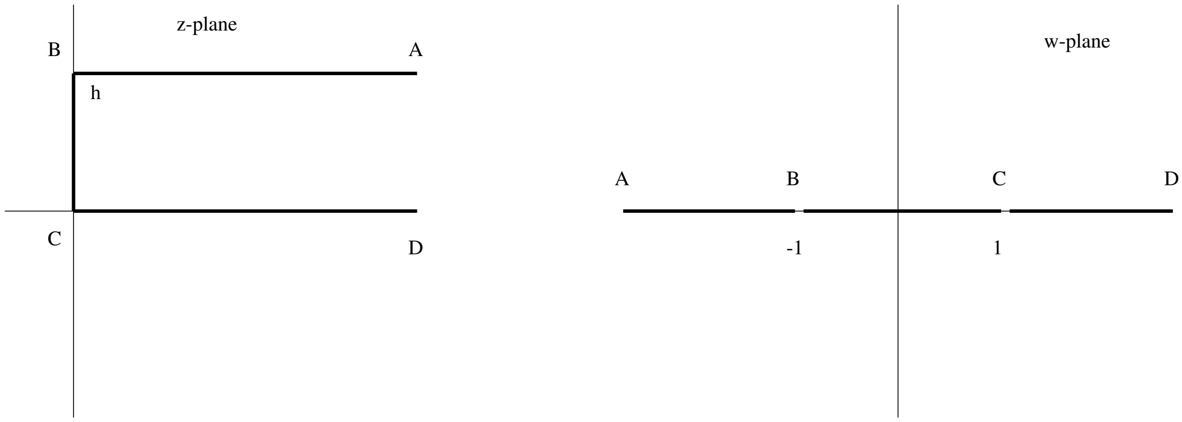

Figure 19: The U-sphaped channel is mapped into the three line segments in the w-plane.

For another example, consider two conductors, one of which consists of the two semi-infinite lines $(x \ge 0, y = 0)$ and $(x = 0, y \ge 0)$ (i.e. the x and y axes in the positive quadrant), and the other consists of the infinite line $y=-d$. Suppose the first conductor is at potential zero, and the second is at potential $V=V_0$. This is an interesting geometry in which to study the electrostatic potential, because one can find an analytical solution everywhere, and it will describe the "fringing field" in the vicinity of the sharp 90-degree angle at the origin. We shall map this geometry onto the real axis in the w-plane. Let us choose constants so that as w runs from $-\infty$ to 0, the z coordinate runs from $z=-\infty-\mathrm{i}\,d$ to $z=+\infty-\mathrm{i}\,d$. Then, as w runs from 0 to 1, the z coordinate runs from $z=+\infty$ to $z=0$. Finally, as w runs from 1 to $+\infty$, the z coordinate runs from 0 to $+\mathrm{i}\,\infty$. We therefore have a 180-degree angle at the point corresponding to $w=w_1=0$, implying $\sigma_1=1$, and a (-90)-degree angle at the point corresponding to $w=w_2=1$, implying $\sigma_2=-1$. Thus the Schwarz-Christoffel transformation is determined by the equation

$$\frac{dz}{dw}=A\,\frac{\sqrt{w-1}}{w}\,. \tag{5.51}$$

which integrates up to give

$$z=z_0+2A\,\sqrt{w-1}\,+\mathrm{i}\,A\,\log\Big(\frac{1+\mathrm{i}\,\sqrt{w-1}}{1-\mathrm{i}\,\sqrt{w-1}}\Big)\,. \tag{5.52}$$

We have to be a little careful here, because of the need to handle the branch cuts properly. First, we may note that $w=1$ is supposed to correspond to $z=0$. This immediately tells us that $z_0=0$. Next, we can determine A from the requirement that z should run along the line from $z=-\infty-\mathrm{i}\,d$ to $z=+\infty-\mathrm{i}\,d$ as w runs from $-\infty$ to 0. In this region we have

$$\sqrt{w-1}=\mathrm{i}\,\lambda\,, \tag{5.53}$$

where λ is real and satisfies $\lambda>1$. Thus the logarithm gives

$$\log\Big(\frac{1+\mathrm{i}\,\sqrt{w-1}}{1-\mathrm{i}\,\sqrt{w-1}}\Big)=\log\Big(\frac{1-\lambda}{1+\lambda}\Big)=\mathrm{i}\,\pi+\log\Big(\frac{\lambda-1}{\lambda+1}\Big)=\mathrm{i}\,\pi+\mu\,, \tag{5.54}$$

where μ is real and runs from 0 to $-\infty$ as w runs from $-\infty$ to 0. So we have

$$z=2A\,\mathrm{i}\,\lambda-A\,\pi+\mathrm{i}\,A\,\mu \tag{5.55}$$

in this region. We are wanting z to have a constant imaginary part $-\mathrm{i}\,d$ along this line, and so we must choose

$$A=\frac{\mathrm{i}\,d}{\pi}\,, \tag{5.56}$$

giving

$$z=-\mathrm{i}\,d-\frac{2d}{\pi}\,\lambda-\frac{d}{\pi}\,\mu\,. \tag{5.57}$$

It is clear, looking at how λ and μ are varying with w, that at large negative w the λ term dominates, sending the real part of z to large negative values. On the other hand as w approaches 0 from the left, the μ term dominates, sending the real part of z to large positive values. So far, so good!

Now, consider what happens for $0 < w < 1$. Here we still have $\sqrt{w-1} = \mathrm{i}\,\lambda$ with λ real and positive, but now $0 < \lambda < 1$. Accordingly, the logarithm is now of the form

$$\log\Big(\frac{1-\lambda}{1+\lambda}\Big) = \mu\,, \tag{5.58}$$

with μ real, running from $\mu = -\infty$ at $w = 0$ to $\mu = 0$ at $w = 1$. It follows from (5.52) that this w segment does indeed map into the required segment in the z-plane, with z running from $+\infty$ to 0.

Finally, consider what happens when $w > 1$. We now have $\sqrt{w-1} = \lambda$ with λ real and positive here, so the region $1 < w \le \infty$ corresponds to $0 < \lambda \le \infty$. Thus we have

$$z = \frac{2\mathrm{i}\,d}{\pi}\,\lambda - \frac{d}{\pi}\,\log\Big(\frac{1+\mathrm{i}\,\lambda}{1-\mathrm{i}\,\lambda}\Big) \tag{5.59}$$

in this region. Now if we let $p = \log((1+\mathrm{i}\,\lambda)/(1-\mathrm{i}\,\lambda))$ then we have $\mathrm{i}\,\lambda = (e^p - 1)/(e^p + 1) = \tanh(p/2)$, and so

$$p = 2\mathrm{i}\,\arctan\lambda\,, \tag{5.60}$$

which is purely imaginary. We can now easily see that as w increases from 1 to ∞, we do indeed have z runnning from $z = 0$ up the imaginary axis to $z = \mathrm{i}\,\infty$.

In summary, we have determined that the required conformal mapping is

$$z = \frac{2\,\mathrm{i}\,d}{\pi}\,\sqrt{w-1} - \frac{d}{\pi}\,\log\Big(\frac{1+\mathrm{i}\,\sqrt{w-1}}{1-\mathrm{i}\,\sqrt{w-1}}\Big)\,, \tag{5.61}$$

with the branch point at $w = 1$ handled as discussed above. The mapping is illustrated in the figure below.

Now, finally, how do we use this transformation? We have mapped the problem of solving Laplaces' equation into one where we have the boundary conditions that the potential $V = 0$ on the positive real w-axis, and $V = V_0$, which is a given constant, on the negative real w-axis. This is easily solved, giving

$$V = \frac{V_0}{\pi}\,\theta = \mathrm{Im}\Big(\frac{V_0}{\pi}\,\log w\Big)\,, \tag{5.62}$$

where θ is the polar angle in the w-plane. In other words, the equipotential surfaces are radial lines coming out from the origin. It is convenient to view the potential V as the

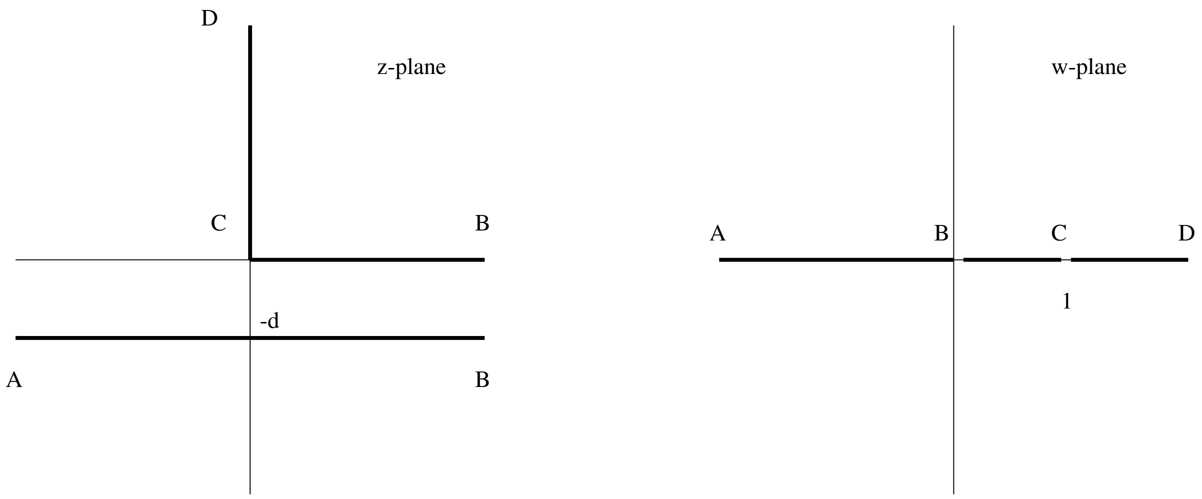

Figure 20: The two conductors in the z-plane are mapped into line segments in the w-plane.

imaginary part of an analytic function W:

$$W = U + \mathrm{i}\, V = \frac{V_0}{\pi}\, \log w\,. \tag{5.63}$$

A question of interest here is to calculate the electric field in the z-plane of the original problem, so that we can see the fringing-fields near the sharp corner at $z = 0$. Things are a little bit tricky here, since we are obviously not going to be able to invert the relation $z = z(w)$ in (5.61) explicitly, to obtain $w = w(z)$. Nonetheless, we can learn a lot from what can be done. To do this, we note from (5.4) that

$$\begin{aligned} \frac{\partial W}{\partial z} &= \tfrac12 \frac{\partial U}{\partial x} + \tfrac{\mathrm{i}}{2} \frac{\partial V}{\partial x} + \tfrac{1}{2\mathrm{i}} \frac{\partial U}{\partial y} + \tfrac12 \frac{\partial V}{\partial y}\,, \\ \frac{\partial W}{\partial \bar z} &= \tfrac12 \frac{\partial U}{\partial x} + \tfrac{\mathrm{i}}{2} \frac{\partial V}{\partial x} - \tfrac{1}{2\mathrm{i}} \frac{\partial U}{\partial y} - \tfrac12 \frac{\partial V}{\partial y} = 0\,, \end{aligned} \tag{5.64}$$

(the second line vanishes because W is analytic). Adding these equations gives

$$\frac{\partial W}{\partial z} = \frac{\partial U}{\partial x} + \mathrm{i}\, \frac{\partial V}{\partial x}\,, \tag{5.65}$$

which can be rewritten using the Cauchy-Riemann equations as

$$\frac{\partial W}{\partial z} = \frac{\partial V}{\partial y} + \mathrm{i}\, \frac{\partial V}{\partial x}\,. \tag{5.66}$$

This is nothing but the statement that

$$E_x - \mathrm{i}\, E_y = \mathrm{i}\, \frac{\partial W}{\partial z}\,, \tag{5.67}$$

where E_x and E_y are the x and y components of the electric field in the z-plane. Using the chain rule, $\partial W/\partial z = (\partial W/\partial w)\, (\partial w/\partial z)$, and (5.51), we therefore find

$$E_x - \mathrm{i}\, E_y = \frac{V_0}{d\, \sqrt{w-1}}\,. \tag{5.68}$$

Now, consider first the region near to $w = 0$, for which we shall have $\sqrt{w-1} \sim \mathrm{i}$, and hence we get

$$E_x \sim 0\,, \qquad E_y \sim \frac{V_0}{d}\,. \tag{5.69}$$

This is what we should expect; far over to the right-hand side, the electric field should look just like the field in a parallel-plate capacitor, with potential difference V_0 and plate-separation d.

In the region where $\mathrm{Re}(w) >> 1$, we see that the field falls away, as it should high up in the region where $\mathrm{Im}(z)$ is very large. In particular, when w is real and large, we see that $E_y = 0$. This is exactly as it should be; the tangential component of electric field at a conductor should vanish.

Now consider the region with $|w| >> 1$, with no particlar restriction on the phase angle. We see from (5.61) that we shall have

$$z \approx \frac{2\,\mathrm{i}\,d}{\pi}\,\sqrt{w}\,, \tag{5.70}$$

so from (5.68) we shall have

$$E_x - \mathrm{i}\,E_y \approx \frac{2\mathrm{i}\,V_0}{\pi\,z}\,. \tag{5.71}$$

Taking $z = R\,e^{\mathrm{i}\,\theta}$, with $R >> 1$, we need to consider the region $\frac{1}{2}\pi \le \theta \le \pi$. Thus we have

$$E_x - \mathrm{i}\,E_y \approx \frac{2\mathrm{i}\,V_0}{\pi\,R}\,e^{-\mathrm{i}\,\theta}\,, \tag{5.72}$$

which implies that

$$E_x = \frac{2V_0}{\pi\,R}\,\sin\theta\,, \qquad E_y = -\frac{2V_0}{\pi\,R}\,\cos\theta\,. \tag{5.73}$$

The electric field lines form large quarter-circles, starting perpendicular to the real z-axis at large negative z, and swinging round to hit the imaginary z axis at large positive-imaginary z.

Finally, the most interesting behaviour is close to the sharp corner at $z = 0$. Since this is close to $w = 1$ we can perform a Taylor expansion of (5.61) around $w = 1$, finding

$$z = \frac{2\mathrm{i}\,d}{3\pi}\,(w-1)^{3/2} + O((w-1)^{5/2})\,. \tag{5.74}$$

This can then used to solve approximately for $(w-1)^{1/2}$ in terms of z, and then substituted into (5.68). The answer is thus of the form

$$E_x - \mathrm{i}\,E_y \sim c\,z^{-1/3}\,. \tag{5.75}$$

The electric fields become singular as z approaches 0, as one would expect, and the precise nature of the fields near to $z = 0$ is determinable.

