PHY 387K

Advanced Classical Electromagnetism

a graduate level course of lectures given by

Richard Fitzpatrick

ASSISTANT PROFESSOR OF PHYSICS

The University of Texas at Austin

Fall 1996

`Email: rfitzp@farside.ph.utexas.edu, Tel.: 512-471-9439`

1 Introduction

1.1 Major sources

The textbooks which I have consulted most frequently whilst developing course material are:

Classical electrodynamics: J.D. Jackson, 2nd edition (John Wiley & Sons, New York NY, 1975).

Classical electricity and magnetism: W.K. Panofsky and M. Phillips, 2nd edition, (Addison-Wesley, Reading MA, 1962).

Special relativity: W. Rindler, 2nd edition (Oliver and Boyd, Edinburg and London, 1966).

Foundations of electromagnetic theory: J.R. Reitz and F.J. Milford, 2nd edition (Addison-Wesley, Reading MA, 1967).

Lectures on theoretical physics: A. Sommerfeld, (Academic Press, New York, 1954).

Wave propagation and group velocity: Léon Brillouin, (Academic Press, New York NY, 1960).

Methods of theoretical physics: P.M. Morse and H. Feshbach, (McGraw-Hill, New York NY, 1953).

An introduction to phase-integral methods: J. Heading, (Meuthen & Co., London, 1962).

Radio waves in the ionosphere: K.G. Budden, (Cambridge University Press, Cambridge, 1961).

Classical electromagnetic radiation: M.A. Heald and J.B. Marion, 3rd edition (Saunders College Publishing, Fort Worth TX, 1995).

1.2 Outline of course

You have all presumably taken the standard undergraduate electromagnetism course in which Maxwell's equations are derived and explained. The basic aim of my course is to cover some material which is usually inadequately treated or omitted altogether in undergraduate courses. In fact, I intend to concentrate on three main topics:

1. The relativistically invariant formulation of the laws of electromagnetism.
2. The effect of dielectric and magnetic materials on electric and magnetic fields.
3. The generation, propagation, and scattering of electromagnetic waves.

1.3 The validity of classical electromagnetism

In this course we shall investigate the classical theory of electromagnetism in Euclidian space-time. This theory is valid over a huge range of different conditions, but, nevertheless, breaks down under certain circumstances. On very large length-scales (or close to collapsed objects such as black holes) the theory must be modified to take general relativistic effects into account. On the other hand, the theory breaks down completely on very small length-scales because of quantum effects. It is legitimate to treat a gas of photons as a classical electromagnetic field provided that we only attempt to resolve space-time into elements that contain a great many photons. In conventional applications of electromagnetic theory (*e.g.,* the generation and propagation of radio waves) this is not a particularly onerous constraint.

1.4 Units

In 1960 physicists throughout the world adopted the so-called S.I. system of units, whose standard measures of length, mass, time, and electric charge are the meter, kilogram, second, and coloumb, respectively. Nowadays, the S.I. system is used almost exclusively in most areas of physics. In fact, only one area of physics has proved at all resistant to the adoption of S.I. units, and that, unfortunately, is electromagnetism, where the previous system of units, the so-called Gaussian system, simply refuses to die out. Admittedly, this is mostly an Anglo-Saxon phenomenon; the Gaussian system is most prevalent in the U.S., followed by Britain (although, the Gaussian system is rapidly dying out in Britain under the benign influence of the European Community). One major exception to this rule is astrophysics, where the Gaussian system remains widely used throughout the world. Incidentally, the standard units of length, mass, time, and electric charge in the Gaussian system are the centimeter, gram, second, and statcoloumb, respectively.

You might wonder why anybody would wish to adopt a different set units in electromagnetism to that used in most other branches of physics. The answer is that in the Gaussian system the laws of electromagnetism look a lot "prettier" than in the S.I. system. There are no ϵ_0 s and μ_0 s in any of the formulae. In fact, in the Gaussian system the only normalizing constant appearing in Maxwell's equations is c, the velocity of light. However, there is a severe price to pay for the aesthetic advantages of the Gaussian system. The standard measures of potential difference and electric current in the S.I. system are the volt and the ampere, respectively. I presume that you all have a fairly good idea how large a voltage 1 volt is, and how large a current 1 ampere is. The standard measures of potential difference and electric current in the Gaussian system are the statvolt and the statampere, respectively. I wonder how many of you have even the slightest idea how large a voltage 1 statvolt is, or how large a current 1 statampere is? Nobody, I bet! Let me tell you: 1 statvolt is 300 volts, and 1 statampere is $1/3 \times 10^{-9}$ amperes. Clearly, these are not particularly convenient units!

In order to decide which system of units we should employ in this course, we essentially have to answer a single question. What is more important to us: that our equations should look pretty, or that the our fundamental units should be sensible? I think that sensible units are of vital importance, especially if we are going to make quantitative calculations (we are!), whereas the prettiness or otherwise of our equations is of marginal concern. For this reason, I intend to use the S.I. system throughout this course.

If, unaccountably, you prefer the Gaussian system of units, there is no reason to despair. Converting formulae from the S.I. system to the Gaussian system is trivial: just use the following transformation

$$\epsilon_0 \rightarrow \frac{1}{4\pi}, \tag{1.1a}$$

$$\mu_0 \rightarrow \frac{4\pi}{c^2}, \tag{1.1b}$$

$$B \rightarrow \frac{B}{c}. \tag{1.1c}$$

The transformation (1.1c) also applies to quantities which are directly related to magnetic field strength, such as the vector potential. Unfortunately, converting formulae from the Gaussian system to the S.I. system is far less straightforward.

As an example of this, consider Coulomb's law in S.I. units:

$$\boldsymbol{f}_2 = \frac{q_1 q_2}{4\pi\epsilon_0}\frac{\boldsymbol{r}_2 - \boldsymbol{r}_1}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|^3}. \tag{1.2}$$

Employing the above transformation, this formula converts to

$$\boldsymbol{f}_2 = q_1 q_2 \, \frac{\boldsymbol{r}_2 - \boldsymbol{r}_1}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|^3} \tag{1.3}$$

in Gaussian units. However, applying the inverse transformation is problematic. In Eq. (1.3) the geometric 4π in the S.I. formula has canceled with the $1/4\pi$ obtained from transforming ϵ_0 to give unity. It is not at all obvious that the reverse transformation should generate a factor $4\pi\epsilon_0$ in the denominator. In fact, the only foolproof way of transforming Eq. (1.3) back into Eq. (1.2) is to use dimensional analysis. This is another good reason for not using the Gaussian system.

There are four fundamental quantities in electrodynamics; mass, length, time, and charge, denoted M, L, T, and Q, respectively. Each of these quantities has its own particular units, since mass, length, time, and charge are fundamentally different from one another. The units of a general physical quantity, such as force or capacitance, can always be expressed as some appropriate power law combination of the four fundamental units, M, L, T, and Q. Equation (1.2) makes dimensional sense because the constant ϵ_0 possesses the units $M^{-1}L^{-3}T^2Q^2$. Likewise, the Biot-Savart law only makes dimensional sense because the constant μ_0 possesses the units MLQ^{-2}. On the other hand, Eq. (1.3) does not make much dimensional sense; *i.e.*, the right-hand side and the left-hand side appear to possess different units. In fact, we can only reconcile Eqs. (1.2) and (1.3) if we divide the right-hand side of (1.3) by some constant, $4\pi\epsilon_0$, say, with dimensions $M^{-1}L^{-3}T^2Q^2$, which happens to have the numerical value unity for the particular choice of units in the Gaussian scheme. Likewise, the Gaussian version of the Biot-Savart law contains a hidden constant with the numerical value unity which also possesses dimensions. It can be seen that the apparent simplicity of the equations of electrodynamics in the Gaussian scheme is only achieved at the expense of wrecking their dimensionality. This is, perhaps, the best reason of all for not using Gaussian units.

2 Relativity and electromagnetism

2.1 The relativity principle

Physical phenomena are conventionally described relative to some *frame of reference* which allows us to define fundamental quantities such as position and time. Of course, there are very many different ways of choosing a reference frame, but it generally convenient to restrict our choice to the set of rigid inertial frames. A classical rigid reference frame is the imagined extension of a rigid body. For instance, the Earth determines a rigid frame throughout all space, consisting of all those points which remain rigidly at rest relative to the Earth and each other. We can associate an orthogonal Cartesian coordinate system S with such a frame, by choosing three mutually orthogonal planes within it and measuring x, y, and z as distances from these planes. A time coordinate must also be defined in order that the system can be used to specify events. A rigid frame, endowed with such properties, is called a *Cartesian frame*. The description given above presupposes that the underlying geometry of space is Euclidian, which is reasonable provided that gravitational effects are negligible (we shall assume that this is the case). An *inertial* frame is a Cartesian frame in which free particles move without acceleration, in accordance with Newton's first law of motion. There are an infinite number of different inertial frames, each moving with some constant velocity with respect to a given inertial frame.

The key to understanding special relativity is Einstein's *relativity principle*, which states that

> All inertial frames are totally equivalent for the performance of all physical experiments.

In other words, it is impossible to perform a physical experiment which differentiates in any fundamental sense between different inertial frames. By definition, Newton's laws of motion take the same form in all inertial frames. Einstein generalized this result in his special theory of relativity by asserting that *all* laws of physics take the same form in all inertial frames.

Consider a wave-like disturbance. In general, such a disturbance propagates at a fixed velocity with respect to the medium in which the disturbance takes place. For instance, sound waves (at S.T.P.) propagate at 343 meters per second with respect to air. So, in the inertial frame in which air is stationary sound waves appear to propagate at 343 meters per second. Sound waves appear to propagate at a different velocity in some other inertial frame which is moving with respect to the first frame. However, this does not violate the relativity principle, since if the air were stationary in the second frame then sound waves would appear to propagate at 343 meters per second in this frame as well. In other words, exactly the same experiment (*e.g.*, the determination of the speed of sound relative to stationary air) performed in two different inertial frames of reference yields exactly the same result, in accordance with the relativity principle.

Consider, now, a wave-like disturbance which is self-regenerating and does not require a medium through which to propagate. The most well known example of such a disturbance is a light wave. Another example is a gravity wave. According to electromagnetic theory the speed of propagation of a light wave through a vacuum is

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} = 2.99729 \times 10^8 \text{ meters per second,} \tag{2.1}$$

where ϵ_0 and μ_0 are physical constants which can be evaluated by performing two simple experiments which involve measuring the force of attraction between two fixed changes and two fixed parallel current carrying wires. According to the relativity principle these experiments must yield the same values for ϵ_0 and μ_0 in all inertial frames. Thus, the speed of light must be the same in all inertial frames. In fact, any disturbance which does not require a medium to propagate through must appear to travel at the same velocity in all inertial frames, otherwise we could differentiate inertial frames using the apparent propagation speed of the disturbance, which would violate the relativity principle.

2.2 The Lorentz transform

Consider two Cartesian frames $S(x, y, z, t)$ and $S'(x', y', z', t')$ in the *standard configuration* in which S' moves in the x-direction of S with uniform velocity v and the corresponding axes of S and S' remain parallel throughout the motion, having coincided at $t = t' = 0$. It is assumed that the same units of distance and time are adopted in both frames. Suppose that an *event* (*e.g.,* the flashing of a light-bulb, or the collision of two point particles) has coordinates (x, y, z, t) relative to S and (x', y', z', t') relative to S'. The "common sense" relationship between these two sets of coordinates is given by the Galilean transformation:

$$\begin{aligned} x' &= x - vt, && \text{(2.2a)} \\ y' &= y, && \text{(2.2b)} \\ z' &= z, && \text{(2.2c)} \\ t' &= t. && \text{(2.2d)} \end{aligned}$$

This transformation is tried and tested and provides a very accurate description of our everyday experience. Nevertheless, it must be wrong! Consider a light wave which propagates along the x-axis in S with velocity c. According to the Galilean transformation the apparent speed of propagation in S' is $c - v$, which violates the relativity principle. Can we construct a new transformation which makes the velocity of light invariant between different inertial frames, in accordance with the relativity principle, but reduces to the Galilean transformation at low velocities, in accordance with our everyday experience?

Consider an event P and a neighbouring event Q whose coordinates differ from those of P by dx, dy, dz, dt in S and by dx', dy', dz', dt' in S'. Suppose that at the event P a flash of light is emitted and that Q is an event in which some particle in space is illuminated by the flash. In accordance with the laws of light-propagation, and the invariance of the velocity of light between different inertial frames, an observer in S will find that

$$dx^2 + dy^2 + dz^2 - c^2 dt^2 = 0 \tag{2.3}$$

for $dt > 0$, and an observer in S' will find that

$$dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2 = 0 \tag{2.4}$$

for $dt' > 0$. Any event near P whose coordinates satisfy *either* (2.3) *or* (2.4) is illuminated by the flash from P and therefore its coordinates must satisfy *both*

(2.3) *and* (2.4). Now, no matter what form the transformation between coordinates in the two inertial frames takes, the transformation between differentials at any fixed event P is linear and homogeneous. In other words, if

$$x' = F(x, y, z, t), \tag{2.5}$$

where F is a general function, then

$$dx' = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy + \frac{\partial F}{\partial z}\,dz + \frac{\partial F}{\partial t}\,dt. \tag{2.6}$$

It follows that

$$\begin{aligned} dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2 &= a\,dx^2 + b\,dy^2 + c\,dz^2 + d\,dt^2 + g\,dx\,dt + h\,dy\,dt \\ &\quad + k\,dz\,dt + l\,dy\,dz + m\,dx\,dz + n\,dx\,dy, \end{aligned} \tag{2.7}$$

where a, b, c, *etc.* are functions of x, y, z, and t. We know that the right-hand side of the above expression vanishes for all real values of the differentials which satisfy Eq. (2.3). It follows that the right-hand side is a multiple of the quadratic in Eq. (2.3); *i.e.*,

$$dx^2 + dy^2 + dz^2 - c^2 dt^2 = K(dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2), \tag{2.8}$$

where K is a function of x, y, z, and t. [We can prove this by substituting into Eq. (2.7) the following obvious zeros of the quadratic in Eq. (2.3): $(\pm 1, 0, 0, 1)$, $(0, \pm 1, 0, 1)$, $(0, 0, \pm 1, 1)$, $(0, 1/\sqrt{2}, 1/\sqrt{2}, 1)$, $(1/\sqrt{2}, 0, 1/\sqrt{2}, 1)$, $(1/\sqrt{2}, 1/\sqrt{2}, 0, 1)$: and solving the resulting conditions on the coefficients.] Note that K at P is also independent of the choice of standard coordinates in S and S'. Since the frames are Euclidian, the values of $dx^2 + dy^2 + dz^2$ and $dx'^2 + dy'^2 + dz'^2$ relevant to P and Q are independent of the choice of axes. Furthermore, the values of dt^2 and dt'^2 are independent of the choice of the origins of time. Thus, without affecting the value of K at P we can choose coordinates such that $P = (0, 0, 0, 0)$ in both S and S'. Since the orientations of the axes in S and S' are, at present, arbitrary, and since inertial frames are isotropic, the relation of S and S' relative to each other, to the event P, and to the locus of possible events Q is now completely symmetric. Thus, we can write

$$dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2 = K(dx^2 + dy^2 + dz^2 - c^2 dt^2), \tag{2.9}$$

in addition to Eq. (2.8). It follows that $K = \pm 1$. $K = -1$ can be dismissed immediately, since the intervals $dx^2+dy^2+dz^2-c^2dt^2$ and $dx'^2+dy'^2+dz'^2-c^2dt'^2$ must coincide exactly when there is no motion of S' relative to S. Thus,

$$dx'^2 + dy'^2 + dz'^2 - c^2dt'^2 = dx^2 + dy^2 + dz^2 - c^2dt^2. \tag{2.10}$$

Equation (2.10) implies that the transformation equations between primed and unprimed coordinates must be *linear*. The proof of this statement is postponed until later.

The linearity of the transformation allows the coordinate axes in the two frames to be orientated so as to give the *standard configuration* mentioned earlier. Consider a fixed plane in S with the equation $lx + my + nz + p = 0$. In S' this becomes, say, $l(a_1x' + b_1y' + c_1z' + d_1t' + e_1) + m(a_2x' + \cdots) + n(a_3x' + \cdots) + p = 0$, which represents a moving plane unless $ld_1 + md_2 + nd_3 = 0$. That is, unless the normal vector to the plane (l, m, n) in S is perpendicular to the vector (d_1, d_2, d_3). All such planes intersect in lines which are fixed in both S and S', and which are parallel to the vector (d_1, d_2, d_3) in S. These lines must correspond to the direction of relative motion of the frames. By symmetry, two such frames which are orthogonal in S must also be orthogonal in S'. This allows the choice of two common coordinate planes.

Under a linear transformation the finite coordinate differences satisfy the same transformation equations as the differentials. It follows from Eq. (2.10), assuming that the events $(0, 0, 0, 0)$ coincide in both frames, that for any event with coordinates (x, y, z, t) in S and (x', y', z', t') in S' the following relation holds:

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2. \tag{2.11}$$

By hypothesis, the coordinate planes $y = 0$ and $y' = 0$ coincide permanently. Thus, $y = 0$ must imply $y' = 0$, which suggests that

$$y' = Ay, \tag{2.12}$$

where A is a constant. We can reverse the directions of the x- and z-axes in S and S', which has the effect of interchanging the roles of these frames. This procedure does not affect Eq. (2.12), but by symmetry we also have

$$y = Ay'. \tag{2.13}$$

It is clear that $A = \pm 1$. The negative sign can again be dismissed, since $y = y'$ when there is no motion between S and S'. The argument for z is similar. Thus, we have

$$y' = y, \tag{2.14a}$$

$$z' = z, \tag{2.14b}$$

as in the Galilean transformation.

Equations (2.11) and (2.14) yield

$$x^2 - c^2 t^2 = x'^2 - c^2 t'^2. \tag{2.15}$$

Since, $x' = 0$ must imply $x = vt$, we can write

$$x' = B(x - vt), \tag{2.16}$$

where B is a constant (possibly depending on v). It follows from the previous two equations that

$$t' = Cx + Dt, \tag{2.17}$$

where C and D are constants (possibly depending on v). Substituting Eqs. (2.16) and (2.17) into Eq. (2.15) and comparing the coefficients of x^2, xt, and t^2, we obtain

$$B = D = \frac{1}{\pm(1 - v^2/c^2)^{1/2}}, \tag{2.18a}$$

$$C = \frac{-v/c^2}{\pm(1 - v^2/c^2)^{1/2}}. \tag{2.18b}$$

We must choose the positive sign in order to ensure that $x' \to x$ as $v/c \to 0$. Thus, collecting our results, the transformation between coordinates in S and S' is given by

$$x' = \frac{x - vt}{(1 - v^2/c^2)^{1/2}}, \tag{2.19a}$$

$$y' = y, \tag{2.19b}$$

$$z' = z, \tag{2.19c}$$

$$t' = \frac{t - vx/c^2}{(1 - v^2/c^2)^{1/2}}, \tag{2.19d}$$

This is the famous *Lorentz transform.* It ensures that the velocity of light is invariant between different inertial frames, and also reduces to the more familiar Galilean transform in the limit $v/c \ll 1$. We can solve Eqs. (2.19) for x, y, z, and t to obtain the *inverse Lorentz transform*:

$$x = \frac{x' + v t'}{(1 - v^2/c^2)^{1/2}}, \tag{2.20a}$$

$$y = y', \tag{2.20b}$$

$$z = z', \tag{2.20c}$$

$$t = \frac{t' + v x'/c^2}{(1 - v^2/c^2)^{1/2}}. \tag{2.20d}$$

Clearly, the inverse transform is equivalent to a Lorentz transform in which the velocity of the moving frame is $-v$ along the x-axis instead of $+v$.

2.3 Transformation of velocities

Consider two frames S and S' in the standard configuration. Let $\boldsymbol{u}$ be the velocity of a particle in S. What is the particle velocity in S'? The components of the velocity are

$$u_1 = \frac{dx}{dt}, \tag{2.21a}$$

$$u_2 = \frac{dy}{dt}, \tag{2.21b}$$

$$u_3 = \frac{dz}{dt}, \tag{2.21c}$$

and, similarly, the components of $\boldsymbol{u}'$ are

$$u_1' = \frac{dx'}{dt'}, \tag{2.22a}$$

$$u_2' = \frac{dy'}{dt'}, \tag{2.22b}$$

$$u_3' = \frac{dz'}{dt'}. \qquad (2.22c)$$

Now we can write Eqs. (2.19) in the form $dx' = \gamma(dx - vdt)$, $dy' = dy$, $dz' = dz$, and $dt' = \gamma(dt - vdx/c^2)$, where

$$\gamma = \frac{1}{(1 - v^2/c^2)^{1/2}} \qquad (2.23)$$

is the well known *Lorentz factor*. If we substitute these differentials into Eqs. (2.22) and make use of Eqs. (2.21), we obtain the transformation formulae

$$u_1' = \frac{u_1 - v}{1 - u_1 v/c^2}, \qquad (2.24a)$$

$$u_2' = \frac{u_2}{\gamma(1 - u_1 v/c^2)}, \qquad (2.24b)$$

$$u_3' = \frac{u_3}{\gamma(1 - u_1 v/c^2)}. \qquad (2.24c)$$

As in the transformation of coordinates, we can obtain the inverse transform by interchanging primed and unprimed symbols and replacing $+v$ with $-v$. Thus,

$$u_1 = \frac{u_1' + v}{1 + u_1' v/c^2}, \qquad (2.25a)$$

$$u_2 = \frac{u_2'}{\gamma(1 + u_1' v/c^2)}, \qquad (2.25b)$$

$$u_3 = \frac{u_3'}{\gamma(1 + u_1' v/c^2)}. \qquad (2.25c)$$

Equations (2.25) can be regarded as giving the resultant, $\boldsymbol{u} = (u_1, u_2, u_3)$, of two velocities, $\boldsymbol{v} = (v, 0, 0)$ and $\boldsymbol{u}' = (u_1', u_2', u_3')$, and are therefore usually referred to as the relativistic *velocity addition formulae*. The following relation between the magnitudes $u = ({u_1}^2 + {u_2}^2 + {u_3}^2)^{1/2}$ and $u' = ({u_1'}^2 + {u_2'}^2 + {u_3'}^2)^{1/2}$ of the velocities is easily demonstrated:

$$c^2 - u^2 = \frac{c^2(c^2 - u'^2)(c^2 - v^2)}{(c^2 + u_1' v)^2}. \qquad (2.26)$$

If $u' < c$ and $v < c$ the right-hand side is positive, implying that $u < c$. In other words, the resultant of two subluminal velocities is another subluminal velocity. It is evident that a particle can never attain the velocity of light relative to a given inertial frame, no matter how many subluminal velocity increments it is given. It follows that no inertial frame can appear to propagate with a superluminal velocity with respect to any other inertial frame (since we can track the origin of a given inertial frame using a particle which remains at rest at the origin in that frame).