5.4 More on the Complex Plane

We shall close this chapter with some further geometrical investigation of the complex plane. This will also serve as an introduction to the topic of the next chapter, which will be some elementary group theory. To begin, let us recall that the complex plane is closely related to the so-called *Riemann Sphere*. The idea here is that by adding a single point, namely *the point at infinity*, to the ordinary complex plane, we find that it now becomes a space that can be mapped into a compact and closed surface, i.e. the Riemann Sphere. It may seem a little strange that infinity is viewed as a single point, but it can easily be understood my making a stereographic projection. The idea was introduced in Part I of the course; here, again, is the figure showing the stereographic projection:

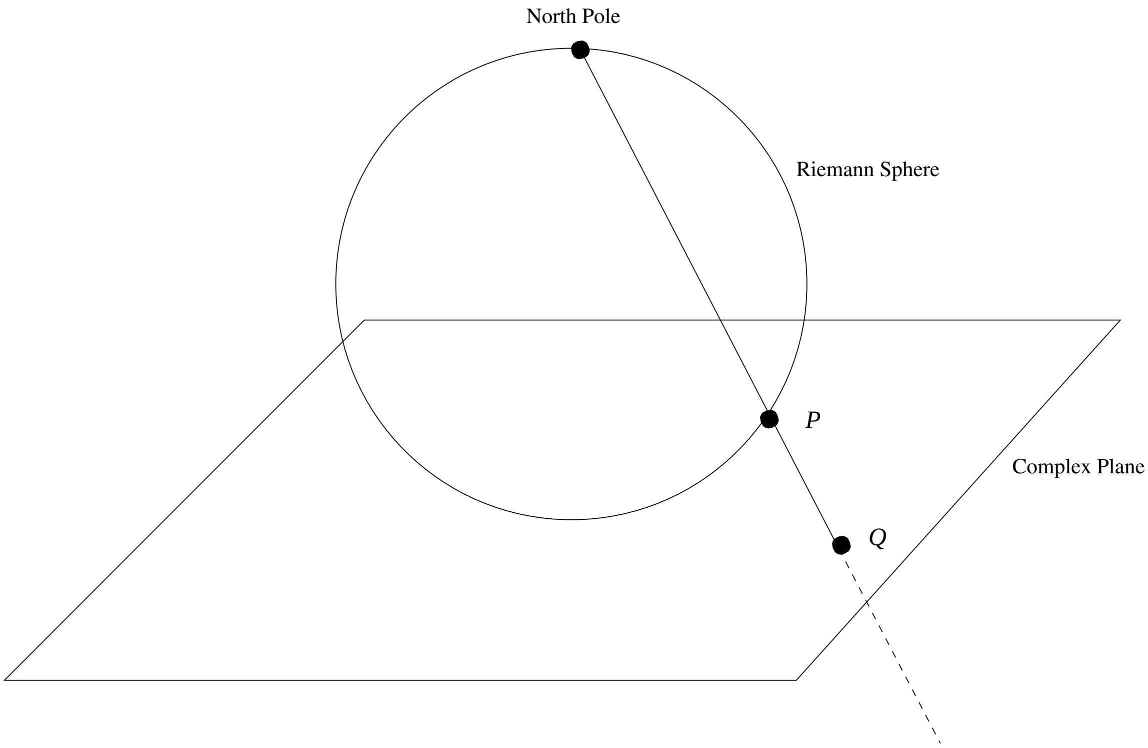

Figure 21: The point Q on the complex plane projects onto the point P on the Riemann sphere.

It is clear that any point Q in the finite complex plane projects onto a well-defined point P on the sphere. As Q moves further and further away from the origin (think of the south pole of the sphere as touching the complex plane at $z = 0$), the corresponding point P gets closer and closer to the north pole. Eventually, as $|z|$ tends to infinity, the corresponding point P reaches the north pole. It doesn't matter in which direction Q heads off to infinity; by the time it gets there, P is at the north pole. Thus by adding *the point at infinity*, the complex plane has been mapped into the compact surface of the sphere. For future reference, let us remark that this is called the 2-sphere, since its surface is 2-dimensional.

Let's now look at the stereographic projection in a little more detail. To do this, it is convenient to take the sphere that sits on the complex plane to have a diameter of 1, which means, of course, that its radius is $\frac{1}{2}$. So if we take the plane to have coordinates (x, y), and take the third direction, perpendicular to the plane, to be the t direction (we can't call it z because that has already been earmarked for another purpose!), then the origin of the sphere sits at $(x, y, t) = (0, 0, \frac{1}{2})$. The north pole sits at $(0, 0, 1)$, and, of course, the south pole is at $(0, 0, 0)$.

What we are going to do now is to work out how the usual spherical polar coordinates (θ, ϕ) for the point P on the sphere are related to the Cartesian coordinates (x, y) for the corresponding point Q in the plane. For this purpose, it is useful to give the names $(\tilde{x}, \tilde{y}, \tilde{t})$ to the Cartesian coordinates of points in the 3-space. The sphere is clearly defined by the equation

$$\tilde{x}^2 + \tilde{y}^2 + (\tilde{t} - \tfrac{1}{2})^2 = \tfrac{1}{4}\,. \tag{5.76}$$

On the other hand the line running from the north pole at $(0, 0, 1)$ to the point Q at $(x, y, 0)$ can be parameterised as

$$(\tilde{x}, \tilde{y}, \tilde{z}) = (\lambda\, x, \lambda\, y, 1 - \lambda)\,, \tag{5.77}$$

so that as λ increases from 0 to 1 we move along the straight line from the north pole to Q. The point P is located at the intersection of the surface (5.76) and the line (5.77), which implies

$$\lambda^2\,(x^2 + y^2) + (\tfrac{1}{2} - \lambda)^2 = \tfrac{1}{4}\,. \tag{5.78}$$

Multiplying out the left-hand side, we see that the $\frac{1}{4}$ on the right is cancelled, and so we get

$$(1 + \rho^2)\,\lambda^2 - \lambda = 0\,, \tag{5.79}$$

where we have defined

$$\rho^2 \equiv x^2 + y^2\,. \tag{5.80}$$

One solution is $\lambda = 0$, which just tells us the obvious fact that the sphere and the line intersect at the north pole. We want the other intersection, which therefore occurs at the value of λ given by

$$\lambda = \frac{1}{1 + \rho^2}\,. \tag{5.81}$$

From (5.77), it therefore follows that the point P is located at

$$(\tilde{x},\tilde{y},\tilde{t})=\Big(\frac{x}{1+\rho^2},\frac{y}{1+\rho^2},\frac{\rho^2}{1+\rho^2}\Big)\,. \tag{5.82}$$

To convert to the spherical polar coordinates (θ,ϕ), we recall that this are related to $(\tilde{x},\tilde{y},\tilde{t})$ by

$$\tilde{x}=\tfrac12\sin\theta\,\cos\phi\,,\qquad \tilde{y}=\tfrac12\sin\theta\,\sin\phi\,,\qquad \tilde{t}-\tfrac12=\tfrac12\cos\theta\,, \tag{5.83}$$

remembering that the sphere has radius $\frac12$ and that its origin is located at $(0,0,\frac12)$. These equations can be better written as

$$\tilde{x}+\mathrm{i}\,\tilde{y}=\tfrac12 e^{\mathrm{i}\,\phi}\,\sin\theta\,,\qquad \tilde{t}=\cos^2\tfrac12\theta\,. \tag{5.84}$$

Comparing with (5.82), and defining $z=x+\mathrm{i}\,y$ in the complex plane (this is why we couldn't use z for the 3'rd axis!), we see that

$$\cos\tfrac12\theta=\frac{|z|}{\sqrt{1+|z|^2}}\,,\qquad e^{\mathrm{i}\,\phi}=\frac{z}{|z|}=\sqrt{\frac{z}{\bar{z}}}\,, \tag{5.85}$$

since $\rho^2=x^2+y^2=|z|^2$. We can neaten up this relation, by noting that the first equation implies $|z|=\cot\frac12\theta$, and so we get

$$z=\cot\tfrac12\theta\,e^{\mathrm{i}\,\phi}\,. \tag{5.86}$$

So (5.86) gives us the required mapping from a point P on the sphere with spherical polar coordinates (θ,ϕ) to the corresponding point z in the complex plane.

Recall that we observed earlier that the way to measure the distance ds between the infinitesimally-separated points (x,y) and $x+dx,y+dy)$ in the complex plane is by Pythagoras' Theorem, giving

$$ds^2=dx^2+dy^2=dz\,d\bar{z}\;=|dz|^2\,. \tag{5.87}$$

This is called the *metric* on the plane, since it is the thing we use in order to measure distances. Suppose now that an ant lives on the sphere, and that its job is to work out the infinitesimal distance between the points Q at (x,y) and Q' at $(x+dx,y+dy)$ on the plane. However, being short-sighted, it can only see the corresponding points P and P' in the surface of the sphere, to which it assigns spherical polar coordinates (θ,ϕ) and $(\theta+d\theta,\phi+d\phi)$. From (5.86), we see that the differentials are related by

$$dz=-\tfrac12\mathrm{cosec}^{\,2}\tfrac12\theta\,e^{\mathrm{i}\,\phi}\,d\theta+\mathrm{i}\,\cot\tfrac12\theta\,e^{\mathrm{i}\,\phi}\,d\phi\,, \tag{5.88}$$

and hence the metric (5.87) in the complex z-plane becomes

$$ds^2=|dz|^2=\frac{1}{4(\sin\frac12\theta)^4}\,(d\theta^2+\sin^2\theta\,d\phi^2)\,. \tag{5.89}$$

This is therefore the rule that the ant must use, for working out the distance between the two points in the complex plane. Notice, however, that it is a different rule from the one that the ant will use if it wants to work out how far it actually has to walk on the surface of the sphere, to get from P to P'. It is a simple geometrical exercise to work out that the distance between the points (θ, ϕ) and $(\theta + d\theta, \phi + d\phi)$ on the sphere of radius $\frac{1}{2}$ is given by $d\tilde{s}$, where

$$d\tilde{s}^2 = \tfrac{1}{4}(d\theta^2 + \sin^2\theta\, d\phi^2)\,. \qquad (5.90)$$

This is just like the metric we would use on the earth, to work out the distance between any two points. (We would do this by integrating up all the infinitesimal contributions along the path, using (5.90).)

There are very important differences between the metric (5.87) on the complex plane, and the metric (5.90) on the sphere. In particular, using the metric (5.90) we would discover that there is *curvature*. This would show up, for example, if we measured the circumference L of a circle of radius R on the surface of the sphere. This is easy to work out. We can exploit the fact (which we shall examine in more detail later on) that the sphere is a completely symmetrical object, and any point on it is just like any other point (before we start attaching cities, and mountians, and things like that). Thus when considering a circle of radius R on the sphere, we may as well take the centre of the circle to be at the north pole, since that makes the calculation easy.

To get a circle of radius R, we must therefore walk from the north pole ($\theta = 0$) to a point at coordinate θ_0 such that $R = \frac{1}{2}\theta_0$ (recalling that we are stuck with a sphere of radius $\frac{1}{2}$ here). We then measure the circumference of this circle by walking around the line of latitude, at fixed $\theta = \theta_0$, until the azimuthal angle ϕ has advanced through 2π. The distance walked around the circumference is therefore $L = \frac{1}{2}\sin\theta_0$, and so the ratio of circumference to radius is given by

$$\frac{L}{R} = 2\pi\, \frac{\sin\theta_0}{\theta_0}\,, \qquad (5.91)$$

where $\theta_0 = 2R$. We see that as expected, if θ_0 is very small, corresponding to a very small circle, it has the usual property that $L/R = 2\pi$. Locally, we don't notice that the earth is curved. As the radius of the circle gets bigger, however, the ratio L/R becomes less than 2π, revealing that the surface of the earth is curved. The most extreme situation occurs when the radius of the circle becomes so big that $\theta_0 = \pi$, i.e. when $R = \pi/2$ on our earth of radius $\frac{1}{2}$. Now, the circumference of the circle is in fact zero. All we have to do to traverse the circumference in this extreme case is to stand at the south pole and not walk at all!

Let us return to our ant, and the stereographic projection from the complex plane. Just like ourselves on the earth, the ant will be aware that it lives on a curved space, since it measures its own walking distances using the sphere metric (5.90). On the other hand, during its working hours when its job is to measure distances in the complex plane, it has been instructed to use the rule given by the metric metric (5.89) for reporting distances. Using this rule, it will find no curvature, and all circles, no matter how big, will have a ratio of circumference to radius that is equal to 2π. The point is that even though it is written in terms of (θ, ϕ) coordinates, the metric (5.89) is nothing but a restatement of the original flat metric $|dz|^2$ on the complex plane.

The point of all this preamble was to draw a distinction between two very different ideas. The first is that we can choose to use any (well-behaved) coordinate system we like in order to specify the locations of points in a space. Thus, for example, on the complex plane we can simply specify a point Q by its Cartesian x and y coordinates, conveniently grouped together into the complex coordinate $z = x + \mathrm{i}\, y$. Alternatively, we can if we wish specify the same point by its *image* in the stereographic projection, with spherical polar coordinates (θ, ϕ) that are related to z by equation (5.86). The mapping between the two coordinate systems works well everywhere except at the north pole itself. This freedom to describe a given geometrical configuration in terms of different possible choices of coordinate system is one of the cornerstones of Einstein's general theory of relativity, which is the theory of gravitation. A crucial ingredient in the theory is that our description of physics, and physical laws, should be formulated in such a way that no preferred choice of coordinate system need be made.

The second idea that our investigation of the stereographic projection has introduced is that there are also genuinely different geometries that can be objectively distinguished from one another. Again, though, the choice of coordinates is not important. In particular, we saw that the flat metric on the plane is geometrically quite different from the curved metric on the 2-sphere. We wrote the flat metric ds^2 in two equivalent ways, using either Cartesian or spherical polar coordinates:

$$ds^2 = dx^2 + dy^2 = \tfrac{1}{4}\mathrm{cosec}^{\,4}\tfrac{1}{2}\theta\,(d\theta^2 + \sin^2\theta\, d\phi^2)\,. \tag{5.92}$$

By the same token, we can write the metric on the sphere in different ways too. On the one hand we have

$$d\tilde{s}^2 = \tfrac{1}{4}(d\theta^2 + \sin^2\theta\, d\phi^2)\,, \tag{5.93}$$

on the sphere of radius $\frac{1}{2}$. From (5.89) we can also therefore write this in terms of the complex coordinate z, related to (θ, ϕ) by (5.86), as

$$d\tilde{s}^2 = \sin^4 \tfrac{1}{2}\theta\, |dz|^2 \,, \tag{5.94}$$

which, after expressing θ in terms of z, becomes

$$d\tilde{s}^2 = \frac{|dz|^2}{(1+|z|^2)^2} \,. \tag{5.95}$$

Notice that the metric $d\tilde{s}^2$ on the sphere, and the metric ds^2 on the plane, are related to one another by a multiplicative factor:

$$d\tilde{s}^2 = \Omega^2\, ds^2 \,. \tag{5.96}$$

Of course the factor is coordinate-dependent, namely

$$\Omega = \frac{1}{1+|z|^2} \,. \tag{5.97}$$

This means that the *conformal structure* is preserved; the shapes of infinitesimal surfaces, and the angles between lines in infinitesimal figures, are the same whether they are measured in the flat metric or the sphere metric.

6 Some Introductory Geometry and Group Theory

6.1 Some Properties of the 2-Sphere

We shall begin by looking in more detail at at some of the properties of the 2-sphere. It is going to become tedious at this stage if we continue to work with a sphere of radius $\frac{1}{2}$; it was the "natural" radius in the context of the stereographic projection, but not otherwise. So consider from now on a sphere of radius 1, which is commmonly called the *unit sphere.* Introduce three coordinates (X, Y, Z) in Euclidean 3-space. We sometimes denote this space by $\mathbb{R}^3$ (indicating three real directions). The unit sphere can then be considered to be the surface

$$X^2 + Y^2 + Z^2 = 1 \tag{6.1}$$

in $\mathbb{R}^3$.

At times it will be convenient to use an index notation for the coordinates, and so we shall define X^a to mean $(X^1, X^2, X^3) = (X, Y, Z)$. Note that we put the index "upstairs" on the coordinates; that is a well-established convention. It does mean, however, that one has to be careful somtimes in order to avoid confusion between, for example, X^2 meaning Y (as it does here), and the total different notion of X^2 meaning X times X. Often, to avoid the confusion, it is convenient to write *explicit* numerical indices on coordinates downstairs, so that we would use X^a for the generic coordinates, but (X_1, X_2, X_3) for the $i = 1$, 2 and 3 values. This is not a perfect resolution either, and one just has to be adaptable.

Let us see how to make precise our observation of a while ago that the 2-sphere is very symmetrical, with each point on the surface looking like each other point. It can be seen very clearly in the defining equation (6.1), in fact, if we write it as

$$X^a\, X^a = 1\,. \tag{6.2}$$

Alternatively, in a vector notation, we could define the column vector $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}\,, \tag{6.3}$$

so that (6.2) becomes

$$\mathbf{X}^T\, \mathbf{X} = 1\,, \tag{6.4}$$

where $\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$.

It is now evident that if we act on the column vector $\mathbf{X}$ with any 3×3 orthogonal matrix M, to give a new column vector $\mathbf{X}' \equiv M\, \mathbf{X}$, then the condition (6.4) will be left unaltered:

$$\mathbf{X}'^T\, \mathbf{X}' = \mathbf{X}^T\, M^T\, M\, \mathbf{X} = \mathbf{X}^T\, \mathbf{X} = 1\,, \tag{6.5}$$

since $M^T\, M = \mathbb{1}$. Expressed in index notation, the equivalent statement is that $X'^a \equiv M_{ab}\, X^b$, and the orthogonality condition on the matrix is $M_{ab}\, M_{ac} = \delta_{bc}$, so that

$$X'^a\, X'^a = M_{ab}\, X^b\, M_{ac}\, X^c = \delta_{bc}\, X^b\, X^c = X^b\, X^b = 1\,. \tag{6.6}$$

Of course M_{ab} here denotes the element at row a and column b in the matrix M. Since M is 3×3 and orthogonal, it is referred to as an $O(3)$ matrix. An orthogonal $n\times n$ matrix is correspondingly called an $O(n)$ matrix.