According to Eq. (2.26), if $u' = c$ then $u = c$ no matter what value v takes; *i.e.,* the velocity of light is invariant between different inertial frames. Note that the Lorentz transform only allows *one* such invariant velocity (*i.e.,* the velocity c which appears in Eqs. (2.19)). Einstein's relativity principle tells us that any disturbance which propagates through a vacuum must appear to propagate at the same velocity in all inertial frames. It is now evident that *all* such disturbances must propagate at the velocity c. It follows immediately that all electromagnetic waves must propagate through the vacuum with this velocity, irrespective of their wavelength. In other words, it is impossible for there to be any dispersion of electromagnetic waves propagating through a vacuum. Furthermore, gravity waves must also propagate with the velocity c. It is convenient to label c as "the velocity of light" since electromagnetic radiation is, by far, the most well known and easily measurable type of disturbance which can propagate through a vacuum.

The Lorentz transformation implies that not only the velocities of material particles but the velocities of propagation of all physical effects are limited by c in deterministic physics. Consider a general process by which an event P *causes* an event Q at a velocity $U > c$ in some frame S. In other words, *information* about the event P appears to propagate to the event Q with a superluminal velocity. Let us choose coordinates such that these two events occur on the x-axis with (finite) time and distance separations $\Delta t > 0$ and $\Delta x > 0$, respectively. The time separation in some other inertial frame S' is given by (see Eq. (2.19d))

$$\Delta t' = \gamma(\Delta t - v\Delta x/c^2) = \gamma\Delta t(1 - vU/c^2). \tag{2.27}$$

Thus, for sufficiently large $v < c$ we obtain $\Delta t' < 0$; *i.e.,* there exist inertial frames in which cause and effect appear to be reversed. Of course, this is impossible in deterministic physics. It follows, therefore, that information can never appear to propagate with a superluminal velocity in any inertial frame, otherwise causality would be violated.

2.4 Tensors

It is now convenient to briefly review the mathematics of tensors. Tensors are of primary importance in connection with coordinate transforms. They serve to isolate intrinsic geometric and physical properties from those that merely depend on coordinates.

A tensor of rank r in an n-dimensional space possesses n^r components which are, in general, functions of position in that space. A tensor of rank zero has one component A and is called a *scalar*. A tensor of rank one has n components $(A_1, A_2, \cdots, A_n)$ and is called a *vector*. A tensor of rank two has n^2 components, which can be exhibited in matrix format. Unfortunately, there is no convenient way of exhibiting a higher rank tensor. Consequently, tensors are usually represented by a typical component; *e.g.,* we talk of the tensor A_{ijk} (rank 3) or the tensor A_{ijkl} (rank 4), *etc.* The suffixes $i, j, k, \cdots$ are always understood to range from 1 to n.

For reasons which will become apparent later on, we shall represent tensor components using both superscripts and subscripts. Thus, a typical tensor might look like A^{ij} (rank 2), or B^i_j (rank 2), *etc.* It is convenient to adopt the Einstein summation convention. Namely, if any suffix appears twice in a given term, once as a subscript and once as a superscript, a summation over that suffix (from 1 to n) is implied.

To distinguish between various coordinate systems we shall use primed and multiply primed suffixes. A first system of coordinates $(x^1, x^2, \cdots, x^n)$ can then be denoted by x^i, a second system $(x^{1'}, x^{2'}, \cdots, x^{n'})$ by $x^{i'}$, *etc.* Similarly the general components of a tensor in various coordinate systems are distinguished by their suffixes. Thus, the components of some third rank tensor are denoted A_{ijk} in the x^i system, by $A_{i'j'k'}$ in the $x^{i'}$ system, *etc.*

When making a coordinate transformation from one set of coordinates x^i to another $x^{i'}$, it is assumed that the transformation in non-singular. In other words, the equations which express the $x^{i'}$ in terms of the x^i can be inverted to express the x^i in terms of the $x^{i'}$. It is also assumed that the functions specifying a transformation are differentiable. It is convenient to write

$$\frac{\partial x^{i'}}{\partial x^i} = p_i^{i'}, \tag{2.28a}$$

$$\frac{\partial x^i}{\partial x^{i'}} = p_{i'}^i. \tag{2.28b}$$

Note that

$$p_{i'}^i p_{i''}^{i'} = p_{i''}^i, \tag{2.29a}$$

$$p_{i'}^i p_j^{i'} = \delta_j^i \tag{2.29b}$$

by the chain rule, where δ_j^i (the *Kronecker delta*) equals 1 or 0 when $i = j$ or $i \neq j$, respectively.

The formal definition of a tensor is as follows:
(i) An entity having components $A_{ij\cdots k}$ in the x^i system and $A_{i'j'\cdots k'}$ in the $x^{i'}$ system is said to behave as a *covariant tensor* under the transformation $x^i \to x^{i'}$ if

$$A_{i'j'\cdots k'} = A_{ij\cdots k}\, p_{i'}^i p_{j'}^j \cdots p_{k'}^k. \tag{2.30}$$

(ii) Similarly, $A^{ij\cdots k}$ is said to behave as a *contravariant tensor* under $x^i \to x^{i'}$ if

$$A^{i'j'\cdots k'} = A^{ij\cdots k} p_i^{i'} p_j^{j'} \cdots p_k^{k'}. \tag{2.31}$$

(iii) Finally, $A_{k\cdots l}^{i\cdots j}$ is said to behave as a *mixed tensor* (contravariant in $i \cdots j$ and covariant in $k \cdots l$) under $x^i \to x^{i'}$ if

$$A_{k'\cdots l'}^{i'\cdots j'} = A_{k\cdots l}^{i\cdots j}\, p_i^{i'} \cdots p_j^{j'}\, p_{k'}^k \cdots p_{l'}^l. \tag{2.32}$$

When an entity is described as a tensor it is generally understood that it behaves as a tensor under *all* non-singular differentiable transformations of the relevant coordinates. An entity which only behaves as a tensor under a certain subgroup of non-singular differentiable coordinate transformations is called a *qualified tensor*, because its name is conventionally qualified by an adjective recalling the subgroup in question. For instance, an entity which only exhibits tensor behaviour under Lorentz transformations is called a Lorentz tensor or, more commonly, a 4-tensor.

When applied to a tensor of rank zero (a scalar), the above definitions imply that $A^{i} = A$. Thus, a scalar is a function of position only, and is independent of the coordinate system. A scalar is often termed an *invariant*.

The main theorem of tensor calculus is as follows:

> If two tensors of the same type are equal in one coordinate system, then they are equal in all coordinate systems.

The simplest example of a contravariant vector (tensor of rank one) is provided by the differentials of the coordinates, dx^{i}, since

$$dx^{i'} = \frac{\partial x^{i'}}{\partial x^{i}} dx^{i} = dx^{i} p^{i'}_{i}. \tag{2.33}$$

The coordinates themselves do not behave as tensors under all coordinate transformations. However, since they transform like their differentials under linear homogeneous coordinate transformations, they do behave as tensors under such transformations.

The simplest example of a covariant vector is provided by the gradient of a function of position $\phi = \phi(x^{1}, \cdots, x^{n})$. Since, if we write

$$\phi_{i} = \frac{\partial \phi}{\partial x^{i}}, \tag{2.34}$$

then we have

$$\phi_{i'} = \frac{\partial \phi}{\partial x^{i'}} = \frac{\partial \phi}{\partial x^{i}} \frac{\partial x^{i}}{\partial x^{i'}} = \phi_{i}\, p^{i}_{i'}. \tag{2.35}$$

An important example of a mixed second rank tensor is provided by the

Kronecker delta introduced previously. Since,

$$\delta^i_j\, p^{i'}_i\, p^j_{j'} = p^{i'}_j\, p^j_{j'} = \delta^{i'}_{j'}. \tag{2.36}$$

Tensors *of the same type* can be added or subtracted to form new tensors. Thus, if A_{ij} and B_{ij} are tensors, then $C_{ij} = A_{ij} \pm B_{ij}$ is a tensor of the same type. Note that the sum of tensors at different points in space is not a tensor if the p's are position dependent. However, under linear coordinate transformations the p's are constant, so the sum of tensors at different points behaves as a tensor under this particular type of coordinate transformation.

If A^{ij} and B_{ijk} are tensors, then $C^{ij}_{klm} = A^{ij} B_{klm}$ is a tensor of the type indicated by the suffixes. The process illustrated by this example is called *outer multiplication* of tensors.

Tensors can also be combined by *inner multiplication*, which implies at least one dummy suffix link. Thus, $C^j_{kl} = A^{ij} B_{ikl}$ and $C_k = A^{ij} B_{ijk}$ are tensors of the type indicated by the suffixes.

Finally, tensors can be formed by *contraction* from tensors of higher rank. Thus, if A^{ij}_{klm} is a tensor then $C^j_{kl} = A^{ij}_{ikl}$ and $C_k = A^{ij}_{kij}$ are tensors of the type indicated by the suffixes. The most important type of contraction occurs when no free suffixes remain: the result is a scalar. Thus, A^i_i is a scalar provided that A^j_i is a tensor.

Although we cannot usefully divide tensors, one by another, an entity like C^{ij} in the equation $A^j = C^{ij} B_i$, where A^i and B_i are tensors, can be formally regarded as the quotient of A^i and B_i. This gives the name to a particularly useful rule for recognizing tensors, the *quotient rule*. This rule states that *if a set of components, when combined by a given type of multiplication with all tensors of a given type yields a tensor, then the set is itself a tensor.* In other words, if the product $A^i = C^{ij} B_j$ transforms like a tensor for *all* tensors B_i then it follows that C^{ij} is a tensor.

Let

$$\frac{\partial A^{i\cdots j}_{k\cdots l}}{\partial x^m} = A^{i\cdots j}_{k\cdots l,m}. \tag{2.37}$$

Then if $A^{i\cdots j}_{k\cdots l}$ is a tensor, differentiation of the general tensor transformation (2.32) yields

$$A^{i'\cdots j'}_{k'\cdots l',m'} = A^{i\cdots j}_{k\cdots l,m}\, p^{i'}_{i} \cdots p^{j'}_{j}\, p^{k}_{k'} \cdots p^{l}_{l'}\, p^{m}_{m'} + P_1 + P_2 + \cdots , \tag{2.38}$$

where P_1, P_2, *etc.,* are terms involving derivatives of the p's. Clearly, $A^{i\cdots j}_{k\cdots l}$ is not a tensor under a general coordinate transformation. However, under a linear coordinate transformation (p's constant) $A^{i'\cdots j'}_{k'\cdots l',m'}$ behaves as a tensor of the type indicated by the suffixes, since the P_1, P_2, *etc.,* all vanish. Similarly, all higher partial derivatives,

$$A^{i\cdots j}_{k\cdots l,mn} = \frac{\partial A^{i\cdots j}_{k\cdots l}}{\partial x^m \partial x^n} \tag{2.39}$$

etc., also behave as tensors under linear transformations. Each partial differentiation has the effect of adding a new covariant suffix.

So far the space to which the coordinates x^i refer has been without structure. We can impose a structure on it by defining the distance between all pairs of neighbouring points by means of a *metric*

$$ds^2 = g_{ij}\, dx^i dx^j , \tag{2.40}$$

where the g_{ij} are functions of position. We can assume that $g_{ij} = g_{ji}$ without loss of generality. The above metric is analogous to, but more general than, the metric of Euclidian n-space, $ds^2 = (dx^1)^2 + (dx^2)^2 + \cdots + (dx^n)^2$. A space whose structure is determined by a metric of the type (2.40) is called *Riemannian*. Since ds^2 is invariant, it follows from a simple extension of the quotient rule that g_{ij} must be a tensor. It is called the *metric tensor*.

The elements of the inverse of the matrix g_{ij} are denoted by g^{ij}. These elements are uniquely defined by the equations

$$g^{ij} g_{jk} = \delta^i_k. \tag{2.41}$$

It is easily seen that the g^{ij} constitute the elements of a contravariant tensor. This tensor is said to be *conjugate* to g_{ij}. The conjugate metric tensor is symmetric (*i.e.,* $g^{ij} = g^{ji}$) just like the metric tensor itself.

The tensors g_{ij} and g^{ij} allow us to introduce the important operations of *raising* and *lowering suffixes*. These operations consist of forming inner products of a given tensor with g_{ij} or g^{ij}. For example, given a contravariant vector A^i, we define its covariant components A_i by the equation

$$A_i = g_{ij} A^j. \tag{2.42}$$

Conversely, given a covariant vector B_i, we can define its contravariant components B^i by the equations

$$B^i = g^{ij} B_j. \tag{2.43}$$

More generally, we can raise or lower any or all of the free suffixes of any given tensor. Thus, if A_{ij} is a tensor we define $A^i{}_j$ by the equation

$$A^i{}_j = g^{ip} A_{pj}. \tag{2.44}$$

Note that once the operations of raising and lowering suffixes has been defined the order of raised suffixes relative to lowered suffixes becomes significant.

By analogy with Euclidian space we define the *squared magnitude* $(A)^2$ of a vector A^i with respect to the metric $g_{ij} dx^i dx^j$ by the equation

$$(A)^2 = g_{ij} A^i A^j = A_i A^i. \tag{2.45}$$

A vector A^i termed a *null vector* if $(A)^2 = 0$. Two vectors A^i and B^i are said to be *orthogonal* if their inner product vanishes, *i.e.*, if

$$g_{ij} A^i B^j = A_i B^i = A^i B_i = 0. \tag{2.46}$$

Finally, let us consider differentiation with respect to distance s. The *tangent vector* dx^i/ds to a given curve in space is a contravariant tensor, since

$$\frac{dx^{i'}}{ds} = \frac{\partial x^{i'}}{\partial x^i} \frac{dx^i}{ds} = \frac{dx^i}{ds}\, p_i^{i'}. \tag{2.47}$$

The derivative $d(A^{i\cdots j}{}_{k\cdots l})/ds$ of some tensor with respect to distance is not, in general, a tensor, since

$$\frac{d(A^{i\cdots j}{}_{k\cdots l})}{ds} = A^{i\cdots j}{}_{k\cdots l,m} \frac{dx^m}{ds}, \tag{2.48}$$

and, as we have seen, the first factor on the right is not generally a tensor. However, under linear transformations it behaves as a tensor, so under linear transformations the derivative of a tensor with respect to distance behaves as a tensor of the same type.

2.5 Transformations

In this course we shall only concern ourselves with coordinate transformations which transform an inertial frame into another inertial frame. This limits us to four classes of transformations: displacements of the coordinate axes, rotations of the coordinate axes, parity reversals (*i.e.*, $x, y, z \to -x, -y, -z$), and Lorentz transformations. All of these transformations possess *group properties*. As a reminder, the requirements for an abstract multiplicative group are:

(i) The product of two elements is an element of the group.
(ii) The associative law $(ab)c = a(bc)$ holds.
(iii) There is a unit element e satisfying $ae = ea = a$ for all a.
(iv) Each element a possesses an inverse a^{-1} such that $a^{-1}a = aa^{-1} = e$.

Consider Lorentz transformations (in the standard configuration). It is easily demonstrated that the resultant of two successive Lorentz transformations, with velocities v_1 and v_2, respectively, is equivalent to a Lorentz transformation with velocity $v = (v_1+v_2)/(1+v_1v_2/c^2)$. Lorentz transformations obviously satisfy the associative law. The unit element of the transformation group is just a Lorentz transformation with $v = 0$. Finally, the inverse of a Lorentz transformation with velocity v is a transformation with velocity $-v$. We can use similar arguments to show that translations, rotations, parity inversions, and general Lorentz transformations (*i.e.*, transformations between frames which are not in the standard configuration) also possess group properties.

If we think carefully, we can see that the group properties of the above mentioned transformations are a direct consequence of the relativity principle. Let us again consider Lorentz transformations. Suppose that we have three inertial frames S, S', and S''. According to (i), if we can get from S to S' by a

Lorentz transformation, and from S' to S'' by a second Lorentz transformation, then it must always be possible to go directly from S to S'' by means of a third Lorentz transformation. Suppose, for the sake of argument, that we can find three frames for which this is not the case. In this situation, the frame S' could be distinguished from the frame S'' because it is possible to make a direct Lorentz transformation from S to the former frame, but not to the latter. This violates the relativity principle and, therefore, this situation can never arise. We can use a similar argument to demonstrate that a Lorentz transformation must possess an inverse. The associative law and the requirement that a unit element exists are trivially satisfied.

2.6 The physical significance of tensors

One of the central tenets of physics is that experiments should be repeatable. In other words, if somebody performs a physical experiment today and obtains a certain result, then somebody else performing the same experiment next week ought to obtain the same result, within the experimental errors. Presumably, in performing these hypothetical experiments both experimentalists find it necessary to set up a coordinate frame. Usually, these two frames do not coincide. After all, the experiments are, in general, performed in different places and at different times. Also, the two experimentalists are likely to orientate their coordinate axes differently. For instance, one experimentalist might align his x-axis with the North Star, whilst the other might align the same axis to point towards Mecca. Nevertheless, we still expect both experiments to yield the same result. What exactly do we mean by this statement? We do not mean that both experimentalists will obtain the same numbers when they measure something. For instance, the numbers used to denote the position of a point (*i.e.*, the coordinates of the point) are, in general, different in different coordinate frames. What we do expect is that any physically significant interrelation between physical quantities (*i.e.*, position, velocity, *etc.*) which appears to hold in the coordinate system of the first experimentalist will also appear to hold in the coordinate system of the second experimentalist. We usually refer to such interrelationships as "laws of physics." So, what we are really saying is that the laws of physics do not depend on our choice of coordinate system. In particular, if a law of physics is true in one coordinate system then it is automatically true in every other coordinate system, subject to the proviso that both coordinate systems are inertial.

Recall that tensors are geometric objects which possess the property that if a certain interrelationship holds between various tensors in one particular coordinate system, then the same interrelationship holds in any other coordinate system which is related to the first system by a certain class of transformations. It follows that *the laws of physics are expressible as interrelationships between tensors.* In special relativity the laws of physics are only required to exhibit tensor behaviour under transformations between different inertial frames; *i.e.*, translations, rotations, and Lorentz transformations. This set of transformations forms a group known as the *Poincaré group.* Parity inversion is a special type of transformation, and will be dealt with later on. In general relativity the laws of physics are required to exhibit tensor behaviour under *all* non-singular coordinate transformations.

Consider Newton's first law of motion. These take the form of three differential equations,

$$m\,\frac{d^2x}{dt^2} = f_x, \tag{2.49a}$$

$$m\,\frac{d^2y}{dt^2} = f_y, \tag{2.49b}$$

$$m\,\frac{d^2z}{dt^2} = f_z, \tag{2.49c}$$

in a general inertial frame. However, we can also write them as a single vector differential equation,

$$m\,\frac{d^2\boldsymbol{r}}{dt^2} = \boldsymbol{f}. \tag{2.50}$$

What is the advantage of the vector notation? Many people would say that it is just a convenient form of shorthand. However, there is another, far more important, advantage. Before we can accept Newton's first law of physics as a proper law of physics we need to convince ourselves that it is coordinate independent; *i.e.*, that it also holds in coordinate frames which are related to the original frame via a general translation or rotation of the coordinate axes. It is indeed possible to prove this, but the demonstration is rather tedious because a general rotation is a rather complicated transformation. A vector is a geometric object (in fact, it is a rank one tensor in three dimensional Euclidean space) whose three components transform under a general translation and rotation of the coordinate axes in an analogous manner to the difference in coordinates between two fixed points in space. This ensures that any vector equation which is true in one coordinate frame is also true in any other coordinate frame which is related to the original frame via a general rotation or translation of the axes. Thus, the main advantage of Eq. (2.50) is that it makes the coordinate independent nature of Newton's first law of motion manifestly obvious. Of course, we cannot deny that Newton's first law also looks simpler when it is expressed in terms of vectors. This is one example of a rather general feature of physical laws. Namely, *when the laws of physics are expressed in a manner which makes their invariance under various transformation groups manifest then they tend to take a particularly simple form.* In general, the larger the group of transformations the simpler the form taken by the laws of physics. One of the major goals of modern physics is to find the largest possible group of transformations under which the laws of physics are invariant, and then prove that when expressed in a manner which makes this invariance manifest these laws reduce to a single unifying principle.

We already know how to write the laws of physics in terms of vectors and vector fields. This means that these laws are automatically invariant under translations and rotations. However, according to the relativity principle, there is a third class of transformations under which the laws of physics must also be invariant; namely, Lorentz transformations. There are two ways in which we could verify that the laws of physics are Lorentz invariant. The direct method is extremely tedious, since Lorentz transformations are rather complicated. An alternative method is to write the laws of physics in terms of geometric objects which transform as tensors under translations, rotations, *and* Lorentz transformations. This method has the advantage that it makes the Lorentz invariant nature of the laws of physics obvious. We also expect that when the laws of physics are written in manifestly Lorentz invariant form then they will look even simpler than they do when written just in terms of vectors. The laws of electromagnetism provide a particularly good illustration of this effect.

2.7 Space-time

In special relativity we are only allowed to use inertial frames to assign coordinates to events. There are many different types of inertial frames. However, it is convenient to adhere to those with *standard coordinates*. That is, spatial coordinates which are right-handed rectilinear Cartesians based on a standard unit of length and time-scales based on a standard unit of time. We shall continue to assume that we are employing standard coordinates. However, from now on we shall make no assumptions, unless specifically stated, about the relative configuration of the two sets of spatial axes and the origins of time when dealing with two inertial frames. Thus, the most general transformation between two inertial frames consists of a Lorentz transformation in the standard configuration plus a translation (this includes a translation in time) and a rotation of the coordinate axes. The resulting transformation is called a *general Lorentz transformation*, as opposed to a Lorentz transformation in the standard configuration which will henceforth be termed a *standard Lorentz transformation*.

In Section 2.2 we proved quite generally that corresponding differentials in two inertial frames S and S' satisfy the relation

$$dx^2 + dy^2 + dz^2 - c^2 dt^2 = {dx'}^2 + {dy'}^2 + {dz'}^2 - c^2 {dt'}^2. \tag{2.51}$$

Thus, we expect this relation to remain invariant under a general Lorentz transformation. Since such a transformation is *linear* it follows that

$$\begin{aligned}(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 - c^2(t_2 - t_1)^2 &= \\ (x_2' - x_1')^2 + (y_2' - y_1')^2 + (z_2' - z_1')^2 - c^2(t_2' - t_1')^2&,\end{aligned} \tag{2.52}$$

where (x_1, y_1, z_1, t_1) and (x_2, y_2, z_2, t_2) are the coordinates of any two events in S and the primed symbols denote the corresponding coordinates in S'. It is convenient to write

$$-dx^2 - dy^2 - dz^2 + c^2 dt^2 = ds^2, \tag{2.53}$$

and

$$-(x_2 - x_1)^2 - (y_2 - y_1)^2 - (z_2 - z_1)^2 + c^2(t_2 - t_1)^2 = s^2. \tag{2.54}$$

The differential ds, or the finite number s, defined by these equations is called the *interval* between the corresponding events. Equations (2.51) and (2.52) express the fact that *the interval between two events is invariant*, in the sense that it has the same value in all inertial frames. In other words, the interval between two events is invariant under a general Lorentz transformation.

Let us consider entities defined in terms of four variables

$$x^1 = x, \quad x^2 = y, \quad x^3 = z, \quad x^4 = ct, \tag{2.55}$$

and which transform as tensors (see Eqs. (2.30)–(2.32)) under a general Lorentz transformation. From now on such entities will be referred to as *4-tensors*.

Tensor analysis cannot proceed very far without the introduction of a non-singular tensor g_{ij}, the so-called *fundamental tensor*, which is used to define the operations of raising and lowering suffixes (see Eqs. (2.42)–(2.44)). The fundamental tensor is usually introduced using a metric $ds^2 = g_{ij}\,dx^i dx^j$, where ds^2 is a differential invariant. We have already come across such an invariant, namely

$$\begin{aligned} ds^2 &= -dx^2 - dy^2 - dz^2 + c^2 dt^2 \\ &= -(dx^1)^2 - (dx^2)^2 - (dx^3)^2 + (dx^4)^2 \\ &= g_{\mu\nu}\,dx^\mu dx^\nu, \end{aligned} \tag{2.56}$$

where μ, ν run from 1 to 4. Note that the use of Greek suffixes is conventional in 4-tensor theory. Roman suffixes are reserved for tensors in three dimensional Euclidian space, so-called 3-tensors. The 4-tensor $g_{\mu\nu}$ has the components $g_{11} = g_{22} = g_{33} = -1, g_{44} = 1$, and $g_{\mu\nu} = 0$ when $\mu \neq \nu$, in all permissible coordinate frames. From now on $g_{\mu\nu}$, as defined above, is adopted as the fundamental tensor for 4-tensors. $g_{\mu\nu}$ can be thought of as the *metric tensor* of the "space" whose points are the events (x^1, x^2, x^3, x^4). This "space" is usually referred to as *space-time*, for obvious reasons. Note that space-time cannot be regarded as a straightforward generalization of Euclidian 3-space to four dimensions, with time as the fourth dimension. The distribution of signs in the metric ensures that the time coordinate x^4 is not on the same footing as the three space coordinates. Thus, space-time has a non-isotropic nature which is quite unlike Euclidian space with its positive definite metric. According to the relativity principle, all physical laws are expressible as interrelationships between 4-tensors in space-time.

A tensor of rank one is called a *4-vector*. We shall also have occasion to use ordinary vectors in three dimensional Euclidian space. Such vectors are called *3-vectors* and are conventionally represented by boldface symbols. We shall use the Latin suffixes i, j, k, *etc.* to denote the components of a 3-vector; these suffixes are understood to range from 1 to 3. Thus, $\boldsymbol{u} = u^i = dx^i/dt$ denotes a velocity vector. For 3-vectors we shall use the notation $u^i = u_i$ interchangeably; *i.e.*, the level of the suffix has no physical significance.

When tensor transformations from one frame to another actually have to be computed, we shall usually find it possible to choose coordinates in the standard configuration, so that the standard Lorentz transform applies. Under it, any contravariant 4-vector T^μ transforms according to the same scheme as the difference in coordinates $x_2^\mu - x_1^\mu$ between two points in space-time. It follows that

$$T^{1'} = \gamma(T^1 - \beta\, T^4), \tag{2.57a}$$

$$T^{2'} = T^2, \tag{2.57b}$$

$$T^{3'} = T^3, \tag{2.57c}$$

$$T^{4'} = \gamma(T^4 - \beta\, T^1), \tag{2.57d}$$

where $\beta = v/c$. Higher rank 4-tensors transform according to the rules (2.30)–(2.32). The transformation coefficients take the form

$$p_\mu^{\mu'} = \left(\begin{array}{cccc} \gamma & 0 & 0 & -\gamma\beta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\gamma\beta & 0 & 0 & \gamma \end{array}\right) \tag{2.58a}$$

$$p_{\mu'}^{\mu} = \left(\begin{array}{cccc} \gamma & 0 & 0 & \gamma\beta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \gamma\beta & 0 & 0 & \gamma \end{array}\right) \tag{2.58b}$$

Often the first three components of a 4-vector coincide with the components of a 3-vector. For example, the x^1, x^2, x^3 in $R^\mu = (x^1, x^2, x^3, x^4)$ are the components of $\boldsymbol{r}$, the position 3-vector of the point at which the event occurs. In such cases we adopt the notation exemplified by $R^\mu = (\boldsymbol{r}, ct)$. The covariant form of such a vector is simply $R_\mu = (-\boldsymbol{r}, ct)$. The squared magnitude of the vector is $(R)^2 = R_\mu R^\mu = -r^2 + c^2 t^2$. The inner product $g_{\mu\nu} R^\mu Q^\nu = R_\mu Q^\mu$ of R^μ with a similar vector $Q^\mu = (\boldsymbol{q}, k)$ is given by $R_\mu Q^\mu = -\boldsymbol{r}\cdot\boldsymbol{q} + ct\, k$. The vectors R^μ and Q^μ are said to be *orthogonal* if $R_\mu Q^\mu = 0$.

Since a general Lorentz transformation is a *linear* transformation, the partial derivative of a 4-tensor is also a 4-tensor;

$$\frac{\partial A^{\nu\sigma}}{\partial x^\mu} = A^{\nu\sigma}{}_{,\mu}. \tag{2.59}$$

Clearly, a general 4-tensor acquires an extra covariant index after partial differentiation with respect to the contravariant coordinate x^μ. It is helpful to define a covariant derivative operator

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left(\nabla, \frac{1}{c}\frac{\partial}{\partial t}\right), \tag{2.60}$$

where

$$\partial_\mu A^{\nu\sigma} \equiv A^{\nu\sigma}{}_{,\mu}. \tag{2.61}$$

There is a corresponding contravariant derivative operator

$$\partial^\mu \equiv \frac{\partial}{\partial x_\mu} = \left(-\nabla, \frac{1}{c}\frac{\partial}{\partial t}\right), \tag{2.62}$$

where

$$\partial^\mu A^{\nu\sigma} \equiv g^{\mu\tau} A^{\nu\sigma}{}_{,\tau}. \tag{2.63}$$

The 4-divergence of a 4-vector $A^\mu = (\boldsymbol{A}, A^0)$ is the invariant

$$\partial^\mu A_\mu = \partial_\mu A^\mu = \nabla\cdot\boldsymbol{A} + \frac{1}{c}\frac{\partial A^0}{\partial t}. \tag{2.64}$$

The four dimensional Laplacian operator, or *d'Alembertian*, is equivalent to the invariant contraction

$$\Box \equiv \partial_\mu \partial^\mu = -\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}. \tag{2.65}$$

Recall that we still need to prove (from Section 2.2) that the invariance of the differential metric,

$$ds^2 = dx'^2 + dy'^2 + dz'^2 - c^2 dt'^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2, \tag{2.66}$$

between two general inertial frames implies that the coordinate transformation between such frames is necessarily linear. To put it another way, we need to demonstrate that a transformation which transforms a metric $g_{\mu\nu}\, dx^\mu dx^\nu$ with constant coefficients into a metric $g_{\mu'\nu'}\, dx^{\mu'} dx^{\nu'}$ with constant coefficients must be linear. Now

$$g_{\mu\nu} = g_{\mu'\nu'}\, p^{\mu'}_{\mu}\, p^{\nu'}_{\nu}. \tag{2.67}$$

Differentiating with respect to x^σ we get

$$g_{\mu'\nu'}\, p^{\mu'}_{\mu\sigma} p^{\nu'}_{\nu} + g_{\mu'\nu'}\, p^{\mu'}_{\mu}\, p^{\nu'}_{\nu\sigma} = 0, \tag{2.68}$$

where

$$p^{\mu'}_{\mu\sigma} = \frac{\partial p^{\mu'}_{\mu}}{\partial x^\sigma} = \frac{\partial^2 x^{\mu'}}{\partial x^\mu \partial x^\sigma} = p^{\mu'}_{\sigma\mu} \tag{2.69}$$

etc. Interchanging the indices μ and σ yields

$$g_{\mu'\nu'}\, p^{\mu'}_{\mu\sigma} p^{\nu'}_{\nu} + g_{\mu'\nu'}\, p^{\mu'}_{\sigma}\, p^{\nu'}_{\nu\mu} = 0. \tag{2.70}$$

Interchanging the indices ν and σ gives

$$g_{\mu'\nu'}\, p^{\mu'}_{\sigma}\, p^{\nu'}_{\nu\mu} + g_{\mu'\nu'}\, p^{\mu'}_{\mu}\, p^{\nu'}_{\nu\sigma} = 0, \tag{2.71}$$

where the indices μ' and ν' have been interchanged in the first term. It follows from Eqs. (2.68), (2.70), and (2.71) that

$$g_{\mu'\nu'}\, p^{\mu'}_{\mu\sigma} p^{\nu'}_{\nu} = 0. \tag{2.72}$$

Multiplication by $p^{\nu}_{\sigma'}$ yields

$$g_{\mu'\nu'}\, p^{\mu'}_{\mu\sigma} p^{\nu'}_{\nu}\, p^{\nu}_{\sigma'} = g_{\mu'\sigma'}\, p^{\mu'}_{\mu\sigma} = 0. \tag{2.73}$$

Finally, multiplication by $g^{\nu'\sigma'}$ gives

$$g_{\mu'\sigma'} g^{\nu'\sigma'}\, p^{\mu'}_{\mu\sigma} = p^{\nu'}_{\mu\sigma} = 0. \tag{2.74}$$

This proves that the coefficients $p^{\nu'}_{\mu}$ are constants and, hence, that the transformation is linear.

2.8 Proper time

It is often helpful to write the invariant differential interval ds^2 in the form

$$ds^2 = c^2\, d\tau^2. \tag{2.75}$$

The quantity $d\tau$ is called the *proper time*. It follows that

$$d\tau^2 = -\frac{dx^2 + dy^2 + dz^2}{c^2} + dt^2. \tag{2.76}$$

Consider a series of events on the world-line of some material particle. If the particle has speed u then

$$d\tau^2 = dt^2 \left[-\frac{dx^2 + dy^2 + dz^2}{c^2 dt^2} + 1 \right] = dt^2 \left(1 - \frac{u^2}{c^2} \right), \tag{2.77}$$

implying that

$$\frac{dt}{d\tau} = \gamma(u). \tag{2.78}$$

It is clear that $dt = d\tau$ in the particle's rest frame. Thus, $d\tau$ corresponds to the time difference between two neighbouring events on the particle's world-line, as measured by a clock attached to the particle (hence, the name "proper time"). According to Eq. (2.78), the particle's clock appears to run slow, by a factor $\gamma(u)$, in an inertial frame in which the particle is moving with velocity u. This is the celebrated *time dilation* effect.

Let us consider how a small 4-dimensional volume element in space-time transforms under a general Lorentz transformation. We have

$$d^4x' = \mathcal{J}\, d^4x, \tag{2.79}$$

where

$$\mathcal{J} = \frac{\partial(x^{1'}, x^{2'}, x^{3'}, x^{4'})}{\partial(x^1, x^2, x^3, x^4)} \tag{2.80}$$

is the Jacobian of the transformation; *i.e.*, the determinant of the transformation matrix $p^{\mu'}_{\mu}$. A general Lorentz transformation is made up of a standard Lorentz transformation plus a displacement and a rotation. Thus, the transformation matrix is the *product* of that for a standard Lorentz transformation, a translation, and a rotation. It follows that the Jacobian of a general Lorentz transformation is the product of that for a standard Lorentz transformation, a translation, and a rotation. It is well known that the Jacobian of the latter two transformations is unity, since they are both volume preserving transformations which do not affect time. Likewise, it is easily seen (*e.g.*, by taking the determinant of the transformation matrix (2.58a)) that the Jacobian of a standard Lorentz transformation is also unity. It follows that

$$d^4x' = d^4x \tag{2.81}$$

for a general Lorentz transformation. In other words, a general Lorentz transformation preserves the volume of space-time. Since time is dilated by a factor γ in a moving frame, the volume of space-time can only be preserved if the volume of ordinary 3-space is reduced by the same factor. As is well known, this is achieved by *length contraction* along the direction of motion by a factor γ.

2.9 4-velocity and 4-acceleration

We have seen that the quantity dx^μ/ds transforms as a 4-vector under a general Lorentz transformation (see Eq. (2.47)). Since $ds \propto d\tau$ it follows that

$$U^\mu = \frac{dx^\mu}{d\tau} \tag{2.82}$$

also transforms as a 4-vector. This quantity is known as the *4-velocity*. Likewise, the quantity

$$A^\mu = \frac{d^2x^\mu}{d\tau^2} = \frac{dU^\mu}{d\tau} \tag{2.83}$$

is a 4-vector, and is called the *4-acceleration*.

For events along the world-line of a particle traveling with 3-velocity $\boldsymbol{u}$ we have

$$U^\mu = \frac{dx^\mu}{d\tau} = \frac{dx^\mu}{dt}\frac{dt}{d\tau} = \gamma(u)(\boldsymbol{u}, c), \tag{2.84}$$

where use has been made of Eq. (2.78). This gives the relationship between a particle's 3-velocity and its 4-velocity. The relationship between the 3-acceleration and the 4-acceleration is less straightforward. We have

$$A^\mu = \frac{dU^\mu}{d\tau} = \gamma\,\frac{dU^\mu}{dt} = \gamma\frac{d}{dt}(\gamma\boldsymbol{u}, \gamma c) = \gamma\left(\frac{d\gamma}{dt}\,\boldsymbol{u} + \gamma\boldsymbol{a}, c\,\frac{d\gamma}{dt}\right), \tag{2.85}$$

where $\boldsymbol{a} = d\boldsymbol{u}/dt$ is the 3-acceleration. In the rest frame of the particle $U^\mu = (\boldsymbol{0}, c)$ and $A^\mu = (\boldsymbol{a}, 0)$. It follows that

$$U_\mu A^\mu = 0 \tag{2.86}$$

(note that $U_\mu A^\mu$ is an invariant quantity). In other words, the 4-acceleration of a particle is always orthogonal to its 4-velocity.

2.10 The current density 4-vector

Let us now consider the laws of electromagnetism. We wish to demonstrate that these laws are compatible with the relativity principle. In order to achieve this it is necessary for us to make an *assumption* about the transformation properties of electric charge. The assumption which we shall make, which is well substantiated experimentally, is that charge, unlike mass, is invariant. That is, the charge carried by a given particle has the same measure in all inertial frames. In particular, the charge carried by a particle does not vary with the particle's velocity.

Let us suppose, following Lorentz, that all charge is made up of elementary particles, each carrying the invariant amount e. Suppose that n is the number density of such charges at some given point and time, moving with velocity $\boldsymbol{u}$, as observed in a frame S. Let n_0 be the number density of charges in the frame S_0 in which the charges are momentarily at rest. As is well known, a volume of measure V in S has measure $\gamma(u)\,V$ in S_0 (because of length contraction). Since observers in both frames must agree on how many particles are contained in the volume, and, hence, on how much charge it contains, it follows that $n = \gamma(u)\,n_0$. If $\rho = en$ and $\rho_0 = en_0$ are the charge densities in S and S_0, respectively, then

$$\rho = \gamma(u)\rho_0. \tag{2.87}$$

The quantity ρ_0 is called the *proper density* and is obviously Lorentz invariant.

Suppose that x^μ are the coordinates of the moving charge in S. The *current density 4-vector* is constructed as follows:

$$J^\mu = \rho_0 \frac{dx^\mu}{d\tau} = \rho_0 U^\mu. \tag{2.88}$$

Thus,

$$J^\mu = \rho_0\, \gamma(u)(\boldsymbol{u}, c) = (\boldsymbol{j}, \rho c), \tag{2.89}$$

where $\boldsymbol{j} = \rho \boldsymbol{u}$ is the current density 3-vector. Clearly, charge density and current density transform as the time-like and space-like components of the same 4-vector.

Consider the invariant 4-divergence of J^μ:

$$\partial_\mu J^\mu = \nabla\cdot\boldsymbol{j} + \frac{\partial\rho}{\partial t}. \tag{2.90}$$

We know that one of the caveats of Maxwell's equations is the charge conservation law

$$\frac{\partial\rho}{\partial t} + \nabla\cdot\boldsymbol{j} = 0. \tag{2.91}$$

It is clear that this expression can be rewritten in the manifestly Lorentz invariant form

$$\partial_\mu J^\mu = 0. \tag{2.92}$$

This equation tells us that there are no net sources or sinks of electric charge in nature; *i.e.*, electric charge is neither created nor destroyed.

2.11 The potential 4-vector

There are many ways of writing the laws of electromagnetism. However, the most obviously Lorentz invariant way is to write them in terms of the vector and scalar potentials. When written in this fashion, Maxwell's equations reduce to

$$\left(-\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\phi = \frac{\rho}{\epsilon_0}, \tag{2.93a}$$

$$\left(-\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{A} = \mu_0 \boldsymbol{j}, \tag{2.93b}$$

where ϕ is the scalar potential and $\boldsymbol{A}$ is the vector potential. Note that the differential operator appearing in these equations is the Lorentz invariant d'Alembertian, defined in Eq. (2.65). The above pair of equations can be rewritten in the form

$$\Box\phi = \frac{\rho\, c}{c\,\epsilon_0}, \tag{2.94a}$$

$$\Box\, c\boldsymbol{A} = \frac{\boldsymbol{j}}{c\,\epsilon_0}. \tag{2.94b}$$

Maxwell's equations can be written in Lorentz invariant form provided that the entity

$$\Phi^\mu = (c\boldsymbol{A}, \phi) \tag{2.95}$$

transforms as a contravariant 4-vector. This entity is known as the *potential 4-vector*. It follows from Eqs. (2.89), (2.94), and (2.95) that

$$\Box\Phi^\mu = \frac{J^\mu}{c\,\epsilon_0}. \tag{2.96}$$

Thus, the field equations which govern classical electromagnetism can all be summed up in a single 4-vector equation.

2.12 Gauge invariance

The electric and magnetic fields are obtained from the vector and scalar potentials according to the prescription

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial\boldsymbol{A}}{\partial t}, \tag{2.97a}$$

$$\boldsymbol{B} = \nabla\wedge\boldsymbol{A}. \tag{2.97b}$$

These fields are important because they determine the electromagnetic forces exerted on charged particles. Note that the above prescription does not uniquely determine the two potentials. It is possible to make the following transformation, known as a *gauge transformation*, which leaves the fields unaltered:

$$\phi \rightarrow \phi + \frac{\partial\psi}{\partial t}, \tag{2.98a}$$

$$\boldsymbol{A} \rightarrow \boldsymbol{A} - \nabla\psi, \tag{2.98b}$$

where $\psi(\boldsymbol{r}, t)$ is a general scalar field. It is necessary to adopt some form of convention, generally known as a *gauge condition*, to fully specify the two potentials. In fact, there is only one gauge condition which is consistent with Eqs. (2.93). This is the *Lorentz gauge condition*,

$$\frac{1}{c^2}\frac{\partial\phi}{\partial t} + \nabla\cdot\boldsymbol{A} = 0. \tag{2.99}$$

Note that this condition can be written in the Lorentz invariant form

$$\partial_\mu \Phi^\mu = 0. \tag{2.100}$$

This implies that if the Lorentz gauge holds in one particular inertial frame then it automatically holds in all other inertial frames. A general gauge transformation can be written

$$\Phi^\mu \to \Phi^\mu + c\,\partial^\mu\psi. \tag{2.101}$$

Note that even after the Lorentz gauge has been adopted the potentials are undetermined to a gauge transformation using a scalar field ψ which satisfies the sourceless wave equation

$$\Box\psi = 0. \tag{2.102}$$

However, if we adopt "sensible" boundary conditions in both space and time then the only solution to the above equation is $\psi = 0$.

2.13 Solution of the inhomogeneous wave equation

Equations (2.93) all have the general form

$$\Box\psi(\boldsymbol{r}, t) = g(\boldsymbol{r}, t). \tag{2.103}$$

Can we find a *unique* solution to the above equation? Let us assume that the source function $g(\boldsymbol{r}, t)$ can be expressed as a Fourier integral

$$g(\boldsymbol{r}, t) = \int_{-\infty}^{\infty} g_\omega(\boldsymbol{r})\, e^{-\mathrm{i}\,\omega t}\, d\omega. \tag{2.104}$$

The inverse transform is

$$g_\omega(\boldsymbol{r}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(\boldsymbol{r},t)\, e^{\,\mathrm{i}\,\omega t}\, dt. \tag{2.105}$$

Similarly, we may write the general potential $\psi(\boldsymbol{r},t)$ as a Fourier integral

$$\psi(\boldsymbol{r},t) = \int_{-\infty}^{\infty} \psi_\omega(\boldsymbol{r})\, e^{-\mathrm{i}\,\omega t}\, d\omega, \tag{2.106}$$

with the corresponding inverse

$$\psi_\omega(\boldsymbol{r}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi(\boldsymbol{r},t)\, e^{\,\mathrm{i}\,\omega t}\, dt. \tag{2.107}$$

Fourier transformation of Eq. (2.103) yields

$$(\nabla^2 + k^2)\psi_\omega = -g_\omega, \tag{2.108}$$

where $k = \omega/c$.