Thus we have the statement that if one acts the on the defining equation (6.2) with any $O(3)$ matrix, the equation is left unaltered. This means that $O(3)$ is the *symmetry group* of the 2-sphere. It may be helpful to look at what infinitesimal $O(3)$ transformations do to the sphere. Suppose M is orthogonal, and infinitesimally close to the identity matrix:

$$M = \mathbb{1} + A\,, \tag{6.7}$$

where the magnitudes of the components of A are infinitesimal. Then the orthogonality condition $M^T\, M = \mathbb{1}$ becomes

$$(\mathbb{1} + A^T)(\mathbb{1} + A) = \mathbb{1}\,, \tag{6.8}$$

and since A is infinitesimal we can neglect the $A^T A$ term in comparison to the terms linear in A, giving $A + A^T = 0$, so

$$A^T = -A\,. \tag{6.9}$$

So the condition for M defined in (6.7) to be orthogonal when A is infinitesimal is that A should be antisymmetric.

This means that we can easily calculate the infinitesimal displacements $\delta X^a \equiv X'^a - X^a$ that result from acting with $M = 1\!\!1 + A$:

$$\delta X^a = M_{ab}\, X^b - X^a = (\delta_{ab} + A_{ab})\, X^b - X^a = A_{ab}\, X^b\,. \tag{6.10}$$

The number of independent components in a 3×3 antisymmetric matrix is clearly $\frac{1}{2}\times 3\times 2$, and so we can say that the symmetry group $O(3)$ of the 2-sphere has 3 parameters.

We can see directly that the defining surface (6.2) is invariant under the infinitesimal transformations, since we shall then have

$$\delta(X^a\, X^a) = 2X^a\, \delta X^a = 2X^a\, A_{ab}\, X^b = 0\,, \tag{6.11}$$

where the last step follows from the fact that A_{ab} is antisymmetric in ab, while $X^a\, X^b$ is symmetric in ab.

Note that not only is the surface (6.2) invariant under the $O(3)$ transformations, but so also is the metric on the 2-sphere. How do we write the metric in terms of the X^a coordinates? After all, there are three of them, but the 2-sphere needs only two coordinates. The point is that when we say *the metric on the 2-sphere*, we are having in mind the metric that we would induce by taking the ordinary Euclidean metric in $\mathbb{R}^3$, and then imposing the rule that all points have to be restricted to lie on the surface defined by (6.2). Thus the 2-sphere metric can be written as

$$ds^2 = dX^a\, dX^a\,, \tag{6.12}$$

subject to the constraint (6.2). Clearly (6.12) is also invariant under the $O(3)$ rotations that we have been considering. Bearing in mind that M is a *constant* matrix, the calculations that showed the invariance of (6.1) will work in exactly the same way to show the invariance of (6.12). Since the metric (6.12) and the constraint (6.2) are both invariant under $O(3)$, it follows that the induced metric on the surface of the sphere is invariant under $O(3)$ also.

To make contact with some earlier discussion, let us confirm that (6.12) together with (6.2) does indeed give us the metric that we expect to see on the 2-sphere. We can do this most easily by solving the constraint equation (6.2) explicitly, which can be done by making the familiar definitions

$$X = \sin\theta\,\cos\phi\,,\qquad Y = \sin\theta\,\sin\phi\,,\qquad Z = \cos\theta\,. \tag{6.13}$$

These are nothing but the usual definitions relating spherical polar coordinates to Cartesian coordinates, but with the r coordinate set equal to 1 since we have $r^2 \equiv X^2 + Y^2 + Z^2 = 1$. Substituting (6.13) into (6.12), we get

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2\,. \tag{6.14}$$

This is exactly what we should get, for the metric on a unit 2-sphere.

We can also now look at what the $O(3)$ symmetry transformations do in terms of the coordinates (θ,ϕ) on the 2-sphere. This is most easily done at the infinitesimal level, so we just take (6.10), and put it together with (6.13). First, consider δZ:

$$\delta Z = A_{31}\, X + A_{32}\, Y\,. \tag{6.15}$$

But $\delta Z = \delta(\cos\theta) = -\sin\theta\,\delta\theta$, so we get

$$-\sin\theta\,\delta\theta = -A_{13}\,\sin\theta\,\cos\phi - A_{23}\,\sin\theta\,\sin\phi\,, \tag{6.16}$$

where we have also used the antisymmetry to re-express A_{31} as $-A_{13}$, and A_{32} as $-A_{23}$. Thus we have

$$\delta\theta = A_{13}\,\cos\phi + A_{23}\,\sin\phi\,. \tag{6.17}$$

Now, we can look at δX, which gives

$$-\sin\theta\,\sin\phi\,\delta\phi + \cos\theta\,\cos\phi\,\delta\theta = A_{12}\,\sin\theta\,\sin\phi + A_{13}\,\cos\theta\,. \tag{6.18}$$

But we already know how θ transforms, from (6.17), so we can plug this back in, and hence read off the transformation for ϕ. Collecting the results together, we then have:

$$\begin{aligned}\delta\theta &= A_{13}\,\cos\phi + A_{23}\,\sin\phi\,,\\ \delta\phi &= -A_{12} - A_{13}\,\cot\theta\,\sin\phi + A_{23}\,\cot\theta\,\cos\phi\,.\end{aligned} \tag{6.19}$$

This gives us the infinitesimal transformations of the θ and ϕ coordinates on the 2-sphere, corresponding to the action of the infinitesimal $O(3)$ transformation with parameters A_{12}, A_{13} and A_{23}.

Notice that the transformation corresponding to the parameter A_{12} is particularly simple; it is just

$$\delta\theta = 0\,,\qquad \delta\phi = -A_{12}\,. \tag{6.20}$$

This means that under this symmetry transformation the θ coordinate is held fixed, and the ϕ coordinate is shifted by an infinitesimal constant. We can easily visualise this symmetry transformation; we just take a little walk along a line of latitude on the sphere. Obviously this is a symmetry. This can also be seen by looking at the metric (6.14) on the sphere; sending $\phi \longrightarrow \phi+$ constant leaves the metric unaltered. The other two symmetry transformations, associated with the parameters A_{13} and A_{23} are a little harder to visualise, in terms of the θ and ϕ coordinates on the 2-sphere, but they again correspond to translations on the surface, which again leave the metric unchanged.

6.2 Vector Fields

In fact the infinitesimal transformations of the coordinates θ and ϕ that we have just seen allow us to introduce the concept of a *vector field.* We should begin this discussion by forgetting certain things about vectors that we learned in kindergarten. There, the concept of a vector was introduced through the notion of the *position vector*, which was an arrow joining a point A to some other point B in three-dimensional Euclidean space. This is fine if one is only going to talk about Euclidean space in Cartesian coordinates, but it is not a valid way describing a vector in general. If the space is curved, such as the sphere, or even if it is flat but described in non-cartesian coordinates, such as Euclidean 3-space described in spherical polar coordinates, the notion of a vector as a line joining two distant points A and B breaks down. What we *can* do is take the infinitesimal limit of this notion, and consider the line joining two points A and $A + \delta A$. In fact what this means is that we think of the *tangent plane* at a point in the space, and imagine vectors in terms of infinitesimal displacements in this plane.

To make the thinking a bit more concrete, consider a 2-sphere, such as the surface of the earth. A line drawn between Ney York and Los Angeles is not a vector; for example, it would not make sense to consider the "sum" of the line from New York to Los Angeles and the line from Los Angeles to Tokyo, and expect it to satisfy any meaningful addition rules. However, we *can* place a small flat sheet on the surface of the earth at any desired point, and draw very short arrows in the plane of the sheet; these are tangent vectors at that particular point on the earth.

The concept of a vector as an infinitesimal displacement makes it sound very like the derivative operator, and indeed this is exactly what a vector is. Suppose we draw a path on the surface of the earth, parameterised by some quantity λ that increases monotonically as we move along the path. The coordinates of a point P on the path will then be given by

$(\theta(\lambda), \phi(\lambda))$, and the tangent vector at that point is

$$V = \frac{\partial}{\partial \lambda} \,. \tag{6.21}$$

Generally, if we are in a space with coordinates x^i, and there is a path $x^i(\lambda)$ parameterised by λ, then the tangent vector at the point P is again given by (6.21). Furthermore, using the chain rule for differentiation, we shall have

$$V = \frac{\partial}{\partial \lambda} = \frac{dx^i(\lambda)}{d\lambda}\, \frac{\partial}{\partial x^i} \,. \tag{6.22}$$

The derivatives $\partial_i \equiv \partial/\partial x^i$, which in fact are what we normally call the *gradient* operator, are acting here as a set of *basis vectors* for the tangent space, and we may write the vector V as

$$V = V^i\, \partial_i \,, \tag{6.23}$$

where V^i are the *components* of the vector V with respect to the basis ∂_i;

$$V^i = \frac{dx^i(\lambda)}{d\lambda} \,. \tag{6.24}$$

(Of course here we are using the Einstein summation convention that any dummy index, which occurs twice in a term, is understood to be summed over the range of the index.)

Notice that there is another significant change in viewpoint here in comparison to the "kindergarten" notion of a vector. We make a clear distinction betwen the vector itself, which is the geometrical object V defined quite independently of any coordinate system by (6.21), and its *components* V^i, which are coordinate-dependent.[22] Indeed, if we imagine now changing to a different set of coordinates x'^i in the space, related to the original ones by $x'^i = x'^i(x^j)$, then we can use the chain rule to convert between the two bases:

$$V = V^j\, \frac{\partial}{\partial x^j} = V^j\, \frac{\partial x'^i}{\partial x^j}\, \frac{\partial}{\partial x'^i} \equiv V'^i\, \frac{\partial}{\partial x'^i} \,. \tag{6.25}$$

In the last step we are, by definition, taking V'^i to be the components of the vector V with respect to the primed coordinate basis. Thus we have the rule

$$V'^i = \frac{\partial x'^i}{\partial x^j}\, V^j \,, \tag{6.26}$$

which tells us how to transform the components of the vector V between the primed and the unprimed coordinate system. This is the fundamental defining rule for how a vector

[22]However, it sometimes becomes cumbersome to use the longer form of words "the vector whose components are V^i," and so we shall sometimes slip into the way of speaking of "the vector V^i." One should remember, however, that this is a slightly sloppy way of speaking, and the more precise distinction between the vector and its components should always be borne in mind.

must transform under arbitrary coordinate transformations. Such transformations are called *General Coordinate Transformations.*

Let us return to the point alluded to previously, about the vector as a linear differential operator. We have indeed been writing vectors as derivative operators, so let's see why that is very natural. Suppose we have a scalar field $\psi(x)$ defined in the space. (We suppress the i index on the coordinates x^i in the argument here; think of the x in $\psi(x)$ as representing the full set of coordinates, $\psi(x_1, x_2, \ldots, x_n)$.) Now, if we wish to evaluate ψ at a nearby point $x^i + \xi^i$, where ξ^i is infinitesimal, we can just make a Taylor expansion:

$$\psi(x+\xi) = \psi(x) + \xi^i\, \partial_i\, \psi(x) + \cdots\,, \tag{6.27}$$

and we can neglect the higher terms since ξ is assumed to be infinitesimal. Thus we see that the change in ψ is given by

$$\delta\psi(x) \equiv \psi(x+\xi) - \psi(x) = \xi^i\, \partial_i\, \psi(x)\,, \tag{6.28}$$

and that the operator that is implementing the translation of $\psi(x)$ is exactly what we earlier called a vector field,

$$\xi^i\, \partial_i\,, \tag{6.29}$$

where

$$\delta x^i \equiv (x^i + \xi^i) - x^i = \xi^i\,. \tag{6.30}$$

Having introduced the concept of the vector field, let's go back to our discussion of the symmetries of the 2-sphere. Recall that we had infinitesimal translations of the (θ, ϕ) coordinates, given by

$$\begin{aligned} \delta\theta &= A_{13}\, \cos\phi + A_{23}\, \sin\phi\,, \\ \delta\phi &= -A_{12} - A_{13}\, \cot\theta\, \sin\phi + A_{23}\, \cot\theta\, \cos\phi\,, \end{aligned} \tag{6.31}$$

where A_{12}, A_{13} and A_{23} are infinitesimal constants. Thinking of θ and ϕ as the two coordinates x^i in the 2-sphere, we see that we have precisely the situation we were just looking at, with infinitesimal components ξ^i of vector fields that can be read off by comparing (6.30) with (6.31). Let us give the names K_{12}, K_{13} and K_{23} to the three vector fields associated with the transformation parameters A_{12}, A_{13} and A_{23} respectively. Thus we have

$$\begin{aligned} K_{12} &= \frac{\partial}{\partial\phi}\,, \\ K_{13} &= -\cos\phi\, \frac{\partial}{\partial\theta} + \cot\theta\, \sin\phi\, \frac{\partial}{\partial\phi}\,, \\ K_{23} &= -\sin\phi\, \frac{\partial}{\partial\theta} - \cot\theta\, \cos\phi\, \frac{\partial}{\partial\phi}\,. \end{aligned} \tag{6.32}$$

(We have introduced an overall factor of (-1) in each case, just for convenience.)

It will be recalled that the three vector fields that we have obtained in (6.32) have a very special property, namely that they desribe transalations on the surface of the sphere which leave the metric invariant. They are in fact the generators of the symmetry group of the 2-sphere. Recall that the symmetry group was $O(3)$. Actually, at the infinitesimal level which we are looking at now, we can't tell the difference between $O(3)$ and $SO(3)$, where the "S" stands for special, and indicates that the orthogonal $O(3)$ matrices are furthermore restricted to have determinant equal to $+1$. The orthogonality condition $M^T\, M = \mathbb{1}$ implies that

$$(\det M^T)\,(\det M) = 1\,, \tag{6.33}$$

and hence $(\det M)^2 = 1$ and so $\det M = \pm 1$, so the additional imposition of the $\det M = +1$ condition amounts to a discrete choice that restricts the matrices M to describing pure rotations, without reflections. So in the context of infinitesimal transformations, it is more appropriate to think of the symmetry group of the sphere as being $SO(3)$.

The set of three vectors (6.32) describe the $SO(3)$ rotational symmetries of the 2-sphere. On any space, the vectors that describe the continous symmetries of the space are called *Killing vectors*[23]. The $SO(3)$ Killing vectors (6.32) may seem rather familiar; they are exactly what one meets in quantum mechanics when studying angular momentum. The angular momentum operators are precisely the generators of rotational translations in Euclidean 3-space, and so not surprisingly, they are synonymous with vector fields. By the same token the ordinary linear momentum operators $\mathbf{P}$ are the generators of linear translations in Euclidean 3-space, and so not surprisingly they are associated with the vector fields

$$\frac{\partial}{\partial x}\,, \qquad \frac{\partial}{\partial y}\,, \qquad \frac{\partial}{\partial z}\,. \tag{6.34}$$

We shall close this discussion of vector fields, and Killing vectors, by looking a little more closely at the sense in which the $SO(3)$ Killing vectors in (6.32) leave the metric

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi \tag{6.35}$$

on the 2-sphere invariant. To do this, we can look first at the more general situation of a metric on some general n-dimensional space. We can write this is

$$ds^2 = g_{ij}\, dx^i\, dx^j\,, \tag{6.36}$$

[23]Named after nothing more sinister than a mathematician called Killing!

where g_{ij} are the components of a 2-index symmetric tensor, called the metric tensor. In general it depends on the coordinates x^i. Thus in the case of the 2-sphere we have $x^1 = \theta$, $x^2 = \phi$, and

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix} . \tag{6.37}$$

Notice that the way we are writing the metric in (6.36) is somewhat reminiscent of the way we wrote the vector field V in (6.23). In that case, the geometrical quantity V was expanded in a coordinate basis, in terms of components V^i multiplying the partial derivatives $\partial/\partial x^i$. Here, we are expanding the geometrical quantity ds^2 in terms of its components g_{ij} which multiply the coordinate differentials dx^i. The key difference here is that the indices on the metric tensor components g_{ij} live *downstairs*, whereas the index on the components of the vector field live *upstairs*. These are two quite distinct types of object that one encounters in geometry. We may consider a simpler example of a 1-index object, say U_i, with

$$U = U_i\, dx^i\, . \tag{6.38}$$

One can again work out how the components U_i transform under a change of coordinate basis by using the chain rule:

$$U = U_j\, dx^j = U_j\, \frac{\partial x^j}{\partial x'^i}\, dx'^i \equiv U'_i\, dx'^i\, , \tag{6.39}$$

from which we read off

$$U'_i = \frac{\partial x^j}{\partial x'^i}\, U_j\, . \tag{6.40}$$

This is the "inverse" of the transformation rule for the vector field that we derived in equation (6.26). In a similar fashion, from the intrinsic coordinate independence of the geometrical quantity ds^2 itself, we can deduce that the components g_{ij} of the metric tensor transform as

$$g'_{ij} = \frac{\partial x^k}{\partial x'^i}\, \frac{\partial x^\ell}{\partial x'^j}\, g_{k\ell}\, , \tag{6.41}$$

under a change of coordinate system.