The above equation, which reduces to Poisson's equation in the limit $k \to 0$, and is called *Helmholtz's equation*, is linear, so we may attempt a Green's function method of solution. Let us try to find a function $G_\omega(\boldsymbol{r},\boldsymbol{r}')$ such that

$$(\nabla^2 + k^2)G_\omega(\boldsymbol{r},\boldsymbol{r}') = -\delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{2.109}$$

The general solution is then

$$\psi_\omega(\boldsymbol{r}) = \int g_\omega(\boldsymbol{r}')\, G_\omega(\boldsymbol{r},\boldsymbol{r}')\, dV'. \tag{2.110}$$

The "sensible" spatial boundary conditions which we impose are that $G_\omega(\boldsymbol{r},\boldsymbol{r}') \to 0$ as $|\boldsymbol{r}-\boldsymbol{r}'| \to \infty$. In other words, the field goes to zero a long way from the source. Since the system we are solving is spherically symmetric about the point $\boldsymbol{r}'$ it is plausible that the Green's function itself is spherically symmetric. It follows that

$$\frac{1}{R}\frac{d^2(R\,G_\omega)}{dR^2} + k^2 G_\omega = -\delta(\boldsymbol{R}), \tag{2.111}$$

where $\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}'$ and $R = |\boldsymbol{R}|$. The most general solution to the above equation in the region $R > 0$ is[1]

$$G_\omega(R) = \frac{A\,e^{\,\mathrm{i}\,kR} + B\,e^{-\mathrm{i}\,kR}}{4\pi\,R}. \tag{2.112}$$

We know that in the limit $k \to 0$ the Green's function for Helmholtz's equation must tend towards that for Poisson's equation, which is

$$G_\omega(R) = \frac{1}{4\pi\,R}. \tag{2.113}$$

This is only the case if $A + B = 1$.

Reconstructing $\psi(\boldsymbol{r}, t)$ from Eqs. (2.106), (2.110), and (2.112), we obtain

$$\psi(\boldsymbol{r}, t) = \frac{1}{4\pi}\int\!\!\int \frac{g_\omega(\boldsymbol{r}')}{R}\left[A\,e^{-\mathrm{i}\,\omega(t-R/c)} + B\,e^{-\mathrm{i}\,\omega(t+R/c)}\right]\,d\omega\,dV'. \tag{2.114}$$

It follows from Eq. (2.104) that

$$\psi(\boldsymbol{r}, t) = \frac{A}{4\pi}\int \frac{g(\boldsymbol{r}', t - R/c)}{R}\,dV' + \frac{B}{4\pi}\int \frac{g(\boldsymbol{r}', t + R/c)}{R}\,dV'. \tag{2.115}$$

Now, the real space Green's function for the inhomogeneous wave equation (2.103) satisfies

$$\Box G(\boldsymbol{r}, \boldsymbol{r}'; t, t') = \delta(\boldsymbol{r} - \boldsymbol{r}')\,\delta(t - t'). \tag{2.116}$$

Hence, the most general solution of this equation takes the form

$$\psi(\boldsymbol{r}, t) = \int\!\!\int g(\boldsymbol{r}', t')\,G(\boldsymbol{r}, \boldsymbol{r}'; t, t')\,dV'dt'. \tag{2.117}$$

Comparing Eqs. (2.115) and (2.117) we obtain

$$G(\boldsymbol{r}, \boldsymbol{r}'; t, t') = A\,G^{(+)}(\boldsymbol{r}, \boldsymbol{r}'; t, t') + B\,G^{(-)}(\boldsymbol{r}, \boldsymbol{r}'; t, t'), \tag{2.118}$$

[1]In principle, $A = A(\omega)$ and $B = B(\omega)$, with $A + B = 1$. However, later on we shall demonstrate that $B = 0$, otherwise causality is violated. It follows that $A = 1$. Thus, it is legitimate to assume, for the moment, that A and B are constants.

where

$$G^{(\pm)}(\boldsymbol{r}, \boldsymbol{r}'; t, t') = \frac{\delta(t' - [t \mp |\boldsymbol{r} - \boldsymbol{r}'|/c])}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}'|}, \tag{2.119}$$

and $A + B = 1$.

The real space Green's function specifies the response of the system to a point source at position $\mathbf{r}'$ which appears momentarily at time t'. According to the *retarded Green's function* $G^{(+)}$ the response consists of a spherical wave, centred on $\boldsymbol{r}'$, which propagates forward in time. In order for the wave to reach position $\boldsymbol{r}$ at time t it must have been emitted from the source at $\boldsymbol{r}'$ at the *retarded time* $t_r = t - |\mathbf{r} - \mathbf{r}'|/c$. According to the *advanced Green's function* $G^{(-)}$ the response consists of a spherical wave, centred on $\boldsymbol{r}'$, which propagates backward in time. Clearly, the advanced potential is not consistent with our ideas about causality, which demand that an effect can never precede its cause in time. Thus, the Green's function which is consistent with our experience is

$$G(\boldsymbol{r}, \boldsymbol{r}'; t, t') = G^{(+)}(\boldsymbol{r}, \boldsymbol{r}'; t, t') = \frac{\delta(t' - [t - |\boldsymbol{r} - \boldsymbol{r}'|/c])}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}'|}. \tag{2.120}$$

We are able to find solutions of the inhomogeneous wave equation (2.103) which propagate backward in time because this equation is time symmetric (*i.e.*, it is invariant under the transformation $t \to -t$).

In conclusion, the most general solution of the inhomogeneous wave equation (2.103) which satisfies sensible boundary conditions at infinity and is consistent with causality is

$$\psi(\boldsymbol{r}, t) = \int \frac{g(\boldsymbol{r}', t - |\boldsymbol{r} - \boldsymbol{r}'|/c)}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}'|}\, dV'. \tag{2.121}$$

This expression is sometimes written

$$\psi(\boldsymbol{r}, t) = \int \frac{[g(\boldsymbol{r}')]}{4\pi\, |\boldsymbol{r} - \boldsymbol{r}'|}\, dV', \tag{2.122}$$

where the rectangular bracket symbol [ ] denotes that the terms inside the bracket are to be evaluated at the retarded time $t - |\boldsymbol{r} - \boldsymbol{r}'|/c$. Note, in particular, from Eq. (2.122) that if there is no source (*i.e.*, $g(\boldsymbol{r}, t) = 0$) then there is no field (*i.e.*, $\psi(\boldsymbol{r}, t) = 0$). But, is the above solution really *unique*? Unfortunately, there is a weak link in our derivation, between Eqs. (2.110) and (2.111), where we *assume* that the Green's function for the Helmholtz equation subject to the boundary condition $G_\omega(\boldsymbol{r},\boldsymbol{r}') \to 0$ as $|\boldsymbol{r}-\boldsymbol{r}'| \to \infty$ is *spherically symmetric.* Let us try to fix this problem.

With the benefit of hindsight, we can see that the Green's function

$$G_\omega(R) = \frac{e^{\mathrm{i}\,kR}}{4\pi\,R} \tag{2.123}$$

corresponds to the retarded solution in real space and is, therefore, the correct physical Green's function. The Green's function

$$G_\omega(R) = \frac{e^{-\mathrm{i}\,kR}}{4\pi\,R} \tag{2.124}$$

corresponds to the advanced solution in real space and must, therefore, be rejected. We can select the retarded Green's function by imposing the following boundary condition at infinity

$$\lim_{R\to\infty} R\left(\frac{\partial G}{\partial R} - \mathrm{i}\,kG\right) = 0. \tag{2.125}$$

This is called the *Sommerfeld radiation condition*; it basically ensures that sources radiate waves instead of absorbing them. But, does this boundary condition *uniquely* select the spherically symmetric Green's function (2.123) as the solution of

$$(\nabla^2 + k^2)G_\omega(R,\theta,\varphi) = -\delta(\boldsymbol{R})? \tag{2.126}$$

Here, (R,θ,φ) are spherical polar coordinates. If it does then we can be sure that Eq. (2.122) represents the *unique* solution of the wave equation (2.103) which is consistent with causality.

Let us suppose that there are two solutions of Eq. (2.126) which satisfy the boundary condition (2.125) and revert to the unique Green's function for Poisson's equation (2.113) in the limit $R \to 0$. Let us call these solutions u_1 and u_2, and let us form the difference $w = u_1 - u_2$. Consider a surface Σ_0 which is a sphere of arbitrarily small radius centred on the origin. Consider a second surface Σ_∞ which is a sphere of arbitrarily large radius centred on the origin. Let V denote the volume enclosed by these surfaces. The difference function w satisfies the homogeneous Helmholtz equation,

$$(\nabla^2 + k^2) w = 0, \tag{2.127}$$

throughout V. According to Green's theorem

$$\int_V (w\nabla^2 w^* - w^*\nabla^2 w)\, dV = \left(\int_{\Sigma_0} + \int_{\Sigma_\infty}\right)\left(w\frac{\partial w^*}{\partial n} - w^*\frac{\partial w}{\partial n}\right) dS, \tag{2.128}$$

where $\partial/\partial n$ denotes a derivative normal to the surface in question. It is clear from Eq. (2.127) that the volume integral is zero. It is also clear that the first surface integral is zero, since both u_1 and u_2 must revert to the Green's function for Poisson's equation in the limit $R \rightarrow 0$. Thus,

$$\int_{\Sigma_\infty}\left(w\frac{\partial w^*}{\partial n} - w^*\frac{\partial w}{\partial n}\right) dS = 0. \tag{2.129}$$

Equation (2.127) can be written

$$\frac{\partial^2 (R\, w)}{\partial R^2} + \frac{D(R\, w)}{R^2} + k^2\, Rw = 0, \tag{2.130}$$

where D is the spherical harmonic operator

$$D = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{2.131}$$

The most general solution of Eq. (2.130) takes the form (see Section 7)

$$w(R,\theta,\varphi) = \sum_{l,m=0}^{\infty}\left[C_{lm}\, h_l^{(1)}(kR) + D_{lm}\, h_l^{(2)}(kR)\right] Y_{lm}(\theta,\varphi). \tag{2.132}$$

Here, the C_{lm} and D_{lm} are arbitrary coefficients, the Y_{lm} are spherical harmonics,[2] and

$$h_l^{(1,2)}(\rho) = \sqrt{\frac{\pi}{2\rho}}\, H^{1,2}_{l+1/2}(\rho), \tag{2.133}$$

[2] J.D. Jackson, *Classical Electrodynamics*, (Wiley, 1962), p. 99

where $H_n^{1,2}$ are Hankel functions of the first and second kind.[3] It can be demonstrated that[4]

$$H_n^1(\rho) = \sqrt{\frac{2}{\pi\rho}}\, e^{\,\mathrm{i}\,(\rho-(n+1/2)\pi/2)} \sum_{m=0,1,2,\cdots} \frac{(n,m)}{(-2\mathrm{i}\,\rho)^m}, \tag{2.134a}$$

$$H_n^2(\rho) = \sqrt{\frac{2}{\pi\rho}}\, e^{-\mathrm{i}\,(\rho-(n+1/2)\pi/2)} \sum_{m=0,1,2,\cdots} \frac{(n,m)}{(+2\mathrm{i}\,\rho)^m}, \tag{2.134b}$$

where

$$(n,m) = \frac{(4n^2-1)(4n^2-9)\cdots(4n^2-\{2m-1\}^2)}{2^{2m}\,m!} \tag{2.135}$$

and $(n,0) = 1$. Note that the summations in Eqs. (2.314) terminate after $n+1/2$ terms.

The large R behaviour of the $h_l^{(2)}$ is clearly inconsistent with the Sommerfeld radiation condition (2.125). It follows that all of the D_{lm} in Eq. (2.132) are zero. The most general solution can now be expressed in the form

$$w(R,\theta,\varphi) = \frac{e^{\,\mathrm{i}\,kR}}{R}\sum_{n=0}^{\infty}\frac{f_n(\theta,\varphi)}{R^n}, \tag{2.136}$$

where the $f_n(\theta,\varphi)$ are various weighted sums of the spherical harmonics. Substitution of this solution into the differential equation (2.130) yields

$$e^{\,\mathrm{i}\,kR}\sum_{n=0}^{\infty}\left(-\frac{2\mathrm{i}\,k\,n}{R^{n+1}} + \frac{n(n+1)}{R^{n+2}} + \frac{D}{R^{n+2}}\right) f_n = 0. \tag{2.137}$$

Replacing the index of summation n in the first term of the parentheses by $n+1$ we obtain

$$e^{\,\mathrm{i}\,kR}\sum_{n=0}^{\infty}\frac{-2\mathrm{i}k\,(n+1)f_{n+1} + [n(n+1)+D]f_n}{R^{n+2}} = 0, \tag{2.138}$$

[3] J.D. Jackson, *Classical Electrodynamics*, (Wiley, 1962), p. 104

[4] A. Sommerfeld, *Partial differential equations in physics*, (Academic Press, New York, 1964), p. 117

which gives us the recursion relation

$$2\mathrm{i}\,k(n+1)f_{n+1} = [n(n+1)+D]f_n. \tag{2.139}$$

It follows that if $f_0 = 0$ then all of the f_n are equal to zero.

Let us now consider the surface integral (2.129). Since we are interested in the limit $R \to \infty$ we can replace w by the first term of its expansion in (2.136), so

$$\int_{\Sigma_\infty} \left(w \frac{\partial w^*}{\partial n} - w^* \frac{\partial w}{\partial n} \right) dS = -2\mathrm{i}\,k \int |f_0|^2\, d\Omega = 0, \tag{2.140}$$

where $d\Omega$ is a unit of solid angle. It is clear that $f_0 = 0$. This implies that $f_1 = f_2 = \cdots = 0$ and, hence, that $w = 0$. Thus, there is only one solution of Eq. (2.126) which is consistent with the Sommerfeld radiation condition, and this is given by Eq. (2.123). We can now be sure that Eq. (2.122) is a *unique* solution of Eq. (2.103) subject to the boundary condition (2.125). This boundary condition basically says that infinity is an absorber of radiation but not an emitter, which seems entirely reasonable.

2.14 Retarded potentials

Equations (2.94) have the same form as the inhomogeneous wave equation (2.103), so we can immediately write the solutions to these equations as

$$\phi(\boldsymbol{r},t) = \frac{1}{4\pi\,\epsilon_0}\int \frac{[\rho(\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}\,dV', \tag{2.141a}$$

$$\boldsymbol{A}(\boldsymbol{r},t) = \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{j}(\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}\,dV'. \tag{2.141b}$$

Moreover, we can be sure that these solutions are *unique*, subject to the reasonable proviso that infinity is an absorber of radiation but not an emitter. This is a crucially important point. Whenever the above solutions are presented in physics textbooks there is a tacit assumption that they are unique. After all, if they were not unique why should we choose to study them instead of one of the other possible solutions? The uniqueness of the above solutions has a physical interpretation. It is clear from Eqs. (2.141) that in the absence of any charges and currents there are no electromagnetic fields. In other words, if we observe an electromagnetic field we can be certain that if we were to trace it backward in time we would eventually discover that it was emitted by a charge or a current. In proving that the solutions of Maxwell's equations are unique, and then finding a solution in which all waves are emitted by sources, we have effectively ruled out the possibility that the vacuum can be "unstable" to the production of electromagnetic waves without the need for any sources.

Equations (2.141) can be combined to form the solution of the 4-vector wave equation (2.96),

$$\Phi^\mu = \frac{1}{4\pi\,\epsilon_0\,c}\int\frac{[J^\mu]}{r}\,dV. \tag{2.142}$$

Here, the components of the 4-potential are evaluated at some event P in space-time, r is the distance of the volume element dV from P, and the square brackets indicate that the 4-current is to be evaluated at the retarded time; *i.e.*, at a time r/c before P.

But, does the right-hand side of Eq. (2.142) really transform as a contravariant 4-vector? This is not a trivial question since volume integrals in 3-space are not, in general, Lorentz invariant due to the length contraction effect. However, the integral in Eq. (2.142) is not a straightforward volume integral because the integrand is evaluated at the retarded time. In fact, the integral is best regarded as an integral over events in space-time. The events which enter the integral are those which intersect a spherical light wave launched from the event P and evolved backwards in time. In other words, the events occur before the event P and have zero interval with respect to P. It is clear that observers in all inertial frames will, at least, agree on which events are to be included in the integral, since both the interval between events and the absolute order in which events occur are invariant under a general Lorentz transformation.

We shall now demonstrate that all observers obtain the same value of dV/r for each elementary contribution to the integral. Suppose that S and S' are two inertial frames in the standard configuration. Let unprimed and primed symbols denote corresponding quantities in S and S', respectively. Let us assign coordinates $(0,0,0,0)$ to P and (x,y,z,ct) to the retarded event Q for which r

and dV are evaluated. Using the standard Lorentz transformation (2.19), the fact that the interval between events P and Q is zero, and the fact that both t and t' are negative, we obtain

$$r' = -c t' = -c\gamma \left(t - \frac{v x}{c^2}\right), \tag{2.143}$$

where v is the relative velocity between frames S' and S, γ is the Lorentz factor, and $r = \sqrt{x^2 + y^2 + z^2}$, *etc.* It follows that

$$r' = r\gamma \left(-\frac{c t}{r} + \frac{v x}{c r}\right) = r\gamma \left(1 + \frac{v}{c}\cos\theta\right), \tag{2.144}$$

where θ is the angle (in 3-space) subtended between the line PQ and the x-axis.

We now know the transformation for r. What about the transformation for dV? We might be tempted to set $dV' = \gamma\, dV$, according to the usual length contraction rule. However, this is wrong. The contraction by a factor γ only applies if the whole of the volume is measured at the same time, which is not the case in the present problem. Now, the dimensions of dV along the $y-$ and $z-$ axes are the same in both S and S', according to Eqs. (2.19). For the x-dimension these equations give $dx' = \gamma(dx - v\, dt)$. The extremities of dx are measured at times differing by dt, where[5]

$$dt = -\frac{dr}{c} = -\frac{dx}{c}\,\cos\theta. \tag{2.145}$$

Thus,

$$dx' = \left(1 + \frac{v}{c}\,\cos\theta\right)\gamma\, dx, \tag{2.146}$$

giving

$$dV' = \left(1 + \frac{v}{c}\,\cos\theta\right)\gamma\, dV. \tag{2.147}$$

It follows from Eqs. (2.144) and (2.147) that $dV'/r' = dV/r$. This result will clearly remain valid even when S and S' are not in the standard configuration.

[5]Note that $dr = dx\,\cos\theta$, despite the fact that $x = r\,\cos\theta$. This comes about because the volume element dV is aligned along a radius vector.

Thus, dV/r is an invariant and, therefore, $[J^\mu]\,dV/r$ is a contravariant 4-vector. For linear transformations, such as a general Lorentz transformation, the result of adding 4-tensors evaluated at different 4-points is itself a 4-tensor. It follows that the right-hand side of Eq. (2.142) is a contravariant 4-vector. Thus, this 4-vector equation can be properly regarded as the solution to the 4-vector wave equation (2.96).

2.15 Tensors and pseudo-tensors

The totally antisymmetric fourth rank tensor is defined

$$\epsilon^{\alpha\beta\gamma\delta} = \begin{cases} +1 & \text{for } \alpha,\beta,\gamma,\delta \text{ any even permutation of } 1,2,3,4 \\ -1 & \text{for } \alpha,\beta,\gamma,\delta \text{ any odd permutation of } 1,2,3,4 \\ 0 & \text{otherwise} \end{cases} \tag{2.148}$$

The components of this tensor are invariant under a general Lorentz transformation, since

$$\epsilon^{\alpha\beta\gamma\delta}\, p_\alpha^{\alpha'}\, p_\beta^{\beta'}\, p_\gamma^{\gamma'}\, p_\delta^{\delta'} = \epsilon^{\alpha'\beta'\gamma'\delta'}\, |p_\mu^{\mu'}| = \epsilon^{\alpha'\beta'\gamma'\delta'}, \tag{2.149}$$

where $|p_\mu^{\mu'}|$ denotes the determinant of the transformation matrix, or the Jacobian of the transformation, which we have already established is unity for a general Lorentz transformation. We can also define a totally antisymmetric third rank tensor ϵ^{ijk} which stands in the same relation to 3-space as $\epsilon^{\alpha\beta\gamma\delta}$ does to space-time. It is easily demonstrated that the elements of ϵ^{ijk} are invariant under a general translation or rotation of the coordinate axes. The totally antisymmetric third rank tensor is used to define the cross product of two 3-vectors,

$$(\boldsymbol{a}\wedge\boldsymbol{b})^i = \epsilon^{ijk}\, a_j\, b_k, \tag{2.150}$$

and the curl of a 3-vector field,

$$(\nabla\wedge\boldsymbol{A})^i = \epsilon^{ijk}\,\frac{\partial A_k}{\partial x^j}. \tag{2.151}$$

The following two rules are often useful in deriving vector identities

$$\epsilon^{ijk}\epsilon_{iab} = \delta_a^j\,\delta_b^k - \delta_b^j\,\delta_a^k, \tag{2.152a}$$

$$\epsilon^{ijk}\epsilon_{ijb} = 2\,\delta^k_b. \tag{2.152b}$$

Up to now we have restricted ourselves to three basic types of coordinate transformation; namely, translations, rotations, and standard Lorentz transformations. An arbitrary combination of these three transformations constitutes a general Lorentz transformation. Let us now extend our investigations to include a fourth type of transformation known as a parity inversion; *i.e.*, $x, y, z, \rightarrow -x, -y, -z$. A reflection is a combination of a parity inversion and a rotation. As is easily demonstrated, the Jacobian of a parity inversion is -1, unlike a translation, rotation, or standard Lorentz transformation, which all possess Jacobians of $+1$.

The prototype of all 3-vectors is the difference in coordinates between two points in space, $\boldsymbol{r}$. Likewise, the prototype of all 4-vectors is the difference in coordinates between two events in space-time, $R^\mu = (\boldsymbol{r}, ct)$. It is not difficult to appreciate that both of these objects are invariant under a parity transformation (in the sense that they correspond to the same geometric object before and after the transformation). It follows that any 3- or 4-tensor which is directly related to $\boldsymbol{r}$ and R^μ, respectively, is also invariant under a parity inversion. Such tensors include the distance between two points in 3-space, the interval between two points in space-time, 3-velocity, 3-acceleration, 4-velocity, 4-acceleration, and the metric tensor. Tensors which exhibit tensor behaviour under translations, rotations, special Lorentz transformations, *and* are invariant under parity inversions, are termed *proper tensors*, or sometimes *polar tensors*. Since electric charge is clearly invariant under such transformations (*i.e.*, it is a proper scalar) it follows that 3-current and 4-current are proper vectors. It is also clear from Eq. (2.96) that the scalar potential, the vector potential, and the potential 4-vector, are proper tensors.

It follows from Eq. (2.149) that $\epsilon^{\alpha\beta\gamma\delta} \rightarrow -\epsilon^{\alpha\beta\gamma\delta}$ under a parity inversion. Tensors like this, which exhibit tensor behaviour under translations, rotations, and special Lorentz transformations, but are *not* invariant under parity inversions (in the sense that they correspond to different geometric objects before and after the transformation), are called *pseudo-tensors*, or sometimes *axial tensors*. Equations (2.150) and (2.151) imply that the cross product of two proper vectors is a pseudo-vector, and the curl of a proper vector field is a pseudo-vector field.

One particularly simple way of performing a parity transformation is to exchange positive and negative numbers on the three Cartesian axes. A proper vector is unaffected by such a procedure (*i.e.*, its magnitude and direction are the same before and after). On the other hand, a pseudo-vector ends up pointing in the opposite direction after the axes are renumbered.

What is the fundamental difference between proper tensors and pseudo-tensors? The answer is that all pseudo-tensors are defined according to a handedness convention. For instance, the cross product between two vectors is conventionally defined according to a right-hand rule. The only reason for this is that the majority of human beings are right-handed. Presumably, if the opposite were true then cross products *etc.* would be defined according to a left-hand rule and would, therefore, take minus their conventional values. The totally antisymmetric tensor is the prototype pseudo-tensor and is, of course, conventionally defined with respect to a right-handed spatial coordinate system. A parity inversion converts left into right and *vice versa* and, thereby, effectively swaps left- and right-handed conventions.

The use of conventions in physics is perfectly acceptable provided that we recognize that they are conventions and are *consistent* in their use. It follows that laws of physics cannot contain mixtures of tensors and pseudo-tensors, otherwise they would depend our choice of handedness convention.[6]

Let us now consider electric and magnetic fields. We know that

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \tag{2.153a}$$

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A}. \tag{2.153b}$$

We have already seen that the scalar and the vector potential are proper scalars and vectors, respectively. It follows that $\boldsymbol{E}$ is a proper vector but that $\boldsymbol{B}$ is a pseudo-vector (since it is the curl of a proper vector). In order to fully appreciate the difference between electric and magnetic fields let us consider a thought

[6]Here, we are assuming that the laws of physics do not possess an intrinsic handedness. This is certainly the case for mechanics and electromagnetism. However, the weak interaction *does* possess an intrinsic handedness; *i.e.*, it is fundamentally different in a parity inverted universe. So, the equations governing the weak interaction do actually contain mixtures of tensors and pseudo-tensors.

experiment first proposed by Richard Feynman. Suppose that we are in radio contact with a race of aliens and are trying to explain to them our system of physics. Suppose, further, that the aliens live sufficiently far away from us that there are no common objects which we both can see. The question is this: could we unambiguously explain to these aliens our concepts of electric and magnetic fields? We could certainly explain electric and magnetic lines of force. The former are the paths of charged particles (assuming that the particles are subject only to electric fields) and the latter can be mapped out using small test magnets. We could also explain how we put arrows on electric lines of force to convert them into electric field lines: the arrows run from positive charges (*i.e.*, charges with the same sign as atomic nuclei) to negative charges. This explanation is unambiguous provided that our aliens live in a matter (rather than an anti-matter) dominated part of the universe. But, could we explain how we put arrows on magnetic lines of force in order to convert them into magnetic field lines? The answer is no. By definition, magnetic field lines emerge from the north poles of permanent magnets and converge on the corresponding south poles. The definition of the north pole of a magnet is simply that it possesses the same magnetic polarity as the north pole of the Earth. This is obviously a convention. In fact, we could redefine magnetic field lines to run from the south poles to the north poles of magnets without significantly altering our laws of physics (we would just have to replace $\boldsymbol{B}$ by $-\boldsymbol{B}$ in all our equations). In a parity inverted universe a north pole becomes a south pole and *vice versa*, so it is hardly surprising that $\boldsymbol{B} \rightarrow -\boldsymbol{B}$.[7]

2.16 The electromagnetic field tensor

Let us now investigate whether we can write the components of the electric and magnetic fields as the components of some *proper* 4-tensor. There is an obvious problem here. How can we identify the components of the magnetic field, which is a pseudo-vector, with any of the components of a proper-4-tensor? The former components transform differently under parity inversion than the latter compo-

[7]Note that it would actually be possible to unambiguously communicate to our concepts of left and right to our hypothetical aliens using the fact that the weak interaction possesses an intrinsic handedness.

nents. Consider a proper-3-tensor whose covariant components are written B_{ik}, and which is antisymmetric:

$$B_{ij} = -B_{ji}. \tag{2.154}$$

This immediately implies that all of the diagonal components of the tensor are zero. In fact, there are only three independent non-zero components of such a tensor. Could we, perhaps, use these components to represent the components of a pseudo-3-vector? Let us write

$$B^i = \frac{1}{2}\,\epsilon^{ijk} B_{jk}. \tag{2.155}$$

It is clear that B^i transforms as a contravariant pseudo-3-vector. It is easily seen that

$$B^{ij} = B_{ij} = \begin{pmatrix} 0 & B_z & -B_y \\ -B_z & 0 & B_x \\ B_y & -B_x & 0 \end{pmatrix}, \tag{2.156}$$

where $B^1 = B_1 \equiv B_x$, *etc.* In this manner, we can actually write the components of a pseudo-3-vector as the components of an antisymmetric proper-3-tensor. In particular, we can write the components of the magnetic field $\boldsymbol{B}$ in terms of an antisymmetric proper magnetic field 3-tensor which we shall denote B_{ij}.

Let us now examine Eqs. (2.153) more carefully. Recall that $\Phi_\mu = (-c\boldsymbol{A}, \phi)$ and $\partial_\mu = (\nabla, c^{-1}\partial/\partial t)$. It follows that we can write Eq. (2.153a) in the form

$$E_i = -\partial_i \Phi_4 + \partial_4 \Phi_i. \tag{2.157}$$

Equation (2.153b) can be written

$$cB^i = \frac{1}{2}\,\epsilon^{ijk}\,cB_{jk} = -\epsilon^{ijk}\,\partial_j \Phi_k. \tag{2.158}$$

Let us multiply this expression by ϵ_{iab}, making use of the identity

$$\epsilon_{iab}\,\epsilon^{ijk} = \delta^j_a \delta^k_b - \delta^j_b \delta^k_a. \tag{2.159}$$

We obtain

$$\frac{c}{2}\,(B_{ab} - B_{ba}) = -\partial_a \Phi_b + \partial_b \Phi_a, \tag{2.160}$$

or

$$cB_{ij} = -\partial_i \Phi_j + \partial_j \Phi_i, \tag{2.161}$$

since $B_{ij} = -B_{ji}$.

Let us define a proper-4-tensor whose covariant components are given by

$$F_{\mu\nu} = \partial_\mu \Phi_\nu - \partial_\nu \Phi_\mu. \tag{2.162}$$

It is clear that this tensor is antisymmetric:

$$F_{\mu\nu} = -F_{\nu\mu}. \tag{2.163}$$

This implies that the tensor only possesses six independent non-zero components. Maybe it can be used to specify the components of $\boldsymbol{E}$ and $\boldsymbol{B}$?

Equations (2.157) and (2.162) yield

$$F_{4i} = \partial_4 \Phi_i - \partial_i \Phi_4 = E_i. \tag{2.164}$$

Likewise, Eqs. (2.161) and (2.162) imply that

$$F_{ij} = \partial_i \Phi_j - \partial_j \Phi_i = -cB_{ij}. \tag{2.165}$$

Thus,

$$F_{i4} = -F_{4i} = -E_i, \tag{2.166a}$$

$$F_{ij} = -F_{ji} = -cB_{ij}. \tag{2.166b}$$

In other words, the completely space-like components of the tensor specify the components of the magnetic field, whereas the hybrid space and time-like components specify the components of the electric field. The covariant components of the tensor can be written

$$F_{\mu\nu} = \begin{pmatrix} 0 & -cB_z & +cB_y & -E_x \\ +cB_z & 0 & -cB_x & -E_y \\ -cB_y & +cB_x & 0 & -E_z \\ +E_x & +E_y & +E_z & 0 \end{pmatrix}. \tag{2.167}$$

Not surprisingly, $F_{\mu\nu}$ is usually called the *electromagnetic field tensor*. The above expression, which appears in all standard textbooks, is very misleading. Taken at face value, it is simply wrong! We cannot form a proper-4-tensor from the components of a proper-3-vector and a pseudo-3-vector. The expression only makes sense if we interpret B_x, say, as representing the component B_{23} of the proper magnetic field 3-tensor B_{ij}

The contravariant components of the electromagnetic field tensor are given by

$$F^{i4} = -F^{4i} = +E^i, \tag{2.168a}$$

$$F^{ij} = -F^{ji} = -cB^{ij}, \tag{2.168b}$$

or

$$F^{\mu\nu} = \left(\begin{array}{cccc} 0 & -cB_z & +cB_y & +E_x \\ +cB_z & 0 & -cB_x & +E_y \\ -cB_y & +cB_x & 0 & +E_z \\ -E_x & -E_y & -E_z & 0 \end{array} \right). \tag{2.169}$$

Let us now consider two of Maxwell's equations:

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{2.170a}$$

$$\nabla\wedge\boldsymbol{B} = \mu_0\left(\boldsymbol{j} + \epsilon_0\,\frac{\partial\boldsymbol{E}}{\partial t}\right). \tag{2.170b}$$

Recall that the 4-current is defined $J^\mu = (\boldsymbol{j}, \rho c)$. The first of these equations can be written

$$\partial_i E^i = \partial_i F^{i4} + \partial_4 F^{44} = \frac{J^4}{c\,\epsilon_0}. \tag{2.171}$$

since $F^{44} = 0$. The second of these equations takes the form

$$\epsilon^{ijk}\,\partial_j\,cB_k - \partial_4 E^i = \epsilon^{ijk}\,\partial_j(1/2\,\epsilon_{kab}\,cB^{ab}) + \partial_4 F^{4i} = \frac{J^i}{c\,\epsilon_0}. \tag{2.172}$$

Making use of Eq. (2.159), the above expression reduces to

$$\frac{1}{2}\,\partial_j(cB^{ij} - cB^{ji}) + \partial_4 F^{4i} = \partial_j F^{ji} + \partial_4 F^{4i} = \frac{J^i}{c\,\epsilon_0}. \tag{2.173}$$

Equations (2.171) and (2.173) can be combined to give

$$\partial_\mu F^{\mu\nu} = \frac{J^\nu}{c\,\epsilon_0}. \tag{2.174}$$

This equation is consistent with the equation of charge continuity, $\partial_\mu J^\mu = 0$, because of the antisymmetry of the electromagnetic field tensor.

2.17 The dual electromagnetic field tensor

We have seen that it is possible to write the components of the electric and magnetic fields as the components of a proper-4-tensor. Is it also possible to write the components of these fields as the components of some *pseudo*-4-tensor? It is obvious that we cannot identify the components of the proper-3-vector $\boldsymbol{E}$ with any of the components of a pseudo-tensor. However, we can represent the components of $\boldsymbol{E}$ in terms of those of an antisymmetric pseudo-3-tensor E_{ij} by writing

$$E^i = \frac{1}{2}\,\epsilon^{ijk} E_{jk}. \tag{2.175}$$

It is easily demonstrated that

$$E^{ij} = E_{ij} = \left(\begin{array}{ccc} 0 & E_z & -E_y \\ -E_z & 0 & E_x \\ E_y & -E_x & 0 \end{array} \right), \tag{2.176}$$

in a right-handed coordinate system.

Consider the *dual electromagnetic field tensor* $G^{\mu\nu}$, which is defined

$$G^{\mu\nu} = \frac{1}{2}\,\epsilon^{\mu\nu\alpha\beta} F_{\alpha\beta}. \tag{2.177}$$

This tensor is clearly an antisymmetric pseudo-4-tensor. We have

$$G^{4i} = \frac{1}{2}\,\epsilon^{4ijk}\,F_{jk} = -\frac{1}{2}\,\epsilon^{ijk4}\,F_{jk} = \frac{1}{2}\,\epsilon^{ijk}\,cB_{jk} = cB^i, \tag{2.178}$$

plus

$$G^{ij} = \frac{1}{2}\left(\epsilon^{ijk4}F_{k4} + \epsilon^{ij4k}F_{4k}\right) = \epsilon^{ijk}F_{k4}, \tag{2.179}$$

where use has been made of $F_{\mu\nu} = -F_{\nu\mu}$. The above expression yields

$$G^{ij} = -\epsilon^{ijk}\,E_k = -\frac{1}{2}\,\epsilon^{ijk}\epsilon_{kab}\,E^{ab} = -E^{ij}. \tag{2.180}$$

It follows that

$$G^{i4} = -G^{4i} = -cB^i, \tag{2.181a}$$

$$G^{ij} = -G^{ji} = -E^{ij}, \tag{2.181b}$$

or

$$G^{\mu\nu} = \left(\begin{array}{cccc} 0 & -E_z & +E_y & -cB_x \\ +E_z & 0 & -E_x & -cB_y \\ -E_y & +E_x & 0 & -cB_z \\ +cB_x & +cB_y & +cB_z & 0 \end{array}\right). \tag{2.182}$$

The above expression is, again, slightly misleading, since E_x stands for the component E^{23} of the pseudo-3-tensor E^{ij} and not for an element of the proper-3-vector $\boldsymbol{E}$. Of course, in this case B_x really does represent the first element of the pseudo-3-vector $\boldsymbol{B}$. Note that the elements of $G^{\mu\nu}$ are obtained from those of $F^{\mu\nu}$ by making the transformation $cB^{ij} \rightarrow E^{ij}$ and $E^i \rightarrow -cB^i$.

The covariant elements of the dual electromagnetic field tensor are given by

$$G_{i4} = -G_{4i} = +cB_i, \tag{2.183a}$$

$$G_{ij} = -G_{ji} = -E_{ij}, \tag{2.183b}$$

or

$$G_{\mu\nu} = \left(\begin{array}{cccc} 0 & -E_z & +E_y & +cB_x \\ +E_z & 0 & -E_x & +cB_y \\ -E_y & +E_x & 0 & +cB_z \\ -cB_x & -cB_y & -cB_z & 0 \end{array}\right). \tag{2.184}$$

The elements of $G_{\mu\nu}$ are obtained from those of $F_{\mu\nu}$ by making the transformation $cB_{ij} \rightarrow E_{ij}$ and $E_i \rightarrow -cB_i$.

Let us now consider the two Maxwell equations

$$\nabla\cdot\boldsymbol{B} = 0, \tag{2.185a}$$

$$\nabla\wedge\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t}. \tag{2.185b}$$

The first of these equations can be written

$$-\partial_i\, cB^i = \partial_i G^{i4} + \partial_4 G^{44} = 0, \tag{2.186}$$

since $G^{44} = 0$. The second equation takes the form

$$\epsilon^{ijk}\partial_j E_k = \epsilon^{ijk}\partial_j(1/2\,\epsilon_{kab}E^{ab}) = \partial_j E^{ij} = -\partial_4\, cB^i, \tag{2.187}$$

or

$$\partial_j G^{ji} + \partial_4 G^{4i} = 0. \tag{2.188}$$

Equations (2.186) and (2.188) can be combined to give

$$\partial_\mu G^{\mu\nu} = 0. \tag{2.189}$$

Thus, we conclude that Maxwell's equations for the electromagnetic fields are equivalent to the following pair of 4-tensor equations:

$$\partial_\mu F^{\mu\nu} = \frac{J^\nu}{c\,\epsilon_0}, \tag{2.190a}$$

$$\partial_\mu G^{\mu\nu} = 0. \tag{2.190b}$$

It is obvious from the form of these equations that the laws of electromagnetism are invariant under translations, rotations, special Lorentz transformations, parity inversions, or any combination of these transformations.

2.18 The transformation of electromagnetic fields

The electromagnetic field tensor transforms according to the standard rule

$$F^{\mu'\nu'} = F^{\mu\nu} p_\mu^{\mu'} p_\nu^{\nu'}. \tag{2.191}$$

This easily yields the celebrated rules for transforming electromagnetic fields:

$$E'_{\parallel} = E_{\parallel}, \tag{2.192a}$$

$$B'_{\parallel} = B_{\parallel}, \tag{2.192b}$$

$$\boldsymbol{E}'_{\perp} = \gamma(\boldsymbol{E}_{\perp} + \boldsymbol{v}\wedge\boldsymbol{B}), \tag{2.192c}$$

$$\boldsymbol{B}'_{\perp} = \gamma(\boldsymbol{B}_{\perp} - \boldsymbol{v}\wedge\boldsymbol{E}/c^2), \tag{2.192d}$$

where $\boldsymbol{v}$ is the relative velocity between the primed and unprimed frames, and the perpendicular and parallel directions are, respectively, perpendicular and parallel to $\boldsymbol{v}$.

At this stage we may conveniently note two important invariants of the electromagnetic field. They are

$$\frac{1}{2}\,F_{\mu\nu}F^{\mu\nu} = c^2B^2 - E^2, \tag{2.193}$$

and

$$\frac{1}{4}\,G_{\mu\nu}F^{\mu\nu} = c\,\boldsymbol{E}\cdot\boldsymbol{B}. \tag{2.194}$$

The first of these quantities is a proper-scalar and the second is a pseudo-scalar.

2.19 The potential due to a moving charge

Suppose that a particle carrying a charge e moves with *uniform* velocity $\boldsymbol{u}$ through a frame S. Let us evaluate the vector potential $\boldsymbol{A}$ and the scalar potential ϕ due to this charge at a given event P in S.

Let us choose coordinates in S so that $P = (0,0,0,0)$ and $\boldsymbol{u} = (u,0,0)$. Let S' be that frame in the standard configuration with respect to S in which the charge is (permanently) at rest, say at the point (x',y',z'). In S' the potential at P is the usual potential due to a stationary charge

$$\boldsymbol{A}' = 0, \tag{2.195a}$$

$$\phi' = \frac{e}{4\pi\epsilon_0\,r'}, \tag{2.195b}$$

where $r' = \sqrt{x'^2 + y'^2 + z'^2}$. Let us now transform these equations directly into the frame S. Since $A^\mu = (c\boldsymbol{A}, \phi)$ is a contravariant 4-vector, its components transform according to the standard rules (2.57). Thus,

$$
\begin{aligned}
cA_1 &= \gamma\left(cA_1' + \frac{u}{c}\,\phi'\right) = \frac{\gamma u\, e}{4\pi\epsilon_0\, c\, r'}, &\text{(2.196a)}\\
cA_2 &= cA_2' = 0, &\text{(2.196b)}\\
cA_3 &= cA_3' = 0, &\text{(2.196c)}\\
\phi &= \gamma\left(\phi' + \frac{u}{c}\, cA_1'\right) = \frac{\gamma\, e}{4\pi\epsilon_0\, r'}, &\text{(2.196d)}
\end{aligned}
$$

since $\beta = -u/c$ in this case. It remains to express the quantity r' in terms of quantities measured in S. The most physically meaningful way of doing this is to express r' in terms of *retarded* values in S. Consider the retarded event at the charge for which, by definition, $r' = -ct'$ and $r = -ct$. Using the standard Lorentz transformation (2.19) we find that

$$r' = -ct' = -c\gamma(t - ux/c^2) = r\gamma(1 + u_r/c), \tag{2.197}$$

where $u_r = ux/r = \boldsymbol{r}\cdot\boldsymbol{u}/r$ denotes the radial velocity of the change in S. We can now rewrite Eqs. (2.196) in the form

$$
\begin{aligned}
\boldsymbol{A} &= \frac{\mu_0\, e}{4\pi}\frac{[\boldsymbol{u}]}{[r + \boldsymbol{r}\cdot\boldsymbol{u}/c]}, &\text{(2.198a)}\\
\phi &= \frac{e}{4\pi\epsilon_0}\frac{1}{[r + \boldsymbol{r}\cdot\boldsymbol{u}/c]}, &\text{(2.198b)}
\end{aligned}
$$

where the square brackets, as usual, indicate that the enclosed quantities must be retarded. For a uniformly moving charge the retardation of $\boldsymbol{u}$ is, of course, superfluous. However, since

$$\Phi^\mu = \frac{1}{4\pi\epsilon_0\, c}\int\frac{[J^\mu]}{r}\, dV, \tag{2.199}$$

it is clear that the potentials depend only on the (retarded) velocity of the charge and not on its acceleration. Consequently, the expressions (2.198) give the correct potentials for an *arbitrarily* moving charge. They are known as the *Liénard-Wiechert potentials.*

2.20 The electromagnetic field due to a uniformly moving charge

Although the field generated by a uniformly moving charge can be calculated from the expressions (2.198) for the potentials, it is simpler to calculate it relativistically from first principles.

Let a charge e, whose position vector at time $t = 0$ is $\boldsymbol{r}$, move with uniform velocity $\boldsymbol{u}$ in a frame S whose x-axis has been chosen in the direction of $\boldsymbol{u}$. We require to find the field strengths $\boldsymbol{E}$ and $\boldsymbol{B}$ at the event $P = (0, 0, 0, 0)$. Let S' be that frame in standard configuration with S in which the charge is permanently at rest. In S' the field is given by

$$\boldsymbol{B}' = 0, \tag{2.200a}$$

$$\boldsymbol{E}' = -\frac{e}{4\pi\epsilon_0}\frac{\boldsymbol{r}'}{r'^3}. \tag{2.200b}$$

This field must now be transformed into the frame S. The direct method, using Eqs. (2.192), is somewhat simpler here, but we shall use a somewhat indirect method because of its intrinsic interest.