We have seen how the components of vector fields, such as V^i and U_i, transform under general coordinate transformations. (See (6.26) and (6.40).) More generally, we can consider tensors whose components comprise p upstairs indices, and q downstairs indices:

$$T^{i_1\cdots i_p}{}_{j_1\cdots j_q}\, . \tag{6.42}$$

These quantities will transform analogously under general coordinate transformations, with one transformation factor like in (6.26) for each upstairs index, and one factor like in (6.40) for each downstairs index:

$$T'^{i_1\cdots i_p}{}_{j_1\cdots j_q} = \frac{\partial x'^{i_1}}{\partial x^{k_1}} \cdots \frac{\partial x'^{i_p}}{\partial x^{k_q}}\, \frac{\partial x^{\ell_1}}{\partial x'^{j_1}} \cdots \frac{\partial x^{\ell_q}}{\partial x'^{j_q}}\, T^{k_1\cdots k_p}{}_{\ell_1\cdots\ell_q}\,. \tag{6.43}$$

In fact we already encountered one such example, namely the metric tensor, with components g_{ij}, in (6.41). Tensors $T^{i_1\cdots i_p}{}_{j_1\cdots j_q}$, which by definition transform according to (6.43), are said to transform *covariantly* under general coordinate transformations. Similarly, a tensor-valued equation where all the terms transform according to this rule are said to be *covariant* equations. This means that the rule for transforming them from the unrpimed coordinate system to the primed coordinate system is simply to put primes on everything. What could be easier!

Notice that if we make a contraction of indices in some tensor expression, then the resulting quantity now has the transformation rule that we should expect for an object with the reduced number of free indices. For example, if we take the vectors V^i and U_i, an make a contraction, we can construct the scalar quantity

$$\phi = V^i\, U_i\,. \tag{6.44}$$

We call this a scalar because it requires no coordinate transformation matrix at all (it couldn't, since there are no indices for the matrix to hook onto!). Thus under general coordinate transformations we find

$$\phi' \equiv V'^i\, U'_i = \frac{\partial x'^i}{\partial x^k}\, V^k\, \frac{\partial x^\ell}{\partial x'^i}\, U_\ell = \frac{\partial x^\ell}{\partial x^k}\, V^k\, U_\ell = \delta^\ell_k\, V^k\, U_\ell = V^k\, U_k = \phi\,. \tag{6.45}$$

More generally, if we contract n of the upper indices in $T^{i_1\cdots i_p}{}_{j_1\cdots j_q}$ with n of the lower indices, we shall end up with an object with $(p-n)$ free upper indices, and $(q-n)$ free lower indices, which transforms exactly as a tensor with those numbers of upper and lower indices should.

To close this section, let us go back to the symmetries of the 2-sphere, or more generally, the symmetries of any metric.[24] If an infinitesimal translation $\delta x^i = \xi^i$ of the coordinates leaves the metric invariant then we shall have $ds^2(x+\delta x) = ds^2(x)$, and so

$$g_{ij}(x+\delta x)\, d(x^i + \xi^i)\, d(x^j+\xi^j) = g_{ij}\, dx^i\, dx^j\,, \tag{6.46}$$

where we need only keep quantities up to first order in the infinitesimal ξ^i. Since from the chain rule we have $d\xi^i = (\partial_k\, \xi^i)\, dx^k$, we get, after appropriate changes of the names of dummy summation indices,

$$g_{ij}\, dx^i\, dx^j + \Big(\xi^k\, \partial_k\, g_{ij} + g_{kj}\, \partial_i\, \xi^k + g_{ik}\, \partial_j\, \xi^k\Big)\, dx^i\, dx^j = g_{ij}\, dx^i\, dx^j\,, \tag{6.47}$$

[24]Not all metric have symmetries, so this discussion applies to such symmetries as they may have.

and so the condition for ξ^i to be the components of a Killing vector is

$$\xi^k\,\partial_k\, g_{ij} + g_{kj}\,\partial_i\,\xi^k + g_{ik}\,\partial_j\,\xi^k = 0\,. \tag{6.48}$$

A vector with components ξ^i that satisfies this equation is what is called a Killing vector, and the equation is Killing's equation.

It is quite easy to verify that the three Killing vectors (6.32) that we obtained earlier on the 2-sphere do indeed satisfy Killing's equation. The easiest one to check is K_{12}, since it corresponds simply to $\xi^1 = 0$, $\xi^2 = 1$. Since these components are constants the last two terms in (6.48) can immediately be seen to be zero, while in the first term the directional derivative $\xi^k\,\partial_k$ is clearly just $\partial/\partial\phi$, and so this gives zero since none of the components of the 2-sphere metric (6.37) depends on ϕ. Checking that the other two Killing vectors in equation (6.32) satisfy (6.48) takes a little more work, and in fact one now gets a non-trivial cancellation between contributions from the various terms. Of course there is, logically-speaking, really no need to verify that the vectors in (6.32) do indeed satisfy (6.48), since they were constructed precisely to have the property of generating symmetries of the metric. But it is sometimes reassuring to check things by different methods, to reaffirm that there is indeed a consistent unity in the universe!

6.3 The Metric Tensor and its Inverse

The metric tensor plays many important rôles in geometry. One of these is that it can be used to lower the index on the components of a vector V^i, to give a quantity whose components $g_{ij}\,V^j$ transform just like the U_i we discussed above. To check this, we just evaluate the quantity $g_{ij}\,V^j$ in the primed coordinate system, which we can easily do since we know exactly how g_{ij} and V^j transform (see (6.41)):

$$g'_{ij}\,V'^j = \frac{\partial x^k}{\partial x'^i}\,\frac{\partial x^\ell}{\partial x'^j}\,g_{k\ell}\,\frac{\partial x'^j}{\partial x^m}\,V^m\,. \tag{6.49}$$

But by the chain rule, we have

$$\frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^j}{\partial x^m} = \frac{\partial x^\ell}{\partial x^m}\,, \tag{6.50}$$

and then by definition this gives us δ^ℓ_m, so we find:

$$g'_{ij}\,V'^j = \frac{\partial x^k}{\partial x'^i}\,g_{km}\,V^m\,. \tag{6.51}$$

This is exactly the way that a vector with downstairs components, like U_i in (6.40) should transform. In fact we can be economical with our use of symbols, and *define*

$$V_i \equiv g_{ij}\,V^j\,. \tag{6.52}$$

At the moment, the use of the metric to lower indices looks a bit like a "one-way street," since having got the index downstairs, we don't yet know how to get it back upstairs again. But this is easily remedied; we just need the inverse metric. This is literally what is sounds like; we view g_{ij} as a matrix, and we define the inverse of the metric to be the matrix inverse. We may write its components as g^{ij}. Since we should have $\mathbf{g}^{-1}\,\mathbf{g} = \mathbb{1}$, this means we should have

$$g^{ij}\, g_{jk} = \delta^i_k \,. \tag{6.53}$$

This can be taken as the definition of the inverse metric. It is easy to see, by manipulations precisely analogous to those we performed above, that in order for (6.53) to be true in all coordinate frames, g^{ij} should indeed transform like the components of a tensor with two upstairs indices (see (6.43)). It is then easily verified that if we take V_i defined in (6.52), and now raise the index using g^{ij}, we get back to where we started:

$$V^i = g^{ij}\, V_j \,. \tag{6.54}$$

More generally, we can use g^{ij} to raise indices on any tensor.

Notice that we can construct a scalar quantity from a vector V^i, by using the metric tensor:

$$V^i\, V^j\, g_{ij} \,. \tag{6.55}$$

This is what we can call the (magnitude)^2 of the vector. It is equivalent to the "dot product" of a vector with itself in traditional vector analysis. In the general context we are discussing here one sees that the metric tensor g_{ij} is essential for being able to construct the scalar from V^i. Of course this was effectively true in the context of Cartesian vector analysis also, but there the metric tensor was just δ_{ij}, and one hardly noticed that one was using it. More generally, we can use the metric to allow us to construct a scalar from any two vectors:

$$V^i\, W^j\, g_{ij} \,. \tag{6.56}$$

6.4 Covariant Differentiation

A familiar concept in Cartesian tensor analysis is that the partial derivatives $\partial_i \equiv \partial/\partial x^i$ can act on a tensor field to give another tensor field.[25] However, a crucial point in Cartesian tensor analysis is that we do not consider general coordinate transformations; rather, we restrict ourselves only to *constant* transformation matrices M_{ij} which, furthermore, are

[25] We now use "tensor" as a generic term, which can include the particular cases of a scalar, and a vector.

orthogonal:

$$x'^i = M_{ij}\, x^j\,, \qquad M_{ij}\, M_{ik} = \delta_{jk}\,. \tag{6.57}$$

In fact we encountered precisely such types of transformation earlier on, when considering the $O(3)$ rotational symmetry of the 2-sphere. This was because we were embedding it in 3-dimensional Euclidean space with Cartesian coordinates. For *Cartesian Tensors*, there is no need to distinguish between upstairs and downstairs indices, since the associated metric tensor is just the Kronecker delta, $g_{ij} = \delta_{ij}$, which is its own inverse. Note that from (6.57) we have

$$\frac{\partial x'^i}{\partial x^j} = M_{ij} = \text{constant}\,. \tag{6.58}$$

In Cartesian tensor analysis a tensor is any quantity whose components transform with the appropriate factors of M_{ij}, as, for example,

$$V'^i = M_{ij}\, V_j\,, \qquad \frac{\partial}{\partial x'^i} = M_{ij}\, \frac{\partial}{\partial x^j}\,. \tag{6.59}$$

(The second equation here shows that the gradient operator $\partial/\partial x^i$ is a vector.)

Now, from the above it is easy to see that if V^i is a Cartesian vector field, then the quantity

$$T^i{}_j \equiv \frac{\partial V^i}{\partial x^j} \tag{6.60}$$

is a Cartesian tensor. We prove this by the standard technique of showing that it transforms properly for a Cartesian tensor:

$$T'^i{}_j \equiv \frac{\partial V'^i}{\partial x'^j} = M_{j\ell}\, \frac{\partial (M_{ik}\, V^k)}{\partial x^\ell} = M_{j\ell}\, M_{ik}\, \frac{\partial V^k}{\partial x^\ell} = M_{j\ell}\, M_{ik}\, T^k{}_\ell\,. \tag{6.61}$$

The crucial step in this proof was the one where the transformation matrix M_{ik} was brought outside the differentiation, *because it is a constant matrix.* This is the step where things are going to be different when we consider the case of tensors under general coordinate transformations.

The above was a review of what happens for Cartesian tensors. Now, let's get back to the much more general case we are really interested in, of quantities that transform as tensors under the completely arbitrary general coordinate transformations, with $x'^i = x'^i(x^j)$. First, let's see what goes wrong with a naive attempt, and then we'll see how to fix it.

Suppose V^i is a vector under general coordinate transformations (so it transforms as in (6.26)). Let us consider the quantity

$$W^i{}_j \equiv \frac{\partial V^i}{\partial x^j}\,. \tag{6.62}$$

Is this a tensor? To test it, we calculate $W'^i{}_j$, to see if it is the proper tensorial transform of $W^i{}_j$. We get:

$$\begin{aligned} W'^i{}_j \equiv \frac{\partial V'^i}{\partial x'^j} &= \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial}{\partial x^\ell}\Big(\frac{\partial x'^i}{\partial x^k}\,V^k\Big) \\ &= \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^i}{\partial x^k}\,\frac{\partial V^k}{\partial x^\ell} + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^k}\,V^k\,, \\ &= \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^i}{\partial x^k}\,W^k{}_\ell + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^k}\,V^k\,. \end{aligned} \tag{6.63}$$

So the answer is no; the first term by itself would have been fine, but the second term here has spoiled the general coordinate transformation behaviour. Of course there is no mystery behind what we are seeing here; the second term has arisen because the derivative operator has not only landed on the vector V^k, giving us what we want, but it has also landed on the transformation matrix $\partial x'^i/\partial x^k$. This problem was avoided in the case of the Cartesian tensors, because we only required that they transform nicely under *constant* transformations (6.58).

The concept of differentiation is too important for us to give it up in this context. Accordingly, what we have to do now is to generalise the notion of a derivative, so that it *does* have the property of yielding tensors when we act with it on tensors. What we need to define now is the *Covariant Derivative.*

To abbreviate the writing, let us start to make use of the notation we briefly introduced earlier, where the usual partial derivatives are written as ∂_i:

$$\partial_i \equiv \frac{\partial}{\partial x^i}\,. \tag{6.64}$$

Now, we shall define the covariant derivative ∇_j of a vector V^i as follows:

$$\nabla_j\, V^i \equiv \partial_j\, V^i + \Gamma^i{}_{jk}\, V^k\,, \tag{6.65}$$

where the quantities $\Gamma^i{}_{jk}$ satisfy the symmetry condition

$$\Gamma^i{}_{jk} = \Gamma^i{}_{kj}\,. \tag{6.66}$$

They are defined to have precisely the correct transformation properties under general coordinate transformations that ensure that the quantity

$$T^i{}_j \equiv \nabla_j\, V^i \tag{6.67}$$

does transform like a tensor under general coordinate transformations. The crucial point here is that $\Gamma^i{}_{jk}$ itself is *not* a tensor. It is called a *Connection*, in fact.

First, let us see how we would *like* $\Gamma^i{}_{jk}$ to transform, and then, we shall show how to construct such an object. By definition, we want it to be such that

$$\frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\nabla_\ell\,V^k = \nabla'_j\,V'^i \equiv \partial'_j\,V'^i + \Gamma'^i{}_{jk}\,V'^k\,. \tag{6.68}$$

Wrtiting out the two sides here, we get the requirement that

$$\begin{aligned}\frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\Big(\partial_\ell\,V^k + \Gamma^k{}_{\ell m}\,V^m\Big) &= \frac{\partial x^\ell}{\partial x'^j}\,\partial_\ell\Big(\frac{\partial x'^i}{\partial x^m}\,V^m\Big) + \Gamma'^i{}_{jk}\,\frac{\partial x'^k}{\partial x^m}\,V^m \\ = \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x'^i}{\partial x^m}\,\partial_\ell\,V^m &+ \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,V^m + \Gamma'^i{}_{jk}\,\frac{\partial x'^k}{\partial x^m}\,V^m\,.\end{aligned} \tag{6.69}$$

The required equality of the left-hand side of the top line and the right-hand side of the bottom line *for all vectors* V^m allows us to deduce that we must have

$$\frac{\partial x'^i}{\partial x^m}\,\frac{\partial x^\ell}{\partial x'^j}\,\Gamma^k{}_{\ell m} = \frac{\partial x'^k}{\partial x^m}\,\Gamma'^i{}_{jk} + \frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,. \tag{6.70}$$

Multiplying this by $\partial x^m/\partial x'^n$ then gives us the result that

$$\Gamma'^i{}_{jn} = \frac{\partial x'^i}{\partial x^k}\,\frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial x^m}{\partial x'^n}\,\Gamma^k{}_{\ell m} - \frac{\partial x^m}{\partial x'^n}\,\frac{\partial x^\ell}{\partial x'^j}\,\frac{\partial^2 x'^i}{\partial x^\ell\,\partial x^m}\,. \tag{6.71}$$

This dog's breakfast is the required transformation rule for $\Gamma^i{}_{jk}$. Notice that the first term on the right-hand side is the "ordinary" type of tensor transformation rule. The presence of the second term shows that $\Gamma^i{}_{jk}$ is not in fact a tensor, because it doesn't transform like one.