In order to express Eq. (2.200) in tensor form, we need the electromagnetic field tensor $F^{\mu\nu}$ on the left, and the position 4-vector $R^\mu = (\boldsymbol{r}, ct)$ and the scalar $e/(4\pi\epsilon_0\, r'^3)$ on the right. (We regard r' as an invariant for all observers.) To get a vanishing magnetic field in S' we multiply on the right by the 4-velocity $U^\mu = \gamma(u)(\boldsymbol{u}, c)$, thus tentatively arriving at the equation

$$F^{\mu\nu} = \frac{e}{4\pi\epsilon_0\, c\, r'^3}\, U^\mu R^\nu. \tag{2.201}$$

Recall that $F^{4i} = -E^i$ and $F^{ij} = -cB^{ij}$. This equation cannot be correct, because the antisymmetric tensor $F^{\mu\nu}$ can only be equated to another antisymmetric tensor. Consequently, we try the equation

$$F^{\mu\nu} = \frac{e}{4\pi\epsilon_0\, c\, r'^3}\,(U^\mu R^\nu - U^\nu R^\mu). \tag{2.202}$$

This is found to give the correct field at P in S' as long as R^μ refers to any event at the charge, no matter which. It only remains to interpret (2.202) in S. It is convenient to choose for R^μ that event at the charge at which $t = 0$ (not the retarded event). Thus,

$$F^{jk} = -cB^{jk} = \frac{e}{4\pi\epsilon_0\, c\, r'^3}\,\gamma(u)\,(u^j r^k - u^k r^j), \tag{2.203}$$

giving

$$B_i = \frac{1}{2}\,\epsilon_{ijk} B^{jk} = -\frac{\mu_0\, e}{4\pi\, r'^3}\,\gamma(u)\,\epsilon_{ijk} u^j r^k, \tag{2.204}$$

or

$$\boldsymbol{B} = -\frac{\mu_0\, e\,\gamma}{4\pi\, r'^3}\,\boldsymbol{u}\wedge\boldsymbol{r}. \tag{2.205}$$

Likewise,

$$F^{4i} = -E^i = \frac{e\,\gamma}{4\pi\epsilon_0\, r'^3}\, r^i, \tag{2.206}$$

or

$$\boldsymbol{E} = -\frac{e\,\gamma}{4\pi\epsilon_0\, r'^3}\,\boldsymbol{r}. \tag{2.207}$$

Lastly, we must find an expression for r'^3 in terms of quantities measured in S at time $t = 0$. If t' is the corresponding time in S' at the charge, we have

$$r'^2 = r^2 + c^2 t'^2 = r^2 + \frac{\gamma^2\, u^2 x^2}{c^2} = r^2\left(1 + \frac{\gamma^2 {u_r}^2}{c^2}\right). \tag{2.208}$$

Thus,

$$\boldsymbol{E} \;=\; -\frac{e}{4\pi\epsilon_0}\,\frac{\gamma}{r^3(1 + {u_r}^2\,\gamma^2/c^2)^{3/2}}\,\boldsymbol{r}, \tag{2.209a}$$

$$\boldsymbol{B} \;=\; -\frac{\mu_0\, e}{4\pi}\,\frac{\gamma}{r^3(1 + {u_r}^2\,\gamma^2/c^2)^{3/2}}\,\boldsymbol{u}\wedge\boldsymbol{r} = \frac{1}{c^2}\,\boldsymbol{u}\wedge\boldsymbol{E}. \tag{2.209b}$$

Note that $\boldsymbol{E}$ acts in line with the point which the charge occupies *at the instant of measurement* despite the fact that, owing to the finite speed of propagation of all physical effects, the behaviour of the charge during a finite period before that instant can no longer affect the measurement. Note also that, unlike Eqs. (2.198), the above expressions for the fields are not valid for an arbitrarily moving charge, not can they be made valid by merely using retarded values. For whereas acceleration does not affect the potentials, it does affect the fields, which involve the derivatives of the potential.

For low velocities, $u/c \to 0$, Eqs. (2.209) reduce to the well known Coulomb and Biot-Savart fields. However, at high velocities, $\gamma(u) \gg 1$, the fields exhibit some interesting behaviour. The peak electric field, which occurs at the point of closest approach of the charge to the observation point, becomes equal to γ times its non-relativistic value. However, the duration of appreciable field strength at the point P is decreased. A measure of the time interval over which the field is appreciable is

$$\Delta t \sim \frac{b}{\gamma c}, \tag{2.210}$$

where b is the distance of closest approach (assuming $\gamma \gg 1$). As γ increases, the peak field increases in proportion, but its duration goes in the inverse proportion. The time integral of the field is independent of γ. As $\gamma \to \infty$ the observer at P sees electric and magnetic fields which are indistinguishable from the fields of a pulse of plane polarized radiation propagating in the x-direction. The direction of polarization is along the radius vector pointing towards the particle's actual position at the time of observation.

2.21 Relativistic particle dynamics

Consider a particle which, in its instantaneous rest frame S_0, has mass m_0 and constant acceleration in the x-direction a_0. Let us transform to a frame S, in the standard configuration with respect to S_0, in which the particle's instantaneous velocity is u. What is the value of a, the particle's instantaneous x-acceleration, in S?

The easiest way in which to answer this question is to consider the acceleration 4-vector (see Eq. (2.85))

$$A^\mu = \gamma\left(\frac{d\gamma}{dt}\,\boldsymbol{u} + \gamma\boldsymbol{a}, c\,\frac{d\gamma}{dt}\right). \tag{2.211}$$

Using the standard transformation (2.57) for 4-vectors, we obtain

$$\frac{d\gamma}{dt}\,u + \gamma a = a_0, \tag{2.212a}$$

$$\frac{d\gamma}{dt} = \frac{u\,a_0}{c^2}. \tag{2.212b}$$

It follows that

$$a = \frac{a_0}{\gamma^3}. \tag{2.213}$$

The above equation can be written

$$f = m_0\gamma^3\,\frac{du}{dt}, \tag{2.214}$$

where $f = m_0 a_0$ is the constant force (in the x-direction) acting on the particle in S_0.

Equation (2.214) is equivalent to

$$f = \frac{d(mu)}{dt}, \tag{2.215}$$

where

$$m = \gamma m_0. \tag{2.216}$$

Thus, we can account for the ever decreasing acceleration of a particle subject to a constant force (see Eq. (2.213)) by supposing that the inertial mass of the particle increases with its velocity according to the rule (2.216). Henceforth, m_0 is termed the *rest mass*, and m the *inertial mass*.

The rate of increase of the particle's energy E satisfies

$$\frac{dE}{dt} = f u = m_0\gamma^3\,u\,\frac{du}{dt}. \tag{2.217}$$

This equation can be written

$$\frac{dE}{dt} = \frac{d(mc^2)}{dt}, \tag{2.218}$$

which can be integrated to yield Einstein's famous formula

$$E = mc^2. \tag{2.219}$$

The 3-momentum of a particle is defined

$$\boldsymbol{p} = m\boldsymbol{u}, \tag{2.220}$$

where $\boldsymbol{u}$ is its 3-velocity. Thus, by analogy with Eq. (2.215), Newton's law of motion can be written

$$\boldsymbol{f} = \frac{d\boldsymbol{p}}{dt}, \tag{2.221}$$

where $\boldsymbol{f}$ is the 3-force acting on the particle.

The 4-momentum of a particle is defined

$$P^\mu = m_0 U^\mu = \gamma m_0(\boldsymbol{u}, c) = (\boldsymbol{p}, E/c), \tag{2.222}$$

where U^μ is its 4-velocity. The 4-force acting on the particle obeys

$$\mathcal{F}^\mu = \frac{dP^\mu}{d\tau} = m_0 A^\mu, \tag{2.223}$$

where A^μ is its 4-acceleration. It is easily demonstrated that

$$\mathcal{F}^\mu = \gamma\left(\boldsymbol{f}, c\,\frac{dm}{dt}\right) = \gamma\left(\boldsymbol{f}, \frac{\boldsymbol{f}\cdot\boldsymbol{u}}{c}\right), \tag{2.224}$$

since

$$\frac{dE}{dt} = \boldsymbol{f}\cdot\boldsymbol{u}. \tag{2.225}$$

2.22 The force on a moving charge

The electromagnetic 3-force acting on a charge e moving with 3-velocity $\boldsymbol{u}$ is given by the well known formula

$$\boldsymbol{f} = e(\boldsymbol{E} + \boldsymbol{u}\wedge\boldsymbol{B}). \tag{2.226}$$

When written in component form this expression becomes

$$f_i = e(E_i + \epsilon_{ijk}\, u^j\, B^k), \tag{2.227}$$

or

$$f_i = e(E_i + B_{ij}\, u^j), \tag{2.228}$$

where use has been made of Eq. (2.155).

Recall that the components of the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields can be written in terms of an antisymmetric electromagnetic field tensor $F_{\mu\nu}$ via

$$F_{i4} = -F_{4i} = -E_i, \tag{2.229a}$$

$$F_{ij} = -F_{ji} = -cB_{ij}. \tag{2.229b}$$

Equation (2.228) can be written

$$f_i = -\frac{e}{\gamma\, c}\,(F_{i4}\, U^4 + F_{ij}\, U^j), \tag{2.230}$$

where $U^\mu = \gamma(\boldsymbol{u}, c)$ is the particle's 4-velocity. It is easily demonstrated that

$$\frac{\boldsymbol{f}\cdot\boldsymbol{u}}{c} = \frac{e}{c}\,\boldsymbol{E}\cdot\boldsymbol{u} = \frac{e}{c}\,E_i\, u^i = \frac{e}{\gamma\, c}(F_{4i}\, U^i + F_{44}\, U^4). \tag{2.231}$$

Thus, the 4-force acting on the particle,

$$\mathcal{F}_\mu = \gamma\left(-\boldsymbol{f}, \frac{\boldsymbol{f}\cdot\boldsymbol{u}}{c}\right), \tag{2.232}$$

can be written in the form

$$\mathcal{F}_\mu = \frac{e}{c}\, F_{\mu\nu}\, U^\nu. \tag{2.233}$$

The skew symmetry of the electromagnetic field tensor ensures that

$$\mathcal{F}_\mu U^\mu = \frac{e}{c}\, F_{\mu\nu}\, U^\mu U^\nu = 0. \tag{2.234}$$

This is an important result since it ensures that electromagnetic fields do not change the rest mass of charged particles. In order to appreciate this, let us assume that the rest mass m_0 is not a constant. Since

$$\mathcal{F}_\mu = \frac{d(m_0 U_\mu)}{d\tau} = m_0\, A_\mu + \frac{dm_0}{d\tau}\, U_\mu, \tag{2.235}$$

we can use the standard results $U_\mu U^\mu = c^2$ and $A_\mu U^\mu = 0$ to give

$$\mathcal{F}_\mu U^\mu = c^2\, \frac{dm_0}{d\tau}. \tag{2.236}$$

Thus, if rest mass is to remain an invariant it is imperative that all laws of physics predict 4-forces acting on particles which are orthogonal to the particles' 4-velocities. The laws of electromagnetism pass this test.

2.23 The electromagnetic energy tensor

Consider a continuous volume distribution of charged matter in the presence of an electromagnetic field. Let there be n_0 particles per unit proper volume (unit volume determined in the local rest frame), each carrying a charge e. Consider an inertial frame in which the 3-velocity field of the particles is $\boldsymbol{u}$. The number density of the particles in this frame is $n = \gamma(u)\, n_0$. The charge density and the 3-current due to the particles are $\rho = en$ and $\boldsymbol{j} = en\, \boldsymbol{u}$, respectively. Multiplying Eq. (2.233) by the proper number density of particles n_0, we obtain an expression

$$f_\mu = \frac{1}{c} F_{\mu\nu} J^\nu \tag{2.237}$$

for the 4-force f_μ acting on unit proper volume of the distribution due to the ambient electromagnetic fields. Here, we have made use of the definition $J^\mu = e n_0 U^\mu$. It is easily demonstrated, using some of the results obtained in the previous section, that

$$f^\mu = \left(\rho \boldsymbol{E} + \boldsymbol{j} \wedge \boldsymbol{B}, \frac{\boldsymbol{E}\cdot\boldsymbol{j}}{c}\right). \tag{2.238}$$

The above expression remains valid when there are many charge species (*e.g.*, electrons and ions) possessing different number density and 3-velocity fields. The 4-vector f^μ is usually called the *Lorentz force density*.

We know that Maxwell's equations reduce to

$$\partial_\mu F^{\mu\nu} = \frac{J^\nu}{c\,\epsilon_0}, \tag{2.239a}$$

$$\partial_\mu G^{\mu\nu} = 0, \tag{2.239b}$$

where $F^{\mu\nu}$ is the electromagnetic field tensor and $G^{\mu\nu}$ is its dual. As is easily verified, Eq. (2.239b) can also be written in the form

$$\partial_\mu F_{\nu\sigma} + \partial_\nu F_{\sigma\mu} + \partial_\sigma F_{\mu\nu} = 0. \tag{2.240}$$

Equations (2.237) and (2.239a) can be combined to give

$$f_\nu = \epsilon_0\, F_{\nu\sigma}\, \partial_\mu F^{\mu\sigma}. \tag{2.241}$$

This expression can also be written

$$f_\nu = \epsilon_0(\partial_\mu(F^{\mu\sigma} F_{\nu\sigma}) - F^{\mu\sigma}\, \partial_\mu F_{\nu\sigma}). \tag{2.242}$$

Now,

$$F^{\mu\sigma}\, \partial_\mu F_{\nu\sigma} = \frac{1}{2} F^{\mu\sigma}(\partial_\mu F_{\nu\sigma} + \partial_\sigma F_{\mu\nu}), \tag{2.243}$$

where use has been made of the antisymmetry of the electromagnetic field tensor. It follows from Eq. (2.240) that

$$F^{\mu\sigma}\, \partial_\mu F_{\nu\sigma} = -\frac{1}{2} F^{\mu\sigma}\, \partial_\nu F_{\sigma\mu} = \frac{1}{4}\partial_\nu (F^{\mu\sigma} F_{\mu\sigma}). \tag{2.244}$$

Thus,

$$f_\nu = \epsilon_0 \left(\partial_\mu(F^{\mu\sigma} F_{\nu\sigma}) - \frac{1}{4}\partial_\nu (F^{\mu\sigma} F_{\mu\sigma}) \right). \tag{2.245}$$

The above expression can also be written

$$f_\nu = -\partial_\mu T^\mu{}_\nu, \tag{2.246}$$

where

$$T^\mu{}_\nu = \epsilon_0 \left(F^{\mu\sigma} F_{\sigma\nu} + \frac{1}{4}\delta^\mu_\nu (F^{\rho\sigma} F_{\rho\sigma}) \right) \tag{2.247}$$

is called the *electromagnetic energy tensor*. Note that $T^\mu{}_\nu$ is a proper-4-tensor. It follows from Eqs. (2.167), (2.169), and (2.193) that

$$T^i{}_j = \epsilon_0 E^i E_j + \frac{B^i B_j}{\mu_0} - \delta^i_j \,\frac{1}{2}\left(\epsilon_0 E^k E_k + \frac{B^k B_k}{\mu_0}\right), \tag{2.248a}$$

$$T^i{}_4 = -T^4{}_i = \frac{\epsilon^{ijk} E_j B_k}{\mu_0\, c}, \tag{2.248b}$$

$$T^4_4 = \frac{1}{2}\left(\epsilon_0 E^k E_k + \frac{B^k B_k}{\mu_0}\right). \tag{2.248c}$$

Equation (2.246) can also be written

$$f^\nu = -\partial_\mu T^{\mu\nu}, \tag{2.249}$$

where $T^{\mu\nu}$ is a symmetric tensor whose elements are

$$T^{ij} = -\epsilon_0 E^i E^j - \frac{B^i B^j}{\mu_0} + \delta^{ij}\,\frac{1}{2}\left(\epsilon_0 E^2 + \frac{B^2}{\mu_0}\right), \tag{2.250a}$$

$$T^{i4} = T^{4i} = \frac{(\boldsymbol{E}\wedge\boldsymbol{B})^i}{\mu_0\, c}, \tag{2.250b}$$

$$T^{44} = \frac{1}{2}\left(\epsilon_0 E^2 + \frac{B^2}{\mu_0}\right). \tag{2.250c}$$

Consider the time-like component of Eq. (2.249). It follows from Eq. (2.238) that

$$\frac{\boldsymbol{E}\cdot\boldsymbol{j}}{c} = -\partial_i T^{i4} - \partial_4 T^{44}. \tag{2.251}$$

This equation can be rearranged to give

$$\frac{\partial W}{\partial t} + \nabla\cdot\boldsymbol{\epsilon} = -\boldsymbol{E}\cdot\boldsymbol{j}, \tag{2.252}$$

where $W = T^{44}$ and $\epsilon^i = c T^{i4}$, so that

$$W = \frac{\epsilon_0 E^2}{2} + \frac{B^2}{2\mu_0}, \tag{2.253}$$

and

$$\boldsymbol{\epsilon} = \frac{\boldsymbol{E} \wedge \boldsymbol{B}}{\mu_0}. \tag{2.254}$$

The right-hand side of Eq. (2.252) represents the rate per unit volume at which energy is transferred from the electromagnetic field to charged particles. It is clear, therefore, that Eq. (2.252) is an energy conservation equation for the electromagnetic field. The proper-3-scalar W can be identified as the energy density of the electromagnetic field, whereas the proper-3-vector $\boldsymbol{\epsilon}$ is the energy flux due to the electromagnetic field. The latter quantity is called the *Poynting vector*.

Consider the space-like components of Eq. (2.249). It is easily demonstrated that these reduce to

$$\frac{\partial \boldsymbol{g}}{\partial t} + \nabla \cdot \boldsymbol{P} = -\rho \boldsymbol{E} - \boldsymbol{j} \wedge \boldsymbol{B}, \tag{2.255}$$

where $P^{ij} = T^{ij}$ and $g^i = T^{4i}/c$, or

$$P^{ij} = -\epsilon_0 E^i E^j - \frac{B^i B^j}{\mu_0} + \delta^{ij}\,\frac{1}{2}\left(\epsilon_0 E^2 + \frac{B^2}{\mu_0}\right), \tag{2.256}$$

and

$$\boldsymbol{g} = \frac{\boldsymbol{\epsilon}}{c^2} = \epsilon_0 \boldsymbol{E} \wedge \boldsymbol{B}. \tag{2.257}$$

Equation (2.255) is basically a momentum conservation equation for the electromagnetic field. The right-hand side represents the rate per unit volume at which momentum is transferred from the electromagnetic field to charged particles. The symmetric proper-3-tensor P^{ij} is called the *Maxwell stress tensor*. The element P^{ij} gives the flux of electromagnetic momentum parallel to the ith axis crossing a surface normal to the jth axis. The proper-3-vector $\boldsymbol{g}$ represents the momentum density of the electromagnetic field. It is clear that the energy conservation law (2.252) and the momentum conservation law (2.255) can be combined together to give the relativistically invariant energy-momentum conservation law (2.249).

2.24 The electromagnetic field due to an accelerated charge

Let us calculate the electric and magnetic fields observed at position x^i and time t due to a charge e whose *retarded* position and time are $x^{i'}$ and t', respectively.

From now on (x^i, t) is termed the *field point* and $(x^{i'}, t')$ is termed the *source point*. It is assumed that we are given the retarded position of the charge as a function of its retarded time; *i.e.*, $x^{i'}(t')$. The retarded velocity and acceleration of the charge are

$$u^i = \frac{dx^{i'}}{dt'}, \tag{2.258}$$

and

$$\dot{u}^i = \frac{du^i}{dt'}, \tag{2.259}$$

respectively. The radius vector $\boldsymbol{r}$ is defined to extend *from* the retarded position of the charge *to* the field point, so that $r^i = x^i - x^{i'}$. (Note that this is the *opposite* convention to that adopted in Sections 2.19 and 2.20). It follows that

$$\frac{d\boldsymbol{r}}{dt'} = -\boldsymbol{u}. \tag{2.260}$$

The field and the source point variables are connected by the retardation condition

$$r(x^i, x^{i'}) = \left[(x^i - x^{i'})(x_i - x_i{}')\right]^{1/2} = c(t - t'). \tag{2.261}$$

The potentials generated by the charge are given by the Liénard-Wiechert formulae

$$\boldsymbol{A}(x^i, t) = \frac{\mu_0\, e}{4\pi} \frac{\boldsymbol{u}}{s}, \tag{2.262a}$$

$$\phi(x^i, t) = \frac{e}{4\pi\epsilon_0} \frac{1}{s}, \tag{2.262b}$$

where $s = r - \boldsymbol{r}\cdot\boldsymbol{u}/c$ is a function both of the field point and the source point variables. Recall that the Liénard-Wiechert potentials are valid for accelerating as well as uniformly moving charges.

The fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are derived from the potentials in the usual manner

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \tag{2.263a}$$

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A}. \tag{2.263b}$$

However, the components of the gradient operator ∇ are partial derivatives at constant time t, and *not* at constant time t'. Partial differentiation with respect to the x^i compares the potentials at neighbouring points at the same time, but these potential signals originate from the charge at different retarded times. Similarly, the partial derivative with respect to t implies constant x^i, and, hence, refers to the comparison of the potentials at a given field point over an interval of time during which the retarded coordinates of the source have changed. Since we only know the time variation of the particle's retarded position with respect to t' we must transform $\partial/\partial t|_{x^i}$ and $\partial/\partial x^i|_t$ to expressions involving $\partial/\partial t'|_{x^i}$ and $\partial/\partial x^i|_{t'}$.

Now, since $x^{i'}$ is assumed to be given as a function of t', we have

$$r(x^i, x^{i'}(t')\,) \equiv r(x^i, t') = c(t - t'), \tag{2.264}$$

which is a functional relationship between x^i, t, and t'. Note that

$$\left(\frac{\partial r}{\partial t'}\right)_{x^i} = -\frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r}. \tag{2.265}$$

It follows that

$$\frac{\partial r}{\partial t} = c\left(1 - \frac{\partial t'}{\partial t}\right) = \frac{\partial r}{\partial t'}\frac{\partial t'}{\partial t} = -\frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r}\,\frac{\partial t'}{\partial t}, \tag{2.266}$$

where all differentiation is at constant x^i. Thus,

$$\frac{\partial t'}{\partial t} = \frac{1}{1 - \boldsymbol{r}\cdot\boldsymbol{u}/rc} = \frac{r}{s}, \tag{2.267}$$

giving

$$\frac{\partial}{\partial t} = \frac{r}{s}\frac{\partial}{\partial t'}. \tag{2.268}$$

Similarly,

$$\nabla r = -c\nabla t' = \nabla' r + \frac{\partial r}{\partial t'}\,\nabla t' = \frac{\boldsymbol{r}}{r} - \frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r}\,\nabla t', \tag{2.269}$$

where ∇' denotes differentiation with respect to x^i at constant t'. It follows that

$$\nabla t' = -\frac{\boldsymbol{r}}{sc}, \tag{2.270}$$

so that

$$\nabla = \nabla' - \frac{\boldsymbol{r}}{s c}\frac{\partial}{\partial t'}. \tag{2.271}$$

Equation (2.263a) yields

$$\frac{4\pi\epsilon_0}{e}\,\boldsymbol{E} = \frac{\nabla s}{s^2} - \frac{\partial}{\partial t}\frac{\boldsymbol{u}}{s c^2}, \tag{2.272}$$

or

$$\frac{4\pi\epsilon_0}{e}\,\boldsymbol{E} = \frac{\nabla' s}{s^2} - \frac{\boldsymbol{r}}{s^3 c}\frac{\partial s}{\partial t'} - \frac{r}{s^2 c^2}\,\dot{\boldsymbol{u}} + \frac{r\,\boldsymbol{u}}{s^3 c^2}\frac{\partial s}{\partial t'}. \tag{2.273}$$

However,

$$\nabla' s = \frac{\boldsymbol{r}}{r} - \frac{\boldsymbol{u}}{c}, \tag{2.274}$$

and

$$\frac{\partial s}{\partial t'} = \frac{\partial r}{\partial t'} - \frac{\boldsymbol{r}\cdot\dot{\boldsymbol{u}}}{c} + \frac{\boldsymbol{u}\cdot\boldsymbol{u}}{c} = -\frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r} - \frac{\boldsymbol{r}\cdot\dot{\boldsymbol{u}}}{c} + \frac{u^2}{c}. \tag{2.275}$$

Thus,

$$\frac{4\pi\epsilon_0}{e}\,\boldsymbol{E} = \frac{1}{s^2 r}\left(\boldsymbol{r} - \frac{r\boldsymbol{u}}{c}\right) + \frac{1}{s^3 c}\left(\boldsymbol{r} - \frac{r\,\boldsymbol{u}}{c}\right)\left(\frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r} - \frac{u^2}{c} + \frac{\boldsymbol{r}\cdot\dot{\boldsymbol{u}}}{c}\right) - \frac{r}{s^2 c^2}\,\dot{\boldsymbol{u}}, \tag{2.276}$$

which reduces to

$$\frac{4\pi\epsilon_0}{e}\,\boldsymbol{E} = \frac{1}{s^3}\left(\boldsymbol{r} - \frac{r\,\boldsymbol{u}}{c}\right)\left(1 - \frac{u^2}{c^2}\right) + \frac{1}{s^3 c^2}\left(\boldsymbol{r}\wedge\left[\left(\boldsymbol{r} - \frac{r\,\boldsymbol{u}}{c}\right)\wedge\dot{\boldsymbol{u}}\right]\right). \tag{2.277}$$

Similarly,

$$\frac{4\pi}{\mu_0\, e}\,\boldsymbol{B} = \nabla\wedge\frac{\boldsymbol{u}}{s} = -\frac{\nabla' s\wedge\boldsymbol{u}}{s^2} - \frac{\boldsymbol{r}}{s c}\wedge\left(\frac{\dot{\boldsymbol{u}}}{s} - \frac{u}{s^2}\frac{\partial s}{\partial t'}\right), \tag{2.278}$$

or

$$\frac{4\pi}{\mu_0\, e}\,\boldsymbol{B} = -\frac{\boldsymbol{r}\wedge\boldsymbol{u}}{s^2 r} - \frac{\boldsymbol{r}}{s c}\wedge\left[\frac{\dot{\boldsymbol{u}}}{s} + \frac{\boldsymbol{u}}{s^2}\left(\frac{\boldsymbol{r}\cdot\boldsymbol{u}}{r} + \frac{\boldsymbol{r}\cdot\dot{\boldsymbol{u}}}{c} - \frac{u^2}{c}\right)\right], \tag{2.279}$$

which reduces to

$$\frac{4\pi}{\mu_0\, e}\,\boldsymbol{B} = \frac{\boldsymbol{u}\wedge\boldsymbol{r}}{s^3}\left(1 - \frac{u^2}{c^2}\right) + \frac{1}{s^3 c}\frac{\boldsymbol{r}}{r}\wedge\left(\boldsymbol{r}\wedge\left[\left(\boldsymbol{r} - \frac{r\,\boldsymbol{u}}{c}\right)\wedge\dot{\boldsymbol{u}}\right]\right). \tag{2.280}$$

A comparison of Eqs. (2.277) and (2.280) yields

$$\boldsymbol{B} = \frac{\boldsymbol{r}\wedge\boldsymbol{E}}{r\,c}. \tag{2.281}$$

Thus, the magnetic field is always perpendicular to $\boldsymbol{E}$ and the *retarded* radius vector $\boldsymbol{r}$. Note that all terms appearing in the above formulae are retarded.

The electric field is composed of two separate parts. The first term in Eq. (2.277) varies as $1/r^2$ for large distances from the charge. We can think of $\boldsymbol{r}_u = \boldsymbol{r} - r\,\boldsymbol{u}/c$ as the *virtual present radius vector*; *i.e.*, the radius vector directed from the position the charge would occupy at time t if it had continued with uniform velocity from its retarded position to the field point. In terms of $\boldsymbol{r}_u$ the $1/r^2$ field is simply

$$\boldsymbol{E}_{\rm induction} = \frac{e}{4\pi\epsilon_0}\frac{1-u^2/c^2}{s^3}\,\boldsymbol{r}_u. \tag{2.282}$$

We can rewrite the expression (2.209a) for the electric field generated by a *uniformly* moving charge in the form

$$\boldsymbol{E} = \frac{e}{4\pi\epsilon_0}\frac{1-u^2/c^2}{{r_0}^3(1-u^2/c^2+{u_r}^2/c^2)^{3/2}}\,\boldsymbol{r}_0, \tag{2.283}$$

where $\boldsymbol{r}_0$ is the radius vector directed from the *present* position of the charge at time t to the field point, and $u_r = \boldsymbol{u}\cdot\boldsymbol{r}_0/r_0$. For the case of uniform motion the relationship between the retarded radius vector $\boldsymbol{r}$ and the actual radius vector $\boldsymbol{r}_0$ is simply

$$\boldsymbol{r}_0 = \boldsymbol{r} - \frac{r}{c}\,\boldsymbol{u}. \tag{2.284}$$

It is straightforward to demonstrate that

$$s = r_0\sqrt{1-u^2/c^2+{u_r}^2/c^2} \tag{2.285}$$

in this case. Thus, the electric field generated by a uniformly moving charge can be written

$$\boldsymbol{E} = \frac{e}{4\pi\epsilon_0}\frac{1-u^2/c^2}{s^3}\,\boldsymbol{r}_0. \tag{2.286}$$

Since $\boldsymbol{r}_u = \boldsymbol{r}_0$ for the case of a uniformly moving charge, it is clear that Eq. (2.282) is equivalent to the electric field generated by a uniformly moving charge located at the position the charge would occupy if it had continued with uniform velocity from its retarded position.

The second term in Eq. (2.277),

$$\boldsymbol{E}_{\rm radiation} = \frac{e}{4\pi\epsilon_0 c^2}\frac{\boldsymbol{r}\wedge(\boldsymbol{r}_u\wedge\dot{\boldsymbol{u}})}{s^3}, \tag{2.287}$$

is of order $1/r$ and, therefore, represents a radiation field in the sense of contributing to the energy flux over a large sphere. Similar considerations hold for the two terms of Eq. (2.280).

2.25 The Lamor formula

Let us transform to the inertial frame in which the charge is instantaneously at rest at the origin at time $t = 0$. In this frame $u \ll c$, so that $\boldsymbol{r}_u \simeq \boldsymbol{r}$ and $s \simeq r$, for events which are sufficiently close to the origin at $t = 0$ that the retarded charge still appears to travel with a velocity which is small compared to that of light. It follows from the previous section that

$$\boldsymbol{E}_{\rm rad} \simeq \frac{e}{4\pi\epsilon_0 c^2}\frac{\boldsymbol{r}\wedge(\boldsymbol{r}\wedge\dot{\boldsymbol{u}})}{r^3}, \tag{2.288a}$$

$$\boldsymbol{B}_{\rm rad} \simeq \frac{e}{4\pi\epsilon_0 c^3}\frac{\dot{\boldsymbol{u}}\wedge\boldsymbol{r}}{r^2}. \tag{2.288b}$$

Let us define spherical polar coordinates whose axis points along the direction of instantaneous acceleration of the charge. It is easily demonstrated that

$$E_\theta \simeq \frac{e}{4\pi\epsilon_0 c^2}\frac{\sin\theta}{r}\,\dot{u}, \tag{2.289a}$$

$$B_\phi \simeq \frac{e}{4\pi\epsilon_0 c^3}\frac{\sin\theta}{r}\,\dot{u}. \tag{2.289b}$$

These fields are identical to those of a radiating dipole whose axis is aligned along the direction of instantaneous acceleration. The Poynting flux is given by

$$\epsilon_r = \frac{E_\theta B_\phi}{\mu_0} = \frac{e^2}{16\pi^2\epsilon_0 c^3}\frac{\sin^2\theta}{r^2}\,\dot{u}^2. \tag{2.290}$$

We can integrate this expression to obtain the instantaneous power radiated by the charge

$$P = \frac{e^2}{6\pi\epsilon_0 c^3}\,\dot{u}^2. \tag{2.291}$$

This is known as *Lamor's formula*. Note that zero net momentum is carried off by the fields (2.289).

In order to proceed further it is necessary to prove two useful theorems. The first theorem states that if a 4-vector field T^μ satisfies

$$\partial_\mu T^\mu = 0, \tag{2.292}$$

and if the components of T^μ are non-zero only in a finite spatial region, then the integral over 3-space,

$$I = \int T^4\, d^3x, \tag{2.293}$$

is an invariant. In order to prove this theorem we need to use the 4-dimensional analog of Gauss's theorem, which states that

$$\int_V \partial_\mu T^\mu\, d^4x = \oint_S T^\mu\, dS_\mu, \tag{2.294}$$

where dS_μ is an element of the 3-dimensional surface S bounding the 4-dimensional volume V. The particular volume over which the integration is performed is indicated in Fig. 1. The surfaces A and C are chosen so that the spatial components of T^μ vanish on A and C. This is always possible because it is assumed that the region over which the components of T^μ are non-zero is of finite extent. The surface B is chosen normal to the x^4-axis whereas the surface D is chosen normal to the $x^{4'}$-axis. Here, the x^μ and the $x^{\mu'}$ are coordinates in two arbitrarily chosen inertial frames. It follows from Eq. (2.294) that

$$\int T^4\, dS_4 + \int T^{4'}\, dS_{4'} = 0. \tag{2.295}$$

Here, we have made use of the fact that $T^\mu\, dS_\mu$ is a scalar and, therefore, has the same value in all inertial frames. Since $dS_4 = -d^3x$ and $dS_{4'} = d^3x'$ it follows that $I = \int T^4\, d^3x$ is an invariant under a Lorentz transformation. Incidentally,

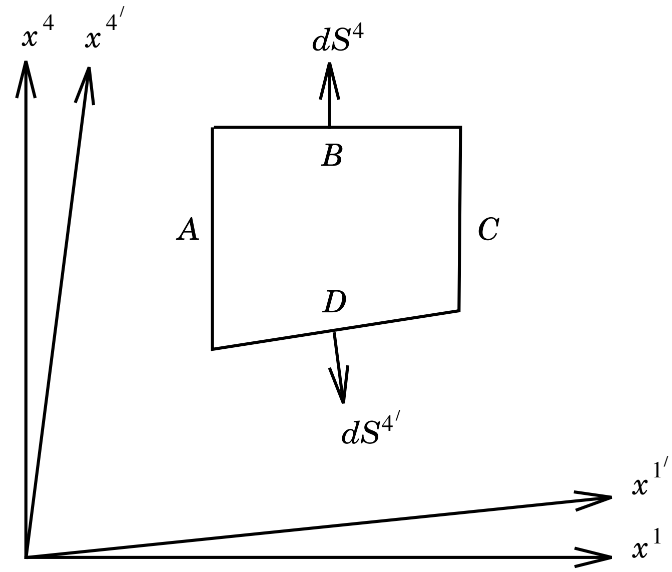


Figure 1: *The region of integration for proving the theorem associated with Eq. (2.293)*

the above argument also demonstrates that I is constant in time (just take the limit in which the two inertial frames are identical).

The second theorem is an extension of the first. Suppose that a 4-tensor field $Q^{\mu\nu}$ satisfies

$$\partial_\mu Q^{\mu\nu} = 0, \tag{2.296}$$

and has components which are only non-zero in a finite spatial region. Let A_μ be a 4-vector whose coefficients do not vary with position in space-time. It follows that $T^\mu = A_\nu Q^{\mu\nu}$ satisfies Eq. (2.292). Therefore,

$$I = \int A_\nu Q^{4\nu}\, d^3x \tag{2.297}$$

is an invariant. However, we can write

$$I = A_\mu B^\mu, \tag{2.298}$$

where

$$B^\mu = \int Q^{4\mu}\, d^3x. \tag{2.299}$$

It follows from the quotient rule that if $A_\mu B^\mu$ is an invariant for arbitrary A_μ then B^μ must transform as a constant (in time) 4-vector.

These two theorems enable us to convert differential conservation laws into integral conservation laws. For instance, in differential form the conservation of electrical charge is written

$$\partial_\mu J^\mu = 0. \tag{2.300}$$

However, from Eq. (2.293) this immediately implies that

$$Q = \frac{1}{c}\int J^4\, d^3x = \int \rho\, d^3x \tag{2.301}$$

is an invariant. In other words, the total electrical charge contained in space is both constant in time and the same in all inertial frames.

Suppose that S is the instantaneous rest frame of the charge. Let us consider the electromagnetic energy tensor $T^{\mu\nu}$ associated with all of the radiation emitted by the charge between times $t = 0$ and $t = dt$. According to Eq. (2.249) this tensor field satisfies

$$\partial_\mu T^{\mu\nu} = 0, \tag{2.302}$$

apart from a region of space of measure zero in the vicinity of the charge. Furthermore, the region of space over which $T^{\mu\nu}$ is non-zero is clearly finite, since we are only considering the fields emitted by the charge in a small time interval, and these fields propagate at a finite velocity. Thus, according to the second theorem

$$P^\mu = \frac{1}{c}\int T^{4\mu}\, d^3x \tag{2.303}$$

is a 4-vector. It follows from Section 2.23 that we can write $P^\mu = (d\boldsymbol{p}, dE/c)$, where $d\boldsymbol{p}$ and dE are the total momentum and energy carried off by the radiation emitted between times $t = 0$ and $t = dt$, respectively. As we have already mentioned, $d\boldsymbol{p} = 0$ in the instantaneous rest frame S. Transforming to an arbitrary inertial frame S' in which the instantaneous velocity of the charge is u, we obtain

$$dE^{'} = \gamma(u)\left(dE + u\, dp^1\right) = \gamma\, dE. \tag{2.304}$$

However, the time interval over which the radiation is emitted in S' is $dt' = \gamma dt$. Thus, the instantaneous power radiated by the charge,

$$P' = \frac{dE'}{dt'} = \frac{dE}{dt} = P, \tag{2.305}$$

is the same in all inertial frames.

We can make use of the fact that the power radiated by an accelerating charge is Lorentz invariant to find a relativistic generalization of the Lamor formula (2.291) which is valid in all inertial frames. We expect the power emitted by the charge to depend only on its 4-velocity and 4-acceleration. It follows that the Lamor formula can be written in Lorentz invariant form as

$$P = -\frac{e^2}{6\pi\epsilon_0 c^3}\, A_\mu A^\mu, \tag{2.306}$$

since the 4-acceleration takes the form $A^\mu = (\dot{\boldsymbol{u}}, 0)$ in the instantaneous rest frame. In a general inertial frame

$$-A_\mu A^\mu = \gamma^2 \left(\frac{d\gamma}{dt}\,\boldsymbol{u} + \gamma\,\dot{\boldsymbol{u}}\right)^2 - \gamma^2 c^2 \left(\frac{d\gamma}{dt}\right)^2, \tag{2.307}$$

where use has been made of Eq. (2.85). Furthermore, it is easily demonstrated that

$$\frac{d\gamma}{dt} = \gamma^3 \frac{\boldsymbol{u}\cdot\dot{\boldsymbol{u}}}{c^2}. \tag{2.308}$$

It follows, after a little algebra, that the relativistic generalization of Lamor's formula takes the form

$$P = \frac{e^2}{6\pi\epsilon_0 c^3}\,\gamma^6 \left[\dot{\boldsymbol{u}}^2 - \frac{(\boldsymbol{u}\wedge\dot{\boldsymbol{u}})^2}{c^2}\right]. \tag{2.309}$$

2.26 Radiation losses in charged particle accelerators

Radiation losses often limit the maximum practical energy attainable in a charged particle accelerator. Let us investigate radiation losses in various different types of accelerator device using the relativistic Lamor formula.

For a linear accelerator the motion is one dimensional. In this case, it is easily demonstrated that

$$\frac{dp}{dt} = m_0 \gamma^3 \, \dot{u}, \tag{2.310}$$

where use has been made of Eq. (2.308), and $p = \gamma m_0 u$ is the particle momentum in the direction of acceleration (the x-direction, say). Here, m_0 is the particle rest mass. Thus, Eq. (2.309) yields

$$P = \frac{e^2}{6\pi\epsilon_0 {m_0}^2 c^3} \left(\frac{dp}{dt}\right)^2. \tag{2.311}$$

The rate of change of momentum is equal to the force exerted on the particle in the x-direction, which in turn equals the change in the energy, E, of the particle per unit distance. Consequently,

$$P = \frac{e^2}{6\pi\epsilon_0 {m_0}^2 c^3} \left(\frac{dE}{dx}\right)^2. \tag{2.312}$$

Thus, in a linear accelerator the radiated power depends on the external force acting on the particle, and not on the actual energy or momentum of the particle. It is obvious from the above formula that light particles such as electrons are going to radiate a lot more than heavier particles such as protons. The ratio of the power radiated to the power supplied by the external sources is

$$\frac{P}{dE/dt} = \frac{e^2}{6\pi\epsilon_0 {m_0}^2 c^3} \frac{1}{u} \frac{dE}{dx} \simeq \frac{e^2}{6\pi\epsilon_0 m_0 c^2} \frac{1}{m_0 c^2} \frac{dE}{dx}, \tag{2.313}$$

since $u \simeq c$ for a highly relativistic particle. It is clear from the above expression that the radiation losses in an electron linear accelerator are negligible unless the gain in energy is of order $m_e c^2 = 0.511$ MeV in a distance of $e^2/(6\pi\epsilon_0 \, m_e c^2) = 1.28 \times 10^{-15}$ meters. That is 3×10^{14} MeV/meter. Typical energy gains are less that 10 MeV/meter. It is, therefore, obvious that radiation losses are completely negligible in linear accelerators, whether for electrons or for other heavier particles.

The situation is quite different in circular accelerator devices such as the synchrotron and the betatron. In such machines the momentum $\boldsymbol{p}$ changes rapidly in direction as the particle rotates, but the change in energy per revolution is small. Furthermore, the direction of acceleration is always perpendicular to the direction of motion. It follows from Eq. (2.309) that

$$P = \frac{e^2}{6\pi\epsilon_0 c^3}\,\gamma^4\,\dot{u}^2 = \frac{e^2}{6\pi\epsilon_0 c^3}\frac{\gamma^4 u^4}{\rho^2}, \tag{2.314}$$

where ρ is the orbit radius. Here, use has been made of the standard result $\dot{u} = u^2/\rho$ for circular motion. The radiative energy loss per revolution is given by

$$\delta E = \frac{2\pi\rho}{u} P = \frac{e^2}{3\epsilon_0 c^3}\frac{\gamma^4 u^3}{\rho}. \tag{2.315}$$

For highly relativistic ($u \simeq c$) electrons this expression yields

$$\delta E(\mathrm{MeV}) = 8.85\times 10^{-2}\,\frac{[E(\mathrm{GeV})]^4}{\rho(\mathrm{meters})}. \tag{2.316}$$

In the first electron synchrotrons, $\rho \sim 1$ meter, $E_{\max} \sim 0.3$ GeV. Hence, $\delta E_{\max} \sim$ 1 keV per revolution. This was less than, but not negligible compared to, the energy gain of a few keV per turn. For modern electron synchrotrons the limitation on the available radio-frequency power needed to overcome radiation losses becomes a major consideration, as is clear from the E^4 dependence of the radiated power per turn.