The above calculation is quite messy, but hopefully the essential point comes across clearly; the purpose of the ugly second term in the transformation rule for $\Gamma^i{}_{jk}$ is precisely to remove the ugly extra term that we encountered which prevented $\partial_j\,V^i$ from being a tensor.

Luckily, it is quite easy to provide an explicit construction for a suitable quantity $\Gamma^i{}_{jk}$ that has the right transformation properties. First, we need to note that we should like to define a covariant derivative for any tensor, and that it should satisfy Leibnitz's rule for the differentiation of products. Now the need for the covariant derivative arise because the transformation of the components of a vector or a tensor from one coordinate frame to another involves non-constant transformation matrices of the form $\partial x'^i/\partial x^j$. Therefore on a scalar, which doesn't have any indices, the covariant derivative must be just the same thing as the usual partial derivative. Combining this fact with the Leibnitz rule, we can work out what the covariant derivative of a vector with a downstairs index must be:

$$\partial_j\,(V^i\,U_i) \quad = \quad (\partial_j\,V^i)\,U_i + V^i\,\partial_j\,U_i\,, \qquad\qquad \text{usual Leibnitz rule}\,,$$

$$= \quad \nabla_j\,(V^i\,U_i) = (\nabla_j\,V^i)\,U_i + V^i\,\nabla_j\,U_i\,, \qquad \text{covariant Leibnitz rule}\,, \tag{6.72}$$
$$= \quad (\partial_j\,V^i + \Gamma^i{}_{jk}\,V^k)\,U_i + V^i\,\nabla_j\,U_i\,, \qquad \text{from definition of } \nabla_j\,V^i\,.$$

Comparing the top line with the bottom line, the two $\partial_j\,V^i$ terms cancel, and we are left with

$$V^i\,\partial_j\,U_i = V^i\,\nabla_j\,U_i + \Gamma^i{}_{jk}\,V^k\,U_i\,. \tag{6.73}$$

Changing the labelling of dummy indices to

$$V^i\,\partial_j\,U_i = V^i\,\nabla_j\,U_i + \Gamma^k{}_{ji}\,V^i\,U_k\,, \tag{6.74}$$

we see that if this is to be true for all possible vectors V^i then we must have

$$\nabla_j\,U_i = \partial_j\,U_i - \Gamma^k{}_{ji}\,U_k\,. \tag{6.75}$$

This gives us what we wanted to know, namely how the covariant derivative acts on vectors with downstairs indices.

It is straightforward to show, with similar techniques to the one we just used, that the covariant derivative of an arbitrary tensor with p upstairs indices and q downstairs indices is given by using the two rules (6.65) and (6.75) for each index; (6.65) for each upstairs index, and (6.75) for each downstairs index.

To make clear what we mean by this, consider the two-index tensor g_{ij}. We use (6.75) for each downstairs index, giving

$$\nabla_k\,g_{ij} = \partial_k\,g_{ij} - \Gamma^\ell{}_{ki}\,g_{\ell j} - \Gamma^\ell{}_{kj}\,g_{i\ell}\,. \tag{6.76}$$

Actually this particular example, if we take g_{ij} to be the metric tensor, is exactly what we need next. We can now give an explicit contruction of the connection $\Gamma^i{}_{jk}$. We do this by making the additional requirement that we should like the metric tensor to be *covariantly constant*, $\nabla_k\,g_{ij} = 0$. This is a very useful property to have, since it means, for example, that if we look at the scalar product $V^i\,W^j\,g_{ij}$ of two vectors, we shall have

$$\nabla_k(V^i\,W^j\,g_{ij}) = (\nabla_k\,V^i)\,W^j\,g_{ij} + V^i\,(\nabla_k\,W^j)\,g_{ij}\,. \tag{6.77}$$

Remembering our rule that we shall in fact freely write $W^j\,g_{ij}$ as W_i, and so on, it should be clear that life would become a nightmare if the metric could not be taken freely through the covariant derivative!

Luckily, it turns out that all the things we have been asking for are possible. We can find a connection $\Gamma^i{}_{jk}$ that is symmetric in jk, gives us a covariant derivative that satisfies the Leibnitz rule, and for which $\nabla_k\, g_{ij} = 0$. We can find it just by juggling around the indices in equation (6.76). To do this, we write out $\nabla_k\, g_{ij} = 0$ using (6.76) three times, with different labellings of the indices:

$$
\begin{aligned}
\partial_k\, g_{ij} - \Gamma^\ell{}_{ki}\, g_{\ell j} - \Gamma^\ell{}_{kj}\, g_{i\ell} &= 0\,,\\
\partial_i\, g_{kj} - \Gamma^\ell{}_{ik}\, g_{\ell j} - \Gamma^\ell{}_{ij}\, g_{k\ell} &= 0\,,\\
\partial_j\, g_{ik} - \Gamma^\ell{}_{ji}\, g_{\ell k} - \Gamma^\ell{}_{jk}\, g_{i\ell} &= 0\,,
\end{aligned}
\tag{6.78}
$$

Now, add the last two equations and subtract the first one from this. Using the fact that $\Gamma^i{}_{jk}$ is symmetric in jk, we therefore get

$$
\partial_i\, g_{kj} + \partial_j\, g_{ik} - \partial_k\, g_{ij} - 2\Gamma^\ell{}_{ij}\, g_{k\ell} = 0\,. \tag{6.79}
$$

Multiplying this by the inverse metric g^{km}, we immediately obtain the following expression for $\Gamma^i{}_{jk}$ (after finally relabelling indices for convenience):

$$
\Gamma^i{}_{jk} = \tfrac{1}{2} g^{i\ell}\, (\partial_j\, g_{\ell k} + \partial_k\, g_{j\ell} - \partial_\ell\, g_{jk})\,. \tag{6.80}
$$

This is known as the *Christoffel Connection*, or sometimes the *Affine Connection*.

It is a rather simple matter to check that $\Gamma^i{}_{jk}$ defined by (6.80) does indeed have the required transformation property (6.71) under general coordinate transformations. Actually, there is really no need to check this point, since it is logically guaranteed from the way we constructed it that it must have this property. So we leave it as an "exercise to the reader," to verify by direct computation. The principle should be clear enough; one simply uses the expression for $\Gamma^i{}_{jk}$ given in (6.80) to calculate $\Gamma'^i{}_{jk}$, in terms of ∂'_i and g'_{ij} (which can be expressed in terms of ∂_i and g_{ij} using their standard tensorial transformation properties). It then turns out that $\Gamma'^i{}_{jk}$ is related to $\Gamma^i{}_{jk}$ by (6.71).

Notice that $\Gamma^i{}_{jk}$ is zero if the metric components g_{ij} are all constants. This explains why we never see the need for $\Gamma^i{}_{jk}$ if we only look at Cartesian tensors, for which the metric is just δ_{ij}. But as soon as we consider any more general situation, where the components of the metric tensor are functions of the coordinates, the Christoffel connection will become non-vanishing. Note that this does not necessarily mean that the metric has to be one on a curved space (such as the 2-sphere that we met earlier); even a flat metric written in "curvilinear coordinates" will have a non-vanishing Christoffel connection. As a simple example, suppose we take the metric on the plane,

$$
ds^2 = dx^2 + dy^2\,, \tag{6.81}
$$

and write it in polar coordinates (r,θ) defined by

$$x = r\,\cos\theta\,, \qquad y = r\,\sin\theta\,. \tag{6.82}$$

It is easy to see that (6.81) becomes

$$ds^2 = dr^2 + r^2\,d\theta^2\,. \tag{6.83}$$

If we label the (r,θ) coordinates as (x^1,x^2) then in the metric $ds^2 = g_{ij}\,dx^i\,dx^j$ we shall have

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}\,, \qquad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}\,. \tag{6.84}$$

Using (6.80), simple algebra leads to the following results:

$$\begin{aligned} &\Gamma^1{}_{11} = 0\,, \qquad \Gamma^1{}_{12} = 0\,, \qquad \Gamma^1{}_{22} = -r\,, \\ &\Gamma^2{}_{11} = 0\,, \qquad \Gamma^1{}_{12} = \frac{1}{r}\,, \qquad \Gamma^2{}_{22} = 0\,. \end{aligned} \tag{6.85}$$

Having obtained the Christoffel connection for this case, we can illustrate how one uses it by taking the example of the Laplacian. In Cartesian coordinates we know that the Laplacian of a function ψ is just $\partial_i\partial_i\,\psi$, which is again a scalar. Obviously, in general, we should find a generalisation of $\partial_i\partial_i\,\psi$ that is again a scalar. The answer, clearly, is that the Laplacian of ψ is

$$g^{ij}\,\nabla_i\,\partial_j\,\psi\,, \tag{6.86}$$

since by construction, we know that this is a scalar under general coordinate transformations. Notice that we don't need a covariant derivative for the ∂_j that acts directly on ψ, since that is already covariant. Thus we have in general that the Laplacian can be written as

$$g^{ij}\,\partial_i\,\partial_j\,\psi - g^{ij}\,\Gamma^k{}_{ij}\,\partial_k\,\psi\,. \tag{6.87}$$

Now, let us apply this to our simple example of the metric on the plane written in polar coordinates. Substituting from (6.84) and (6.85), we get

$$\partial_1^2\,\psi + \frac{1}{r^2}\,\partial_2^2\,\psi + \frac{1}{r}\,\partial_1\,\psi \tag{6.88}$$

where the last term is the one coming from the contribution of the Christoffel connection. Re-expressing this in a more readable language, we have

$$\frac{\partial^2\psi}{\partial r^2} + \frac{1}{r}\,\frac{\partial\psi}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2\psi}{\partial\theta^2}\,, \tag{6.89}$$

which can also be written as

$$\frac{1}{r}\frac{\partial}{\partial r}\Big(r\,\frac{\partial\psi}{\partial r}\Big)+\frac{1}{r^2}\,\frac{\partial^2\psi}{\partial\theta^2}\,. \tag{6.90}$$

This was, of course, an elaborate way to derive a simple and well-known result, but that was the whole point of the illustrative exercise; to show first how the new method works in a simple "toy" example.

In fact there is a nice way to express the Laplacian operator in general that doesn't require us to grind out all the components of the Christoffel connection. Notice from (6.87) that what we need for the Laplacian is the contracted set of quantities

$$g^{ij}\,\Gamma^k{}_{ij}\,, \tag{6.91}$$

and so from (6.80) we have

$$\begin{aligned} g^{ij}\,\Gamma^k{}_{ij} &= \tfrac12 g^{ij}\,g^{k\ell}\,(\partial_i\,g_{\ell j}+\partial_j\,g_{i\ell}-\partial_\ell\,g_{ij})\,,\\ &= g^{ij}\,g^{k\ell}\,\partial_i\,g_{\ell j}-\tfrac12 g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\ &= -g^{ij}\,g_{\ell j}\,\partial_i\,g^{k\ell}-\tfrac12 g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\ &= -\delta^i_\ell\,\partial_i\,g^{k\ell}-\tfrac12 g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,,\\ &= -\partial_\ell\,g^{k\ell}-\tfrac12 g^{k\ell}\,g^{ij}\,\partial_\ell\,g_{ij}\,. \end{aligned} \tag{6.92}$$

Note that in getting to the third line, we have use that $g^{k\ell}\,g_{\ell j}=\delta^k_j$, which is constant, and so $(\partial_i\,g^{k\ell})\,g_{\ell j}+g^{k\ell}\,(\partial_i\,g_{\ell j})=0$.

Now we use one further trick, which is to note that as a matrix expression, $g^{ij}\,\partial_\ell\,g_{ij}$ is just $\mathrm{tr}(\mathbf{g}^{-1}\,\partial_\ell\,\mathbf{g})$. But for any symmetric matrix we can write[26]

$$\det\mathbf{g}=e^{\mathrm{tr}\log\mathbf{g}}\,, \tag{6.93}$$

and so

$$\partial_\ell\,\det\mathbf{g}=(\det\mathbf{g})\,\mathrm{tr}(\mathbf{g}^{-1}\,\partial_\ell\,\mathbf{g})\,. \tag{6.94}$$

Thus we have

$$\tfrac12 g^{ij}\,\partial_\ell\,g_{ij}=\frac{1}{\sqrt{g}}\,\partial_\ell\,\sqrt{g}\,, \tag{6.95}$$

where we use the symbol g here to mean the determinant of the metric g_{ij}.

Putting all this together, we have

$$g^{ij}\,\nabla_i\,\partial_j\,\psi=g^{ij}\,\partial_i\,\partial_j\,\psi+(\partial_i\,g^{ij})\,\partial_j\,\psi+g^{ij}\,\frac{1}{\sqrt{g}}\,(\partial_i\,\sqrt{g})\,\partial_j\,\psi\,, \tag{6.96}$$

[26]Prove by diagonalising the matrix, so that $\mathbf{g}\longrightarrow\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$. This means that $\det\mathbf{g}=\prod_i\lambda_i$, while $e^{\mathrm{tr}\log\mathbf{g}}=e^{\sum_i\log\lambda_i}$, and so the result is proven.

after making some convenient relabellings of dummy indices. Now we can see that all the terms on the right-hand side assemble together very nicely, giving us the following simple expression for the Laplacian:

$$g^{ij}\,\nabla_i\,\partial_j\,\psi = \frac{1}{\sqrt{g}}\,\partial_i\Big(\sqrt{g}\,g^{ij}\,\partial_j\,\psi\Big)\,. \tag{6.97}$$

This general expression gives us the Laplacian in an arbitrary coordinate system, for an arbitrary metric.

As a first check, let us test it on the previous example of the two-dimensional plane with the metric $ds^2 = dr^2 + r^2\,d\theta^2$ in polar coordinates. From (6.84) we instantly see that the determinant of the metric is $g = r^2$, so plugging into (6.97) we get

$$\begin{aligned} g^{ij}\,\nabla_i\,\partial_j\,\psi &= \frac{1}{r}\,\partial_i\Big(r\,g^{ij}\,\partial_j\,\psi\Big)\,, \\ &= \frac{1}{r}\,\frac{\partial}{\partial r}\Big(r\,\frac{\partial\psi}{\partial r}\Big) + \frac{1}{r^2}\,\frac{\partial^2\psi}{\partial\theta^2}\,, \end{aligned} \tag{6.98}$$

in agreement with our previous result.

As a slightly less trivial example, consider Euclidean 3-space, written in terms of spherical polar coordinates (r,θ,ϕ). These, of course, are related to the Cartesian coordinates (X,Y,Z) by

$$X = r\,\sin\theta\,\cos\phi\,,\qquad Y = r\,\sin\theta\,\sin\phi\,,\qquad Z = \cos\theta\,. \tag{6.99}$$

The metric, written in terms of the spherical polar coordinates, is therefore

$$ds^2 = dr^2 + r^2\,d\theta^2 + r^2\,\sin^2\theta\,d\phi^2\,. \tag{6.100}$$

The determinant is therefore $g = r^4\,\sin^2\theta$ and so from (6.97) we get that the Laplacian is

$$\frac{1}{r^2}\,\frac{\partial}{\partial r}\Big(r^2\,\frac{\partial\psi}{\partial r}\Big) + \frac{1}{r^2}\,\Big[\frac{1}{\sin\theta}\,\frac{\partial}{\partial\theta}\,\Big(\sin\theta\,\frac{\partial\psi}{\partial\theta}\Big) + \frac{1}{\sin^2\theta}\,\frac{\partial^2\psi}{\partial\phi^2}\Big]\,. \tag{6.101}$$

6.5 The n-sphere, $SO(n+1)$ and Spherical Harmonics

6.5.1 The n-sphere and its symmetries

In an earlier discussion we looked in considerable detail at the construction of the 2-sphere, described as the surface $X^2 + Y^2 + Z^2 = 1$ in $\mathbb{R}^3$. All of that discussion can easily be generalised to the case of an n-dimensional sphere, defined by the surface

$$X^a\,X^a = 1\,, \tag{6.102}$$

in $\mathbb{R}^{n+1}$, where now of course the index a is understood to be summed over $(n+1)$ values. For convenience, we sometimes refer to the n-sphere as S^n.

Obviously much of our previous discussion of the symmetries carries over straightforwardly to the case of the n-sphere. The condition (6.102) is invariant under rotations defined by

$$X'^a = M_{ab}\, X^b\,, \tag{6.103}$$

where M_{ab} is an $O(n+1)$ matrix satisfying

$$M_{ab}\, M_{ac} = \delta_{bc}\,. \tag{6.104}$$

Infinitesimally we can again write $M_{ab} = \delta_{ab} + A_{ab}$, where the infinitesimal matrix A_{ab} is antisymmetric. This matrix has $\frac{1}{2}n(n+1)$ independent components, so we conclude that the dimension of the group $O(n+1)$ is

$$\dim(O(n+1)) = \tfrac{1}{2}n\,(n+1)\,. \tag{6.105}$$

By the dimension of the group, we mean the number of continuous parameters needed to specify a group element; we saw for $O(3)$ that the answer was 3. As in the case of $O(3)$, the group elements divide into those that have determinant $+1$, and those that have determinant -1. The former correspond to pure rotations in $\mathbb{R}^{n+1}$, while the latter correspond to rotations together with a reflection. Since the identity element obviously has determinant $+1$ it follows that all the infinitesimal transformations must be contained in $SO(n+1)$ too.

It would be quite complicated to generalise the spherical polar coordinates that we used on S^2 to the case of S^n, but in fact for many purposes we can perfectly well just use the Cartesian coordinates X^a on $\mathbb{R}^{n+1}$, together with the constraint (6.102). For example, we can write the infinitesimal $SO(n+1)$ transformations as $\delta X^a = \xi^a$, where $\xi^a = A_{ab}\, X^b$. Thus we are led to the Killing vectors K_{ab}, defined by

$$K_{ab} \equiv X^a\, \frac{\partial}{\partial X^b} - X^b\, \frac{\partial}{\partial X^a}\,, \tag{6.106}$$

where the ab indices here are labels, indicating which Killing vector we mean. By construction we have $\frac{1}{2}n(n+1)$ Killing vectors, since $K_{ab} = -K_{ba}$. This is the correct number for the $SO(n+1)$ symmetry of the n-sphere. If we specialise to the 2-sphere, it is easy to verify that the three Killing vectors K_{12}, K_{13} and K_{23} defined by (6.106) in this case are just the same, after the change to spherical polar coordinates, as the Killing vectors (6.32) that we derived previously.

Notice that the Killing vectors (6.106) are nothing but the angular momentum operators in $(n+1)$-dimensional Euclidean space. In 3 dimensions we would more commonly use the totally-antisymmetric epsilon tensor ϵ_{abc} to re-express the angular momentum operators in terms of a vector index:

$$L_a = \tfrac{1}{2}\epsilon_{abc}\, K_{bc} = \epsilon_{abc}\, X^b\, \frac{\partial}{\partial X^c}\,. \tag{6.107}$$

Observe, though, that it is a very special feature of 3 dimensions that one can replace an antisymmetric 2-index quantity like K_{ab} by a vector. In higher dimensions, where the corresponding totally-antisymmetric epsilon tensor has more indices, one cannot turn a 2-index antisymmetric tensor into a tensor with fewer indices. In fact this serves to emphasise that in a general dimension one should think of rotations as occurring *in planes*, rather than *about axes*. It is a coincidence of 3 dimensions that a rotation in the (x, y) plane can also be thought of as a rotation about the z axis.

6.5.2 Spherical Harmonics

When one first meets the spherical harmonics on the 2-sphere, it is generally in the context of performing a separation of variables in Laplace's equation or the wave equation, when using spherical polar coordinates. In fact we just re-derived the expression for this Laplacian in the previous section, in (6.101). After a standard separation of variables in which a function $\psi(r, \theta, \phi)$ is written as

$$\psi(r, \theta, \phi) = R(r)\, Y(\theta, \phi)\,, \tag{6.108}$$

Laplace's equation $\nabla^2\psi = 0$ becomes

$$\frac{1}{R}\,\frac{d}{dr}\Big(r^2\,\frac{dR}{dr}\Big) + \frac{1}{Y}\,\nabla^2_{S^2}\, Y = 0\,, \tag{6.109}$$

where $\nabla^2_{S^2}$ is the operator appearing in the large square brackets in (6.101), namely

$$\nabla^2_{S^2} = \frac{1}{\sin\theta}\,\frac{\partial}{\partial\theta}\Big(\sin\theta\,\frac{\partial}{\partial\theta}\Big) + \frac{1}{\sin^2\theta}\,\frac{\partial^2}{\partial\phi^2}\,. \tag{6.110}$$

In fact this operator is precisely the Laplacian for the unit 2-sphere, as may easily be checked by using our general formula (6.97), with the metric $ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2$. Introducing a separation constant λ in the usual way, one is led from (6.109) to consider the equation

$$-\nabla^2_{S^2}\, Y(\theta, \phi) = \lambda\, Y(\theta, \phi)\,. \tag{6.111}$$

This is the equation that determines the spherical harmonics.

A standard way to solve for the spherical harmonics is to write out the S^2 Laplacian $\nabla^2_{S^2}$ explicitly using (6.110), and perform a further separation of variables by writing $Y(\theta, \phi) =$

$P(\theta)\,\Phi(\phi)$. This introduces another separation constant m^2, an one is left to solve the equations

$$\begin{aligned} \sin\theta\,\frac{d}{d\theta}\Big(\sin\theta\,\frac{dP}{d\theta}\Big) + (\lambda\,\sin^2\theta - m^2)\,P &= 0\,, \\ \frac{d^2\Phi}{d\phi^2} + m^2\,\Phi &= 0\,. \end{aligned} \tag{6.112}$$

The latter has solutions of the form $e^{\mathrm{i}\,m\,\phi}$, and to get the proper periodicity under complete rotations $\phi \longrightarrow \phi + 2\pi$ on the sphere, we deduce that m must be an integer. After letting $x = \cos\theta$ the first equation becomes the generalised Legendre equation,

$$\frac{d}{dx}\Big((1-x^2)\,\frac{dP}{dx}\Big) + \Big(\lambda - \frac{m^2}{1-x^2}\Big)\,P = 0\,. \tag{6.113}$$

After a considerable labour, involving, for example, a careful study of the solutions for this equation obtained as a series expansion (discussed at length in Part I of the course), one concludes that for the functions $P(\theta)$ to be regular at $\theta = 0$ and π (the north and south poles of the sphere), the separation constant λ must be of the form $\lambda = \ell\,(\ell+1)$, where ℓ is an integer, and $-\ell \le m \le \ell$. Thus after a rather involved chain of argument, one arrives at the spherical harmics $Y_{\ell m}(\theta,\phi)$ being the complet set of regular eigenfunctions of the Laplacian $\nabla^2_{S^2}$ on S^2, with

$$-\nabla^2_{S^2}\,Y_{\ell m} = \ell\,(\ell+1)\,Y_{\ell m}\,. \tag{6.114}$$

Of course one has the feature that since m does not appear in the expression for the eigenvalues, there is a $(2\ell+1)$-fold degeneracy for the spherical harmonics with a given value of ℓ, since m can take any of the integer values between $-\ell$ and $+\ell$.

This traditional approach to constructing the spherical harmonics is a rather calculational one, which provides very little group-theoretic insight into what is going on. We are in fact now in a position to give a much simpler, and more elegant, construction of the spherical harmonics, which provides us with a rather clear picture of them as representations of the symmetry group $SO(3)$ of the 2-sphere. In fact it is just as easy to construct the spherical harmonics on all the spheres S^n, for arbitrary dimension n, so there is that advantage too.

We have described the unit n-sphere as the surface $X^a\,X^a = 1$ in $\mathbb{R}^{n+1}$. Let us write the metric on the unit n-sphere as $d\Omega^2$. It is evident that this is related to the Cartesian metric ds^2 on $\mathbb{R}^{n+1}$ by

$$ds^2 = dr^2 + r^2\,d\Omega^2\,, \tag{6.115}$$

where $X^a\,X^a = r^2$. This is clear, if you think about how we would measure distances in $\mathbb{R}^{n+1}$ if it were written in "hyperspherical polar coordinates," r and y^α, where y^α represent the set of angular that one would use to parameterise points on the unit n-sphere. The square of the distance between two infinitesimally separated points in $\mathbb{R}^{n+1}$ is therefore the sum of the square of the radial-coordinate separation dr, and the square of the distance in the surface of the sphere that separates the two points. Since $d\Omega^2$ is the metric on the unit sphere, the distance on the sphere of radius r, where the two points are located, will be scaled by the factor r. It is easy to see that (6.115) reduces to familiar cases if we consider $\mathbb{R}^2$ and $\mathbb{R}^3$, since the metrics on the unit 1-sphere and 2-sphere are just

$$\begin{aligned} \text{1-sphere :} \qquad & d\Omega^2 = d\theta^2\,, \\ \text{2-sphere :} \qquad & d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2\,. \end{aligned} \tag{6.116}$$

Luckily we don't ever need to define the angular coordinates on S^n explicitly, in order to solve for the spherical harmonics. We can just let them be called y^α, with $1 \le \alpha \le n$, but we don't need to define how they are related to the Cartesian coordinates X^a in $\mathbb{R}^{n+1}$. (One can usefully have in mind, though, the picture that they will be defined very analogously to the way spherical polar coordinates are related to the (X, Y, Z) coordinates on $\mathbb{R}^3$.) The metric on the unit n-sphere can then be written as

$$d\Omega^2 = h_{\alpha\beta}\, dy^\alpha\, dy^\beta\,. \tag{6.117}$$

The full set of $(n+1)$ hyperspherical coordinates on $\mathbb{R}^{n+1}$ will be (r, y^α). Let us call these hyperspherical coordinates x^i, with i running from 0 to n:

$$x^0 \equiv r\,, \qquad x^\alpha \equiv y^\alpha\,. \tag{6.118}$$

Now, using (6.117), the metric (6.115) on $\mathbb{R}^{n+1}$ is

$$ds^2 = dr^2 + r^2\, h_{\alpha\beta}\, dy^\alpha\, dy^\beta\,. \tag{6.119}$$

Clearly therefore the determinant g of this metric is given by

$$g = r^n\, h\,, \tag{6.120}$$

where h is the determinant of the metric $h_{\alpha\beta}$ on the unit n-sphere. Plugging into our general expression (6.97) for the Laplacian, we therefore find that in these hyperspherical polar coordinates, the Laplacian on $\mathbb{R}^{n+1}$ is given by

$$\nabla^2_{R^{n+1}} = \frac{1}{r^n}\,\frac{\partial}{\partial r}\Big(r^n\,\frac{\partial}{\partial r}\Big) + \frac{1}{r^2}\,\nabla^2_{S^n}\,, \tag{6.121}$$

where

$$\nabla^2_{S^n} \equiv \frac{1}{\sqrt{h}}\,\frac{\partial}{\partial y^\alpha}\Big(\sqrt{h}\, h^{\alpha\beta}\,\frac{\partial}{\partial y^\beta}\Big) \tag{6.122}$$

is the Laplacian on the unit n-sphere. (The special cases for $n=1$ and $n=2$ appear in our examples (6.98) and (6.101) that we looked at previously.)

Having obtained this relation between the Laplacians on $\mathbb{R}^{n+1}$ and S^n, the problem of constructing the spherical harmonics is almost solved. First, we introduce the following functions Ψ on $\mathbb{R}^{n+1}$:

$$\Psi(X) = T_{a_1 a_2 \cdots a_\ell}\, X^{a_1}\, X^{a_2} \cdots X^{a_\ell}\,, \tag{6.123}$$

where $T_{a_1 a_2 \cdots a_\ell}$ is an ℓ-index constant tensor in $\mathbb{R}^{n+1}$ which is completely arbitrary except for satisfying the following two conditions:

(1) $T_{a_1 a_2 \cdots a_\ell}$ is *totally symmetric* in all its indices.

(2) The tensor T is totally traceless, in the sense that the contraction of any pair of indices on $T_{a_1 a_2 \cdots a_\ell}$ gives zero:

$$\delta_{a_1 a_1}\, T_{a_1 a_2 \cdots a_\ell} = 0\,, \quad \text{etc.}\,. \tag{6.124}$$

Clearly condition (1) is simply making sure that we eliminate all the "redundant baggage" in $T_{a_1 a_2 \cdots a_\ell}$. Since it appears in (6.123) contracted onto the totally symmetrical product $X^{a_1}\, X^{a_2} \cdots X^{a_\ell}$, it is obvious that any part of $T_{a_1 a_2 \cdots a_\ell}$ that was not totally symmetrical in the indices would give no contribution anyway.

Condition (2) serves a different purpose. It implies that if we act with the $\mathbb{R}^{n+1}$ Laplacian $\nabla^2_{R^{n+1}}$ on Ψ, we shall get zero:

$$\nabla^2_{R^{n+1}}\, \Psi = 0\,. \tag{6.125}$$

This is because from the definition of Ψ in (6.123), we shall clearly have

$$\begin{aligned} \frac{\partial \Psi}{\partial X_a} &= T_{a a_2 \cdots a_\ell}\, X^{a_2} \cdots X^{a_\ell} + T_{a_1 a \cdots a_\ell}\, X^{a_1}\, X^{a_3} \cdots X^{a_\ell} + \cdots T_{a_1 a_2 \cdots a_\ell}\, X^{a_1}\, X^{a_2} \cdots X^{a_{\ell-1}} \\ &= \ell\, T_{a a_2 \cdots a_\ell}\, X^{a_2} \cdots X^{a_\ell}\,, \end{aligned} \tag{6.126}$$

(all the ℓ terms are equal, because of the total symmetry). Acting with another derivative, we therefore get

$$\frac{\partial^2 \Psi}{\partial X^a\, \partial X^b} = \ell\,(\ell-1)\, T_{a b a_3 \cdots a_\ell}\, X^{a_3} \cdots X^{a_\ell}\,. \tag{6.127}$$

(This time, we have immediately used the symmetry of T to collect the $(\ell-1)$ terms that appear from the second differentiation together. Now we see that the $\mathbb{R}^{n+1}$ Laplacian acting on Ψ gives zero:

$$\nabla^2_{R^{n+1}}\, \Psi = \frac{\partial^2 \Psi}{\partial X^a \partial X^a} = \ell\,(\ell-1)\, \delta_{ab}\, T_{a b a_3 \cdots a_\ell}\, X^{a_3} \cdots X^{a_\ell} = 0\,, \tag{6.128}$$

by virtue of condition (2) above.

Now, it only remains to make the following observation. Since the function Ψ defined in (6.123) involes a product of ℓ Cartesian coordinates X^a, it is evident that it must be expressible as

$$\Psi(X) = r^\ell\, \psi(y)\,, \tag{6.129}$$

where y represents the angular coordinates y^α on the unit n-sphere, and $\psi(y)$ is *independent of* r. Again, it is helpful to have in mind the $\mathbb{R}^3$ example here, where we have

$$X = r\, \sin\theta\, \cos\phi\,, \qquad Y = r\, \sin\theta\, \sin\phi\,, \qquad Z = r\, \cos\theta\,. \tag{6.130}$$

Finally, since we have established that the $\mathbb{R}^{n+1}$ Laplacian annihilates Ψ we simply have to substitute it into (6.121) to deduce that

$$\frac{1}{r^n}\,\frac{d}{dr}\Big(r^n\,\frac{dr^\ell}{dr}\Big)\,\psi + \frac{1}{r^2}\,r^\ell\,\nabla^2_{S^n}\,\psi = 0\,. \tag{6.131}$$

Hence we arrive at the conclusion that ψ is an eigenfunction of the Laplacian on the unit n-sphere, satisfying

$$-\nabla^2_{S^n}\,\psi = \ell\,(\ell+n-1)\,\psi\,. \tag{6.132}$$

Notice that is we take $n = 2$, corresponding to the 2-sphere, we reproduce the familiar eigenvalues $\ell\,(\ell+1)$.

Two issues remain to be discussed here. The first is that we have certainly produced *some* eigenfunctions on the n-sphere by this method, but have we obtained them all? The answer is yes, and it can be seen as follows. Clearly, any regular function on the unit n-sphere can be smoothly extended out as a regular function on $\mathbb{R}^{n+1}$. Conversely, if we consider the set of all regular functions on $\mathbb{R}^{n+1}$, they will project down so as to provide us with all possible regular functions on S^n. Now, the regular functions $f(X)$ on $\mathbb{R}^{n+1}$ can certainly be expanded in a Taylor series, which will give a sum of terms of the form (6.123), summed over all $\ell \ge 0$ (without yet imposing the tracelessness of condition (2) above):

$$f(X) = \sum_{\ell=0}^\infty f_\ell(X)\,, \tag{6.133}$$

where

$$f_\ell(X) \equiv T_{a_1 a_2\cdots a_\ell}\, X^{a_1}\, X^{a_2}\cdots X^{a_\ell}\,, \tag{6.134}$$

But the imposition of tracelessness on $T_{a_1 a_2\cdots a_\ell}$ is just a matter of organising the terms in the sum, since a pure trace contribution in the term $f_\ell(X)$ would correspond to r^2 times a term of the form $f_{\ell-2}(X)$. By the time we restricted to the unit n-sphere, by setting $r=1$, this from f_ℓ term would therefore just be repeating what had already been constructed in $f_{\ell-2}$. So from the viewpoint of constructing regular functions on the n-sphere, the imposition of tracelessness on the tensors $T_{a_1 a_2 \cdots a_\ell}$ is just a matter of avoiding double-counting. Thus we can be sure that our construction of scalar eigenfunctions of the Laplacian on S^n has given *all* all the eigenfunctions. The functions ψ, defined by (6.123) and (6.129), then, give the complete set of *spherical harmonics* on S^n.

The second issue that we must still address concerns the degneracies of the eigenvalues, or, equivalently, the *multiplicities* of the eigenfunctions ψ for a given value of the integer ℓ. This is easily worked out, since it is just a matter of counting how many independent components the constant tensor $T_{a_1 a_2 \cdots a_\ell}$ has, bearing in mind the two conditions of symmetry and tracelessness that we imposed. It is easy to see that a totally-symmetric tensor with ℓ indices that each run over $(n+1)$ values has

$$\frac{(n+1)(n+2)\cdots(n+\ell)}{\ell!} \tag{6.135}$$

independents components. When we impose the traceless condition on such a tensor, we therefore impose a number of conditions equal to the number of independent components in a similar tensor that has only $(\ell-2)$ indices. Thus the number of independent components in our tensor $T_{a_1 a_2 \cdots a_\ell}$ that is totally symmetric and traceless is

$$\begin{aligned} d_\ell &= \frac{(n+1)(n+2)\cdots(n+\ell)}{\ell!} - \frac{(n+1)(n+2)\cdots(n+\ell-2)}{(\ell-2)!}\,, \\ &= \frac{(n+1)(n+2)\cdots(n+\ell-2)}{\ell!}\Big((n+\ell-1)({\scriptstyle (n)}+\ell) - \ell\,(\ell-1)\Big)\,, \\ &= \frac{n\,(n+1)(n+2)\cdots(n+\ell-2)(2\ell+n-1)}{\ell!}\,, \end{aligned} \tag{6.136}$$

which can be written as

$$d_\ell = \frac{(2\ell+n-1)\,(n+\ell-2)!}{\ell!\,(n-1)!}\,. \tag{6.137}$$

This gives us the multiplicity of the eigenfunctions ψ with the specific eigenvalue

$$\lambda_\ell = \ell\,(\ell+n-1) \tag{6.138}$$

that we found above. Notice that if we specialise to the case of the 2-sphere, equation (6.137) reduces to

$$\text{2-sphere:} \qquad d_\ell = 2\ell+1\,, \tag{6.139}$$

as we know it should.

6.5.3 Irreducible Representations of $SO(N)$

The construction of the eigenfunctions that we have obtained here, and the results for the multiplicities of the eigenvalues, have a deeper significance than might at first be apparent. What we have actually been doing here is constructing *irreducible representations* of the the symmetry groups $SO(n+1)$ of the n-spheres. To be a bit more precise, the sets of tensors $T_{a_1 a_2 \cdots a_\ell}$ that we have been using are themselves irreducible representations of $SO(n+1)$. More generally, one can consider many different classes of constant tensor $H_{a_1 a_2 \cdots a_p}$ in $\mathbb{R}^{n+1}$, and associate them with irreducible representations.

To make life a little simpler, let us talk about $SO(N)$ rather than $SO(n+1)$. If we begin with the tensor $H_{a_1 a_2 \cdots a_p}$ in $\mathbb{R}^N$, and make no symmetry or tracelessness requirement at all on it, then the number of independent components for such a tensor will simply be N^p, since each index can range over N values. This set of tensors with N^p components is a representation of $SO(N)$, but it is not irreducible; we can divide it into smaller self-contained subsets of components. The rules for how such subdivisions can be made are very simple. We can do anything as long as it respects $SO(N)$ covariance. What this means is that we have to treat the indices in a totally "democratic" way, and we cannot single out any one index value, or subset of index values, for special treatment.

Let us take a concrete example. Suppose we take a 2-index tensor H_{ab} in $\mathbb{R}^N$, which has N^2 independent components. Is this reducible, or is it already as irreducible as can be? First, the sort of things we *cannot* do is to pick an index value, say $a=1$, and treat that as special. We cannot divide H_{ab} into $H_{\alpha\beta}$, $H_{1\alpha}$, $H_{\alpha 1}$ and H_{11}, where $2 \le \alpha \le N$, and claim that we are decomposing H_{ab} into representations of $SO(N)$; clearly what we are doing here is not covariant from an $SO(N)$ point of view. What we *can* do, however, is to write H_{ab} as the sum of its symmetric and antisymmetric parts:

$$H_{ab} = S_{ab} + A_{ab}\,, \tag{6.140}$$

where

$$S_{ab} \equiv \tfrac{1}{2}(H_{ab} + H_{ba})\,, \qquad A_{ab} \equiv \tfrac{1}{2}(H_{ab} - H_{ba})\,. \tag{6.141}$$

Now, we can count the number of independent components in S_{ab}, namely $\frac{1}{2}N(N+1)$, and the number of independent components in A_{ab}, namely $\frac{1}{2}N(N-1)$. Of course the sum of these two gives us back the original number of components for the unrestricted tensor H_{ab}:

$$\tfrac{1}{2}(N(N+1) + \tfrac{1}{2}N(N-1) = N^2\,. \tag{6.142}$$

Clearly the decomposition in (6.140) is completely covariant with respect to $SO(N)$, since it is a tensorial equation, so it is a perfectly allowable subdivision for us to make.

Have we finished? Not quite, because there is one more thing we can do that respects the covariance, and that is to extract the trace from the symmetric tensor S_{ab}. Thus we can write

$$S_{ab} = \widetilde{S}_{ab} + \frac{1}{N}\, S\, \delta_{ab}\,, \tag{6.143}$$

where S is the trace of S_{ab}, namely

$$S \equiv \delta_{ab}\, S_{ab}\,.\,, \tag{6.144}$$

and so by construction $\widetilde{S}_{ab}$ is traceless,

$$\delta_{ab}\, \widetilde{S}_{ab} = 0\,. \tag{6.145}$$

Clearly (6.143) and (6.144) are both perfectly $SO(N)$-covariant equations; they transform covariantly under $SO(N)$ rotations. (We are really back to "kindergarten" Cartesian tensors here!)

With this extraction of the trace, we have reached the end of the road for the decomposition of the original 2-index tensor H_{ab}. In other words, we have found that it splits into three irreducible representations of $SO(N)$, with dimensions

$$\dim(A_{ab}) = \tfrac{1}{2}N(N-1)\,, \qquad \dim(\widetilde{S}_{ab}) = \tfrac{1}{2}(N-1)(N+2)\,, \qquad \dim(S) = 1\,, \tag{6.146}$$

These are the dimensions of the 2-index antisymmetric representation, the 2-index symmetric traceless representation, and the singlet of $SO(N)$ respectively.

The original H_{ab} representation is really to be thought of as the product of two 1-index representations. The 1-index, or *vector representation* of $SO(N)$ corresponds, as its name implies, to taking an arbitrary constant vector H_a in $\mathbb{R}^n$. It is clear that we cannot subdivide this representation any further by means of any allowable covariant rules, and so it is an N-dimensional irreducible representation.

We have just met four different irreducible representations of $SO(N)$, and we have seen that the following multiplication rule applies:

$$\underline{N} \times \underline{N} = \underline{\tfrac{1}{2}N(N-1)} + \underline{\tfrac{1}{2}(N-1)(N+2)} + \underline{1}\,. \tag{6.147}$$

What this is saying is that the product of the vector representation of $SO(N)$ with itself gives the three irreducible representations whose dimensions are listed on the right-hand side. For example, in $SO(3)$ we have

$$\underline{3} \times \underline{3} = \underline{3} + \underline{5} + \underline{1}\,. \tag{6.148}$$

Note that we use the underlining notation to indicate that we are talking about group representation here.[27]

One can continue the process of examining $SO(N)$ tensors with more and more indices, in each case making a covariant decomposition into the largest possible number of irreducible pieces, and thereby one builds up the complete set of irreducible representations of $SO(N)$. It gets a little trickier than the examples we have looked at so far, once the tensor has several indices. For example, consider a 3-index tensor H_{abc}. This certainly contains a totally-symmetric piece, and a totally antisymmetric piece, but it also has more. This can easily be seen by noting that sum of the independent components $\frac{1}{6}N(N+1)(N+2)$ of a symmetric 3-index tensor and the independent components $\frac{1}{6}N(N-1)(N-2)$ of an antisymmetric 3-index tensor does not add up to the N^3 components of an arbitrary 3-index tensor. There is nothing deep or mysterious about this, of course, and it is really just an exercise in symmetries and combinatorics to work out what the "extra" pieces are. Of course one also needs to extract all trace terms where appropriate, and count those as separate irreducible pieces. A very hand diagrammatic method, known as *Young Tableaux*, has been developed for doing all this. However, it takes us beyond the scope of this introductory discussion, so we shall leave it at that.

For our present purposes we don't need anything terribly exotic, because we saw that in the contruction of the spherical harmonics it was the totally symmetric and traceless $SO(n+1)$ tensors $T_{a_1 a_2 \cdots a_\ell}$ that were relevant. What we have now learned from the above discussion is that these tensors are actually giving us irreducible representations of $SO(n+1)$, and we have already worked out their dimensions d_ℓ in (6.137). For the 2-sphere, this became $d_\ell = 2\ell+1$, and so what we are seeing is that the spherical harmonics on S^2 occur in the following irreducible representations of $SO(3)$:

$$d_\ell = 2\ell + 1 = \underline{1}\,,\underline{3}\,,\underline{5}\,,\underline{7}\,,\ldots \tag{6.149}$$

As the dimension $d_\ell = 2\ell+1$ of the representation gets bigger, so, correspondingly, does the eigenvalue $\lambda_\ell = \ell\,(\ell+1)$.

For the higher-dimensional n-spheres the dimensions of the symmetric traceless irreducible $SO(n+1)$ representations become a bit more interesting. For example, from d_ℓ given in (6.137) we have the following:

$$SO(4): \qquad d_\ell = (\ell+1)^2 = \underline{1}\,,\underline{4},\,,\underline{9}\,,\underline{16}\,,\ldots$$

[27] It also serves to show that we are doing profound mathematics here, and that we have not reverted to the kindergarten arithmetic class!

$$SO(5): \qquad d_\ell = \tfrac{1}{6}(\ell+1)(\ell+2)(2\ell+3) = \underline{1}\,,\underline{5}\,,\underline{14}\,,\underline{30}\,,\ldots \tag{6.150}$$
$$SO(6): \qquad d_\ell = \tfrac{1}{12}(\ell+1)(\ell+2)^2(\ell+3) = \underline{1}\,,\underline{6}\,,\underline{20}\,,\underline{50}\,,\ldots$$

These examples are the first few representations of the spherical harmonics on the 3-sphere, 4-sphere and 5-sphere respectively.

We shall bring this course to a conclusion with a brief discussion of two topics related closely to what has gone before. Each deserves an entire course in its own right, so clearly what will be said here will be very sketchy. The first of the topics is local gauge symmetries, and the second is Riemann curvature, and general relativity.

6.6 Gauge Invariance and Covariant Derivative in Quantum Mechanics

We met the covariant derivative in the context of the differentiation of general-coordinate tensors; it was necessary to introduce it in order to be able take derivatives of tensors and get tensors again. Exactly the same basic notion of a covariant derivative arises also in other contexts. Perhaps the simplest of these is in quantum mechanics, when we consider a wavefunction for a charged particle in the presence of an electromagnetic field.

Consider first the very simple situation of the Schrödinger equation for a free particle,[28]

$$-\frac{\hbar^2}{2m}\,\vec\nabla^2\,\psi = \mathrm{i}\,\hbar\,\frac{\partial\psi}{\partial t}\,. \tag{6.151}$$

Obviously we are free to multiply the wavefunction ψ by an arbitrary constant complex number of modulus 1, without changing anything physically;

$$\psi \longrightarrow \psi' = U\,\psi\,, \qquad |U| = 1\,. \tag{6.152}$$

We can write such a constant as

$$U = e^{\mathrm{i}\,\alpha}\,, \tag{6.153}$$

where α is a constant real number, which may as well be restricted to lie in the range $0 \le \alpha < 2\pi$. The constant U is a 1×1 unitary matrix, since it satisfies $U^\dagger\,U = 1$. It is in fact an element of the group $U(1)$.

It was important in the transformation (6.152) that U should be a *constant*, so that it can pass freely through the derivatives in the Schrödinger equation (6.151), thus ensuring that ψ' satisfies the same equation:

$$-\frac{\hbar^2}{2m}\,\vec\nabla^2\,\psi' = \mathrm{i}\,\hbar\,\frac{\partial\psi'}{\partial t}\,. \tag{6.154}$$

[28]In this section we shall be assuming that we are working in flate Euclidean space, with Cartesian coordinates, so $\vec\nabla$ here just means the usual gradient operator of Cartesian vector analysis.

Such constant phase transformations are called *global* $U(1)$ transformations, or sometimes *rigid* $U(1)$ transformations.

Suppose, now, that we want to generalise the idea of the phase transformation (6.152), to the case where we allow the unit-phase quantity U to be dependent on the spatial position, and on time. Such a transformation is then called a *local* $U(1)$ transformation. Obviously as it stands this will give trouble in the Schrödinger equation, since now when we substitute (6.152) into (6.151), we will pick up terms where the space and time derivatives land on the phase factor U. These terms will prevent the transformed wavefunction ψ' from satisfying the simple primed equation (6.154).

This discussion should sound rather familiar. It is exactly like the situation we faced with derivatives of general-coordinate tensors, where the derivative landing on the transformation matrix $\partial x'^i/\partial x^j$ spoilt the tensor-transformation properties. Here, the problem is analogous, namely that $(\partial_i\psi')$ is not coming out to be the same as $(\partial_i\psi)'$. In the case of general-coordinate tensor, we introduced a covariant derivative to solve the problem, and that is exactly what we can do here too. Thus we shall define[29]

$$D_i\,\psi \equiv \partial_i\,\psi - \frac{\mathrm{i}\,e}{\hbar}\,A_i\,\psi\,,\qquad D_0\,\psi \equiv \frac{\partial\psi}{\partial t} + \frac{\mathrm{i}\,e}{\hbar}\,\phi\,\psi\,. \tag{6.155}$$

We now require that A_i and ϕ should transform under the local $U(1)$ transformation, in precisely such a way as to give us what we want, which is

$$(D_i\,\psi)' = U\,D_i\,\psi\,,\qquad (D_0\psi)' = U\,D_0\psi\,. \tag{6.156}$$

Let us look at D_i first. Writing out what we require for D_i in (6.156) we have

$$\begin{aligned} D_i'\,\psi' &= (\partial_i - \frac{\mathrm{i}\,e}{\hbar}\,A_i')\,(U\,\psi)\,,\\ &= U\,\Big(\partial_i\,\psi - \frac{\mathrm{i}\,e}{\hbar}\,A_i'\,\psi + U^{-1}\,(\partial_i U)\,\psi\Big)\,,\\ &= U\,\Big(\partial_i - \frac{\mathrm{i}\,e}{\hbar}\,A_i\Big)\,\psi + \Big[U^{-1}\,(\partial_i U) + \frac{\mathrm{i}\,e}{\hbar}\,(A_i - A_i')\Big]\,\psi\,,\\ &= U\,D_i\,\psi\Big[U^{-1}\,(\partial_i U) + \frac{\mathrm{i}\,e}{\hbar}\,(A_i - A_i')\Big]\,\psi\,. \end{aligned} \tag{6.157}$$

The first term on the bottom line is exactly what we want, so we must require that the quantity in square brackets be zero. In other words, A_i should have the following transformation

[29]For now, the quantities A_i and ϕ are just a 3-vector and a scalar, introduced for the purpose of allowing us to make local $U(1)$ transformations. Any similarity to things that may be familiar from electromagnetism is entirely non-coincidental, but here we are going to *derive* electromagetism from the requirement of local $U(1)$ invariance.

property under the local $U(1)$ transformation:

$$A_i' = A_i - \frac{\mathrm{i}\,\hbar}{e}\, U^{-1}\,\partial_i U\,. \tag{6.158}$$

If we parameterise U in the following way,

$$U = e^{\mathrm{i}\,e\lambda/\hbar}\,, \tag{6.159}$$

where λ is the local parameter, then we see that (6.158) is nothing but

$$A_i' = A_i + \partial_i\,\lambda\,. \tag{6.160}$$

In an identical fashion, we can derive the required local $U(1)$ transformation of the function ϕ in the covariant time derivative D_0 in (6.155), from the second equation in (6.156). We find

$$\phi' = \phi - \frac{\partial\lambda}{\partial t}\,. \tag{6.161}$$

We can recognise (6.160) and (6.161) as being precisely the gauge transformation rules of the magnetic vector potential $\vec{A}$ and the electrostatic potential ϕ of electrodynamics:

$$\vec{A}' = \vec{A} + \vec{\nabla}\,\lambda\,, \qquad \phi' = \phi - \frac{\partial\lambda}{\partial t}\,. \tag{6.162}$$

We have effectively *derived* electromagnetism, but purely from the considerations of local $U(1)$ invariance in quantum mechanics.

The final step is to write out our new version of the Schrödinger equation, using the covariant derivative. Thus in (6.151) we replace the ordinary derivatives by covariant derivatives:

$$-\frac{\hbar^2}{2m}\,D_i\,D_i\,\psi = \mathrm{i}\,\hbar\,D_0\,\psi\,. \tag{6.163}$$

It is now manifest, from the known covariance properties of the transformations in (6.156), that after performing an arbitrary local $U(1)$ transformation the Schrödinger equation (6.163) will simply take the same form, but now with primes on ψ and the covariant derivatives. Note that (6.163) is nothing but

$$-\frac{\hbar^2}{2m}\Big(\vec{\nabla} - \frac{\mathrm{i}\,e}{\hbar}\,\vec{A}\Big)^2\,\psi + e\,\phi\,\psi = \mathrm{i}\,\hbar\,\frac{\partial\psi}{\partial t}\,, \tag{6.164}$$

which is the Schrödinger equation for a charge particle in an electromagnetic field.

6.7 Curvature, the Riemann Tensor, and General Relativity

We have seen how the Christoffel coonection $\Gamma^i{}_{jk}$ allows us to define a covariant derivative, thereby permitting an extension of the idea that is familiar in Cartesian tensor analysis that the derivative operator provides a mapping from tensors into new tensors. We have seen also that the Christoffel connection is non-vanishing not only for a metric on a curved space such as a sphere, but even for a flat metric that happens to be expressed in a non-Cartesian coordinate system, such as polar coordinates on the plane.

So, for example, if we start with the flat metric on the plane written in Cartesian coordinates, $ds^2 = dx^2 + dy^2$, and then make the standard transformation to polar coordinates, we find that the originally-vanishing Christoffel connection becomes non-vanishing after the coordinate transformation. The fact that this can happen is a reflection of the non-tensorial nature of the connection. By contrast, if a *tensor* were vanishing in one coordinate frame, it would have to remain zero in all coordinate frames. This can be seen immediately from its transformation law, (6.43).

How *do* we characterise the idea of whether the space is intrinsically curved, or not? Of course one approach would be to take the given metric and try making coordinate transformations in order to see whether it can be re-expressed as the flat metric in some Cartesian coordinate system. But that would be a very clumsy thing to do in general, and the mere fact that one failed to find a coordinate transformation that did the job might mean nothing more than that one had not tried hard enough. Besides, it would not be an approach that would provide very much insight into the structure of the metric, especially if it turned out that it was *not* merely flat space in a funny coordinate system.

It should come as no surprise, in the light of the previous observations, that the way to characterise the curvature of a space is by means of a *tensor* quantity. The required object, called the *Riemann Tensor*, has four indices, with certain symmetry properties, and is denoted by $R^i{}_{jk\ell}$. If the metric is flat then the Riemann tensor is zero. Since it *is* a tensor, this vanishing is unaltered under any general coordinate transformation, and so it provides a genuinely coordinate-independent test for whether the metric is capable of being transformed into the standard Cartesian metric by a suitable coordinate transformation. At least as importantly, however, a non-vanishing Riemann tensor provides very useful information about a space that *is* curved.

How do we define the Riemann tensor? It turns out that it can be constructed by taking a derivative of the Christoffel connection, in an appropriate way. Specifically, it is given by

$$R^i{}_{jk\ell} = \partial_k \Gamma^i{}_{j\ell} - \partial_\ell \Gamma^i{}_{jk} + \Gamma^i{}_{km}\Gamma^m{}_{j\ell} - \Gamma^i{}_{\ell m}\Gamma^m{}_{jk}\,. \tag{6.165}$$

Looking at this, it is not manifestly apparent that it should be a tensor at all. After all, it is constructed by taking partial derivatives of something that is itself not a tensor. Remarkably, however, it turns out that this *is* a tensor. In principle, it can be proven by the time-honoured method of calculating it in a primed coordinate frame, using the knownm transformation properties of ∂_i and $\Gamma^i{}_{jk}$, and showing that it is related to the components in the original unprimed frame in the way it should be for a tensor. There is nothing conceptually difficult involved in checking this, but it is somewhat tedious. We shall leave it as an exercise for the interested reader.

The first thing to notice from (6.165) is that the Riemann tensor is indeed obviously zero if we take g_{ij} to be the flat metric in Cartesian coordinates, $g_{ij} = \delta_{ij}$, since already that means that $\Gamma^i{}_{jk} = 0$, as we saw before. Together with the knowledge that $R^i{}_{jk\ell}$ really is a tesnor, this shows that $R^i{}_{jk\ell} = 0$ for flat space in *any* coordinate system.

There are further tensor quantities that can be construced from the Riemann tesnor, by making index contractions. These therefore contain less information than the full Riemann tensor, but they are nevertheless of great importance. First, we can define the *Ricci Tensor*,

$$R_{ij} = R^k{}_{ikj}\,. \tag{6.166}$$

One can show from the definition of the Riemann tensor that R_{ij} is actually symmetric in its two indices, $R_{ij} = R_{ji}$. By contracting with the inverse metric we can also form a scalar, called the *Ricci Scalar* R, given by

$$R = g^{ij}\,R_{ij}\,. \tag{6.167}$$

The Riemann tensor itself also has certain symmetries. To state these, it is convenient we lower the first index, defining (in the standard way)

$$R_{ijk\ell} = g_{im}\,R^m{}_{jk\ell}\,. \tag{6.168}$$

The symmetries are then:

$$\begin{aligned} R_{ijk\ell} &= R_{k\ell ij} = -R_{jik\ell} = -R_{ij\ell k}\,, \\ R_{ijk\ell} &+ R_{ik\ell j} + R_{i\ell jk} = 0\,, \end{aligned} \tag{6.169}$$

all of which can, with some algebra, be proven from the previous definitions. Thus $R_{ijk\ell}$ is symmetric under the interchange of the first pair of indices with the second pair, and it is antisymmetric under the exchange of the first two indices, and under the exchange of the last two indices. It also has the cyclic symmetric given in the second line.

Let us consider the 2-sphere, with the metric $ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2$, as an example. Taking the coordinates to be $x^1 = \theta$, $x^2 = \phi$, we have

$$g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}, \qquad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2\theta} \end{pmatrix}. \tag{6.170}$$

Simple algebra using (6.80) leads to the following results for the comonents of the Christoffel connection:

$$\begin{aligned} &\Gamma^1{}_{11} = 0\,, \qquad \Gamma^1{}_{12} = 0\,, \qquad \Gamma^1{}_{22} = -\sin\theta\,\cos\theta\,, \\ &\Gamma^2{}_{11} = 0\,, \qquad \Gamma^2{}_{12} = \cot\theta\,, \qquad \Gamma^2{}_{22} = 0\,. \end{aligned} \tag{6.171}$$

From the symmetries of the Riemann tensor given above, it follows that in two dimensions there is only one independnet component, and one easily finds that this is given by

$$R_{1212} = \sin^2\theta\,. \tag{6.172}$$

The Ricci tensor R_{ij} and Ricci scalar R then turn out to be

$$R_{11} = 1\,, \qquad R_{22} = \sin^2\theta\,, \qquad R_{12} = R_{21} = 0\,, \qquad R = 2\,. \tag{6.173}$$

Notice that by comparing with (6.170), we see that the Ricci tensor can be written as

$$R_{ij} = g_{ij}\,. \tag{6.174}$$

Metrics whose Ricci tensors satisfy this type of equation, $R_{ij} = \Lambda\, g_{ij}$, are called *Einstein Metrics*, and they are of great importance in mathematics and in theoretical physics.

We conclude this section with some remarks about one of the most important physical applications of the geometrical theory of tensors that we have been studying, namely Einstein's theory of *General Relativity.* This is the theory that describes the phenomenon of gravity, superseding the Newtonian theory of gravity. One of the cornerstones of general relativity is the fact that the "force of gravity" is a frame-dependent concept, being indistinguishable (by means of local experiments) from the effects of acceleration. Thus one can, for example, always render the force of gravity vanishing at some point, by putting oneself in a freely-falling frame (not necessarily a wise thing to do!). Conversely, one can produce a gravitational force that is locally indistinguishable from the force of gravity on the surface of the earth, even out in the far reaches of space, by turning on the rocket-motor of a spacecraft so that is accelerates at 32 feet per second per second.[30]

In general relativity the four-dimensional Minkowsi spacetime metric of special relativity is replaced by a more general four-dimensional spacetime metric. As in our previous discussions, in some cases this might be just a rewriting of the Minkowski metric after some change of coordinates. On the other hand, it might be a genuinly curved metric. It should perhaps come as no surprise, in the light of previous remarks, that the "force of gravity" is characterised by the Christoffel connection ${\Gamma^i}_{jk}$. The frame-dependence of the concept of the gravitational force is now understandable, since it is described by the non-tensorial quantities ${\Gamma^i}_{jk}$. For instance, in a small local region any space looks nearly like a patch of flat space (think of a small region on the surface of the earth, for example), and this means that one can find a coordinate transformation in which the metric becomes like the Minkowski metric at a point, and its first derivatives vanish at that same point. This implies that in this coordinate system the Christoffel connection vanishes at that point, and then there is no "force of gravity." The coordinate system that one has picked that does this job is the "local inertial frame" or "free-fall frame."

The precise way in which the Christoffel connection describes the "force of gravity" is as follows. Consider the worldline of a particle that is acted on by no forces other than gravity. Assuming the particle is massive, we can use the elapse of proper time τ, as measured in the rest frame of the particle, to parameterise its path in spacetime, $x^i = x^i(\tau)$. The equation that governs its motion, called the *Geodesic Equation*, is then

$$\frac{d^2 x^i}{d\tau^2} + {\Gamma^i}_{jk}\, \frac{dx^j}{d\tau}\, \frac{dx^k}{d\tau} = 0\,. \tag{6.175}$$

This equation is the analogue in general relativity of Newton's second law of motion, applied to a massive particle in a gravitational field. In the Newtonian limit of weak gravitational fields and low velocities, the first term in the geodesic equation becomes the acceleration

[30]These evident facts, which are such important foundations in General Relativity, are still, curiously, often denied by the "old guard" of adherents to the Newtonian school of thought. Thus one still frequently encounters, especially in undergraduate mechanics classes, the counter-Einsteinian assertion that "centrifugal forces are fictitious." The trouble stems from an uneasiness, in the old Newtonian picture, with the modern concept that all coordinate frames should be equally valid. Thus "inertial frames" were singled out as the only ones that were kosher, and so forces resulting from acceleration relative to these were deemed to be fictitious. It is interesting to note that the Newtonian and the Einsteinian physicist will disagree on what constitutes an inertial frame. A Newtonian physicist will say that an observer standing in a laboratory on the earth is in an inertial frame, whereas the Einsteinian physicist will say that an observer who is in free-fall, having jumped out of the laboratory window, is in a (local) inertial frame.

of the particle, while in the second term the components $\Gamma^a{}_{00}$ of the Christoffel connection become the dominant ones, where 0 represents the time direction, and the a index ranges over the three spatial directions. In fact in the Newtonian limit, in Cartesian coordinates, these components are given by $\Gamma^a{}_{00} = \partial_a \Phi$, where Φ is the Newtonian gravitational potential. Furthermore, at low velocities we have $dx^0/d\tau \sim 1$, $|dx^a/d\tau| << 1$ (we use units where the speed of light is $c = 1$), and so the geodesic equation limits to

$$\frac{d^2 x^a}{dt^2} = -\frac{\partial \Phi}{\partial x^a}\,, \tag{6.176}$$

which is Newton's second law for the motion of a particle in a gravitational field. Even in the Newtonian limit, however, we see the radically different interpretations of the Newtonian and the Einsteinian viewpoints. The Newtonian physicist will only interpret the right-hand side of (6.176) as a gravitational force if he has first checked to see that the coordinate system is one that is deemed to "inertial" in the Newtonian sense. By contrast, the general relativist places all coordinate systems on a democratic footing, and universally interprets (6.175) as the equation describing the motion of the particle in the gravitational field, without any preference for one coordinate system over another.

Although we can make gravity vanish "at a point," we cannot in general make it vanish everywhere by choice of coordinate frame, except in the special case of a flat spacetime. This is like the difference between the flat 2-plane and the 2-sphere; locally, they both look like bits of flat space, but larger excursions reveal that the plane is flat, while the sphere is curved. In general relativity the curvature of spacetime is brought about by the presence of matter, or other disturbances (such as gravitational waves). The precise way in which this happens is described by the Einstein field equations, which read

$$R_{ij} - \tfrac{1}{2} R\, g_{ij} = 8\pi\, G\, T_{ij}\,. \tag{6.177}$$

The quantities on the left-hand side are the Ricci tensor R_{ij} and Ricci scalar R of the spacetime metric g_{ij}. On the right-hand side T_{ij} is the energy-momentum tensor of the matter, which describes the distribution of energy, and momentum, in the spacetime. Finally, G is Newton's constant.[31] These field equations are the gravitational analogue of the Maxwell field equations

$$\partial_\mu F^{\mu\nu} = -4\pi\, J^\nu\,, \tag{6.178}$$

(or $\vec\nabla \cdot \vec E = 4\pi\,\rho$, $\vec\nabla \times \vec B - \partial \vec E/\partial t = 4\pi\,\vec J$ if you prefer). Just as the Maxwell field equations describe how the distribution of charges and currents generates electromagnetic fields, so

[31] So there is still a place for Newton in the New Order!

the Einstein field equations describe how the distribution of masses and momentum flux generate curvature. Unlike electrodynamics, however, the general theory of relativity is a *non-linear* theory, which makes it considerably more complicated and subtle. Between them, the geodesic equation (6.175) which tells matter how to respond to the geometry, and the Einstein equation (6.177) which tells geometry how to respond to the matter, constitute one of the most elegant and intriguing of our fundamental physical laws.