2.27 The angular distribution of radiation emitted by an accelerated charge

In order to calculate the angular distribution of the energy radiated by an accelerated charge we must think carefully about what is meant by the "rate of radiation" of the charge. This quantity is actually the amount of energy lost by the charge in a retarded time interval dt' during the emission of the signal. Thus,

$$P(t') = -\frac{dE}{dt'}, \tag{2.317}$$

where E is the energy of the charge. The Poynting vector

$$\boldsymbol{\epsilon} = \frac{\boldsymbol{E}_{\rm rad} \wedge \boldsymbol{B}_{\rm rad}}{\mu_0} = \epsilon_0 c\, E_{\rm rad}^{\,2}\, \frac{\boldsymbol{r}}{r}, \tag{2.318}$$

where use has been made of $\boldsymbol{B}_{\rm rad} = (\boldsymbol{r} \wedge \boldsymbol{E}_{\rm rad})/rc$ (see Eq. (2.281)), represents the energy flux per unit actual time, t. Thus, the energy loss rate of the charge into a given element of solid angle $d\Omega$ is

$$\frac{dP(t')}{d\Omega}\, d\Omega = -\frac{dE(\theta,\varphi)}{dt'}\, d\Omega = |\boldsymbol{\epsilon}|\, \frac{dt}{dt'}\, r^2\, d\Omega = \epsilon_0 c\, E_{\rm rad}^{\,2}\, \frac{s}{r}\, r^2\, d\Omega, \tag{2.319}$$

where use has been made of Eq. (2.267). Here, θ and φ are angular coordinates used to locate the element of solid angle. It follows from Eq. (2.287) that

$$\frac{dP(t')}{d\Omega} = \frac{e^2 r}{16\pi^2 \epsilon_0 c^3} \frac{[\boldsymbol{r} \wedge (\boldsymbol{r}_u \wedge \dot{\boldsymbol{u}})]^2}{s^5}. \tag{2.320}$$

Consider the special case where the direction of acceleration coincides with the direction of motion. Let us define spherical polar coordinates whose axis points along this common direction. It is easily demonstrated that the above expression reduces to

$$\frac{dP(t')}{d\Omega} = \frac{e^2\, \dot{u}^2}{16\pi^2 \epsilon_0 c^3} \frac{\sin^2\theta}{[1 - (u/c)\, \cos\theta]^5} \tag{2.321}$$

in this case. In the non-relativistic limit $u/c \rightarrow 0$ the radiation pattern has the familiar $\sin^2\theta$ dependence of dipole radiation. In particular, the pattern is symmetric in the forward ($\theta < \pi/2$) and backward ($\theta > \pi/2$) directions. However, as $u/c \rightarrow 1$ the radiation pattern becomes more and more concentrated in the forward direction. The angle $\theta_{\rm max}$ for which the intensity is a maximum is

$$\theta_{\rm max} = \cos^{-1}\left[\frac{1}{3\, u/c}(\sqrt{1 + 15\, u^2/c^2} - 1)\right]. \tag{2.322}$$

This expression yields $\theta_{\rm max} \rightarrow \pi/2$ as $u/c \rightarrow 0$ and $\theta_{\rm max} \rightarrow 1/(2\gamma)$ as $u/c \rightarrow 1$. Thus, for a highly relativistic charge the radiation is emitted in a narrow cone whose axis is aligned along the direction of motion. In this case, the angular distribution (2.321) reduces to

$$\frac{dP(t')}{d\Omega} \simeq \frac{2e^2\, \dot{u}^2}{\pi^2 \epsilon_0 c^3}\, \gamma^8 \frac{(\gamma\theta)^2}{[1 + (\gamma\theta)^2]^5}. \tag{2.323}$$

The total power radiated by the charge is obtained by integrating Eq. (2.321) over all solid angles. We obtain

$$P(t') = \frac{e^2\,\dot{u}^2}{8\pi\epsilon_0 c^3}\int_0^\pi \frac{\sin^3\theta\,d\theta}{[1-(u/c)\,\cos\theta]^5} = \frac{e^2\,\dot{u}^2}{8\pi\epsilon_0 c^3}\int_{-1}^{+1}\frac{(1-\mu^2)\,d\mu}{[1-(u/c)\,\mu]^5}. \tag{2.324}$$

It is easily verified that

$$\int_{-1}^{+1}\frac{(1-\mu^2)\,d\mu}{[1-(u/c)\,\mu]^5} = \frac{4}{3}\,\gamma^6. \tag{2.325}$$

Hence,

$$P(t') = \frac{e^2}{6\pi\epsilon_0 c^3}\,\gamma^6\,\dot{u}^2, \tag{2.326}$$

which agrees with Eq. (2.309) provided that $\boldsymbol{u}\wedge\dot{\boldsymbol{u}} = 0$.

2.28 Synchrotron radiation

Synchrotron radiation (*i.e.*, radiation emitted by a charged particle constrained to follow a circular orbit by a magnetic field) is of particular importance in astrophysics, since much of the observed radio frequency emission from supernova remnants and active galactic nuclei is thought to be of this type.

Consider a charged particle moving in a circle of radius a with constant angular velocity ω_0. Suppose that the orbit lies in the x-y plane. The radius vector pointing from the centre of the orbit to the retarded position of the charge is defined

$$\boldsymbol{\rho} = a\,(\cos\phi, \sin\phi, 0), \tag{2.327}$$

where $\phi = \omega_0\,t'$ is the angle subtended between this vector and the x-axis. The retarded velocity and acceleration of the charge take the form

$$\boldsymbol{u} = \frac{d\boldsymbol{\rho}}{dt'} = u\,(-\sin\phi, \cos\phi, 0), \tag{2.328a}$$

$$\dot{\boldsymbol{u}} = \frac{d\boldsymbol{u}}{dt'} = -\dot{u}\,(\cos\phi, \sin\phi, 0), \tag{2.328b}$$

where $u = a\,\omega_0$ and $\dot{u} = a\,{\omega_0}^2$. The observation point is chosen such that the radius vector $\boldsymbol{r}$, pointing from the retarded position of the charge to the observation point, is parallel to the y-z plane. Thus, we can write

$$\boldsymbol{r} = r\,(0, \sin\alpha, \cos\alpha), \tag{2.329}$$

where α is the angle subtended between this vector and the z-axis. As usual, we define θ as the angle subtended between the retarded radius vector $\boldsymbol{r}$ and the retarded direction of motion of the charge $\boldsymbol{u}$. It follows that

$$\cos\theta = \frac{\boldsymbol{u}\cdot\boldsymbol{r}}{u\,r} = \sin\alpha\,\cos\phi. \tag{2.330}$$

It is easily seen that

$$\dot{\boldsymbol{u}}\cdot\boldsymbol{r} = -\dot{u}r\,\sin\alpha\,\sin\phi. \tag{2.331}$$

A little vector algebra shows that

$$[\boldsymbol{r}\wedge(\boldsymbol{r}_u\wedge\dot{\boldsymbol{u}})]^2 = -(\boldsymbol{r}\cdot\dot{\boldsymbol{u}})^2\,r^2\,(1-u^2/c^2) + \dot{u}^2\,r^4\,(1-\boldsymbol{r}\cdot\boldsymbol{u}/rc)^2, \tag{2.332}$$

giving

$$[\boldsymbol{r}\wedge(\boldsymbol{r}_u\wedge\dot{\boldsymbol{u}})]^2 = \dot{u}^2\,r^4\left[\left(1-\frac{u}{c}\,\cos\theta\right)^2 - \left(1-\frac{u^2}{c^2}\right)\tan^2\phi\cos^2\theta\right]. \tag{2.333}$$

Making use of Eq. (2.320), we obtain

$$\frac{dP(t')}{d\Omega} = \frac{e^2\,\dot{u}^2}{16\pi^2\epsilon_0 c^3}\,\frac{[1-(u/c)\,\cos\theta)]^2-(1-u^2/c^2)\tan^2\phi\,\cos^2\theta}{[1-(u/c)\,\cos\theta]^5}. \tag{2.334}$$

It is convenient to write this result in terms of the angles α and ϕ, instead of θ and ϕ. After a little algebra we obtain

$$\frac{dP(t')}{d\Omega} = \frac{e^2\,\dot{u}^2}{16\pi^2\epsilon_0 c^3}\,\frac{[1-(u^2/c^2)]\cos^2\alpha + [(u/c)-\sin\alpha\,\cos\phi]^2}{[1-(u/c)\,\sin\alpha\,\cos\phi]^5}. \tag{2.335}$$

Let us consider the radiation pattern emitted in the plane of the orbit; *i.e.*, $\alpha = \pi/2$, with $\cos\phi = \cos\theta$. It is easily seen that

$$\frac{dP(t')}{d\Omega} = \frac{e^2\,\dot{u}^2}{16\pi^2\epsilon_0 c^3}\,\frac{[(u/c)-\cos\theta]^2}{[1-(u/c)\,\cos\theta]^5}. \tag{2.336}$$

In the non-relativistic limit the radiation pattern has a $\cos^2\theta$ dependence. Thus, the pattern is like that of dipole radiation where the axis is aligned along the instantaneous direction of acceleration. As the charge becomes more relativistic the radiation lobe in the forward direction (*i.e.*, $0 < \theta < \pi/2$) becomes more more focused and more intense. Likewise, the radiation lobe in the backward direction (*i.e.*, $\pi/2 < \theta < \pi$) becomes more diffuse. The radiation pattern has zero intensity at the angles

$$\theta_0 = \cos^{-1}(u/c). \tag{2.337}$$

These angles demark the boundaries between the two radiation lobes. In the non-relativistic limit $\theta_0 = \pm\pi/2$, so the two lobes are of equal angular extents. In the highly relativistic limit $\theta_0 \to \pm 1/\gamma$, so the forward lobe becomes highly concentrated about the forward direction ($\theta = 0$). In the latter limit Eq. (2.336) reduces to

$$\frac{dP(t')}{d\Omega} \simeq \frac{e^2\,\dot{u}^2}{2\pi^2\epsilon_0 c^3}\,\gamma^6\,\frac{[1-(\gamma\theta)^2]^2}{[1+(\gamma\theta)^2]^5}. \tag{2.338}$$

Thus, the radiation emitted by a highly relativistic charge is focused into an intense beam of angular extent $1/\gamma$ pointing in the instantaneous direction of motion. The maximum intensity of the beam scales like γ^6.

Integration of Eq. (2.335) over all solid angle (using $d\Omega = \sin\alpha\,d\alpha\,d\phi$) yields (not very easily!)

$$P(t') = \frac{e^2}{6\pi\epsilon_0 c^3}\,\gamma^4\,\dot{u}^2, \tag{2.339}$$

which agrees with Eq. (2.309) provided that $\boldsymbol{u}\cdot\dot{\boldsymbol{u}} = 0$. This expression can also be written

$$\frac{P}{m_0 c^2} = \frac{2}{3}\frac{{\omega_0}^2 r_0}{c}\,\beta^2\,\gamma^4, \tag{2.340}$$

where $r_0 = e^2/(4\pi\epsilon_0\, m_0 c^2) = 2.82\times 10^{-15}$ meters is the *classical electron radius*, m_0 is the rest mass of the charge, and $\beta = u/c$. If the circular motion takes place in an orbit of radius a perpendicular to a magnetic field $\boldsymbol{B}$, then ω_0 satisfies $\omega_0 = eB/m_0\gamma$. Thus, the radiated power is

$$\frac{P}{m_0 c^2} = \frac{2}{3}\left(\frac{eB}{m_0}\right)^2\frac{r_0}{c}\,(\beta\gamma)^2, \tag{2.341}$$

and the radiated energy ΔE per revolution is

$$\frac{\Delta E}{m_0 c^2} = \frac{4\pi}{3}\frac{r_0}{a}\,\beta^3\,\gamma^4. \tag{2.342}$$

Let us consider the frequency distribution of the emitted radiation in the highly relativistic limit. Suppose, for the sake of simplicity, that the observation point lies in the plane of the orbit (*i.e.*, $\alpha = \pi/2$). Since the radiation emitted by the charge is beamed very strongly in the charge's instantaneous direction of motion, a fixed observer is only going to see radiation (at some later time) when this direction points almost directly towards the point of observation. This occurs once every rotation period when $\phi \simeq 0$, assuming that $\omega_0 > 0$. Note that the point of observation is located many orbit radii away from the centre of the orbit along the positive y-axis. Thus, our observer sees short periodic pulses of radiation from the charge. The repetition frequency of the pulses (in radians per second) is ω_0. Let us calculate the duration of each pulse. Since the radiation emitted by the charge is focused into a narrow beam of angular extent $\Delta\theta \sim 1/\gamma$, our observer only sees radiation from the charge when $\phi \lesssim \Delta\theta$. Thus, the observed pulse is emitted during a time interval $\Delta t' = \Delta\theta/\omega_0$. However, the pulse is received in a somewhat shorter time interval

$$\Delta t = \frac{\Delta\theta}{\omega_0}\left(1 - \frac{u}{c}\right), \tag{2.343}$$

because the charge is slightly closer to the point of observation at the end of the pulse than at the beginning. The above equation reduces to

$$\Delta t \simeq \frac{\Delta\theta}{2\omega_0\gamma^2} \sim \frac{1}{\omega_0\,\gamma^3}, \tag{2.344}$$

since $\gamma \gg 1$ and $\Delta\theta \sim 1/\gamma$. The width $\Delta\omega$ of the pulse in frequency space obeys $\Delta\omega\,\Delta t \sim 1$. Hence,

$$\Delta\omega = \gamma^3\,\omega_0. \tag{2.345}$$

In other words, the emitted frequency spectrum contains harmonics of frequency up to γ^3 times that of the fundamental, ω_0.

More involved calculations[8] show that in the ultra-relativistic limit $\gamma \gg 1$ the power radiated in the lth harmonic (whose frequency is $\omega = l\omega_0$) is given by

$$P_l = 0.52 \left(\frac{e^2}{4\pi\epsilon_0 c}\right) {\omega_0}^2 \, l^{1/3} \tag{2.346}$$

for $1 \ll l \ll \gamma^3$, and

$$P_l = \frac{1}{2\sqrt{\pi}} \left(\frac{e^2}{4\pi\epsilon_0 c}\right) {\omega_0}^2 \left(\frac{l}{\gamma}\right)^{1/2} \exp[(-2/3)(l/\gamma^3)] \tag{2.347}$$

for $l \gg \gamma^3$. Note that the spectrum cuts off approximately at the harmonic order γ^3, as predicted earlier. It can also be demonstrated[9] that *seven* times as much energy is radiated with a polarization parallel to the orbital plane than with a perpendicular polarization. A $P(\omega) \propto \omega^{1/3}$ power spectrum at low frequencies coupled with a high degree of polarization are the hallmarks of synchrotron radiation. In fact, these two features are used in astrophysics to identify synchrotron radiation from supernova remnants, active galactic nuclei, *etc.*

[8] L. Landau, and E. Lifshitz, *The classical theory of fields*, (Addison-Wesley, 1951), pp. 215 ff.

[9] J.D. Jackson, *Classical electrodynamics*, (Wiley, 1962), pp. 672 ff.

3 The effect of dielectric and magnetic media on electric and magnetic fields

3.1 Polarization

The terrestrial environment is characterized by dielectric media (*e.g.*, air, water) which are, for the most part, electrically neutral, since they are made up of neutral atoms and molecules. However, if these atoms and molecules are placed in an electric field they tend to *polarize*. Suppose that when a given neutral molecule is placed in an electric field $\boldsymbol{E}$ the centre of charge of its constituent electrons (whose total charge is $-q$) is displaced by a distance $-\boldsymbol{r}$ with respect to the centre of charge of its constituent atomic nuclei. The *dipole moment* of the molecule is defined $\boldsymbol{p} = q\,\boldsymbol{r}$. If there are N such molecules per unit volume then the *electric polarization* $\boldsymbol{P}$ (*i.e.*, the dipole moment per unit volume) is given by $\boldsymbol{P} = N\boldsymbol{p}$. More generally,

$$\boldsymbol{P}(\boldsymbol{r}) = \sum_i N_i \langle \boldsymbol{p}_i \rangle, \tag{3.1}$$

where $\langle \boldsymbol{p}_i \rangle$ is the average dipole moment of the ith type of molecule in the vicinity of point $\boldsymbol{r}$, and N_i is the average number of such molecules per unit volume at $\boldsymbol{r}$.

It is easily demonstrated that any divergence of the polarization field $\boldsymbol{P}(\boldsymbol{r})$ gives rise to an effective charge density ρ_b in the medium. In fact,

$$\rho_b = -\nabla\cdot\boldsymbol{P}. \tag{3.2}$$

This charge density is attributable to *bound charges* (*i.e.*, charges which arise from the polarization of neutral atoms), and is usually distinguished from the charge density ρ_f due to *free charges*, which represents a net surplus or deficit of electrons in the medium. Thus, the total charge density ρ in the medium is

$$\rho = \rho_f + \rho_b. \tag{3.3}$$

It must be emphasized that both terms in this equation represent real physical charge. Nevertheless, it is useful to make the distinction between bound and free charges, especially when it comes to working out the energy associated with electric fields in dielectric media.

Gauss' law takes the differential form

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0} = \frac{\rho_f + \rho_b}{\epsilon_0}. \tag{3.4}$$

This expression can be rearranged to give

$$\nabla\cdot\boldsymbol{D} = \rho_f, \tag{3.5}$$

where

$$\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P} \tag{3.6}$$

is termed the *electric displacement*, and has the same dimensions as $\boldsymbol{P}$ (dipole moment per unit volume). The divergence theorem tells us that

$$\oint_S \boldsymbol{D}\cdot d\boldsymbol{S} = \int_V \rho_f \, dV. \tag{3.7}$$

In other words, the flux of $\boldsymbol{D}$ out of some closed surface S is equal to the total free charge enclosed within that surface. Unlike the electric field $\boldsymbol{E}$ (which is the force acting on unit charge) or the polarization $\boldsymbol{P}$ (the dipole moment per unit volume), the electric displacement $\boldsymbol{D}$ has no clear physical meaning. The only reason for introducing it is that it enables us to calculate fields in the presence of dielectric materials without first having to know the distribution of polarized charges. However, this is only possible if we have a *constitutive relation* connecting $\boldsymbol{E}$ and $\boldsymbol{D}$. It is conventional to assume that the induced polarization $\boldsymbol{P}$ is directly proportional to the electric field $\boldsymbol{E}$, so that

$$\boldsymbol{P} = \epsilon_0\chi_e\boldsymbol{E}, \tag{3.8}$$

where χ_e is termed the *electric susceptibility* of the medium. It follows that

$$\boldsymbol{D} = \epsilon_0\epsilon\boldsymbol{E}, \tag{3.9}$$

where

$$\epsilon = 1 + \chi_e \tag{3.10}$$

is termed the *dielectric constant* or *relative permittivity* of the medium. (Likewise, ϵ_0 is termed the *permittivity of free space.*) It follows from Eqs. (3.5) and (3.9) that

$$\nabla\cdot\boldsymbol{E} = \frac{\rho_f}{\epsilon_0\epsilon}. \tag{3.11}$$

Thus, the electric fields produced by free charges in a dielectric medium are analogous to those produced by the same charges in a vacuum, except that they are reduced by a factor ϵ. This reduction can be understood in terms of a polarization of the atoms or molecules of the dielectric medium that produces electric fields in opposition to that of given charge. One immediate consequence is that the capacitance of a capacitor is increased by a factor ϵ if the empty space between the electrodes is filled with a dielectric medium of dielectric constant ϵ (assuming that fringing fields can be neglected).

It must be understood that Eqs. (3.8)–(3.11) are just an *approximation* which is generally found to hold under terrestrial conditions (provided that the fields are not too large) for *isotropic* media. For anisotropic media (*e.g.*, crystals) Eq. (3.9) generalizes to

$$\boldsymbol{D} = \epsilon_0\, \boldsymbol{\epsilon}\cdot\boldsymbol{E}, \tag{3.12}$$

where $\boldsymbol{\epsilon}$ is a second rank tensor known as the *dielectric tensor*. For strong electric fields $\boldsymbol{D}$ ceases to vary linearly with $\boldsymbol{E}$. Indeed, for sufficiently strong electric fields neutral molecules are disrupted and the whole concept of a dielectric medium becomes meaningless.

3.2 Boundary conditions for $\boldsymbol{E}$ and $\boldsymbol{D}$

When the space near a set of charges contains dielectric material of non-uniform dielectric constant then the electric field no longer has the same form as in vacuum. Suppose, for example, that the space is occupied by two dielectric media whose uniform dielectric constants are ϵ_1 and ϵ_2. What are the matching conditions on $\boldsymbol{E}$ and $\boldsymbol{D}$ at the boundary between the two media?

Imagine a Gaussian pill-box enclosing part of the boundary surface between the two media. The thickness of the pill-box is allowed to tend towards zero, so that the only contribution to the outward flux of $\boldsymbol{D}$ comes from its flat faces. These faces are parallel to the bounding surface and lie in each of the two media. Their outward normals are $d\boldsymbol{S}_1$ (in medium 1) and $d\boldsymbol{S}_2$, where $d\boldsymbol{S}_1 = -d\boldsymbol{S}_2$. Assuming that there is no free charge inside the disk (which is reasonable in the limit where the volume of the disk tends towards zero), then Eq. (3.7) yields

$$\boldsymbol{D}_1 \cdot d\boldsymbol{S}_1 + \boldsymbol{D}_2 \cdot d\boldsymbol{S}_2 = 0, \tag{3.13}$$

where $\boldsymbol{D}_1$ is the electric displacement in medium 1 at the boundary with medium 2, *etc.* The above equation can be rewritten

$$(\boldsymbol{D}_2 - \boldsymbol{D}_1) \cdot \boldsymbol{n}_{21} = 0, \tag{3.14}$$

where $\boldsymbol{n}_{21}$ is the normal to the boundary surface, directed from medium 1 to medium 2. If the fields and charges are non time varying then Maxwell's equations yield $\nabla \wedge \boldsymbol{E} = 0$, which give the familiar boundary condition (obtained by integrating around a small loop which straddles the boundary surface)

$$(\boldsymbol{E}_2 - \boldsymbol{E}_1) \wedge \boldsymbol{n}_{21} = 0. \tag{3.15}$$

In other word, the normal component of the electric displacement and the tangential component of the electric field are both continuous across any boundary between two dielectric materials.

3.3 Boundary value problems with dielectrics - I

Consider a point charge q embedded in a semi-infinite dielectric ϵ_1 a distance d away from a plane interface which separates the first medium from another semi-infinite dielectric ϵ_2. The interface is assumed to coincide with the plane $z = 0$. We need to find solutions to the equations

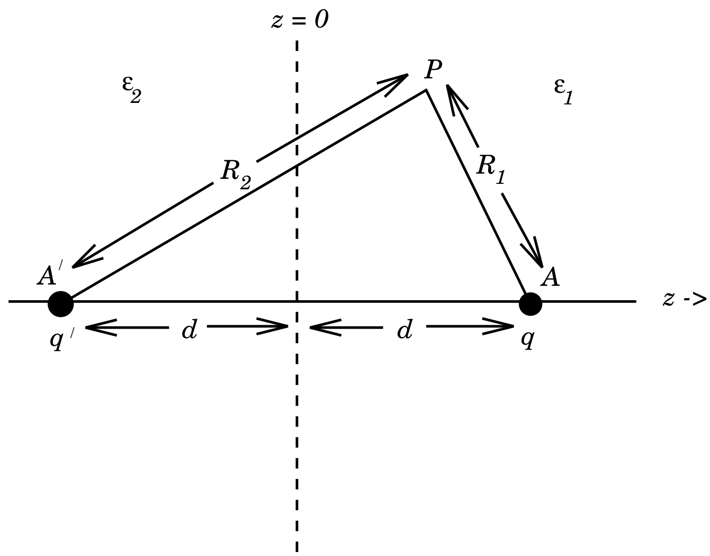
$$\epsilon_1 \nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{3.16}$$

for $z > 0$,

$$\epsilon_2 \nabla \cdot \boldsymbol{E} = 0 \tag{3.17}$$

for $z < 0$, and

$$\nabla \wedge \boldsymbol{E} = 0 \tag{3.18}$$



everywhere, subject to the boundary conditions at $z = 0$ that

$$\begin{aligned} \epsilon_1 E_z(z = 0^+) &= \epsilon_2 E_z(z = 0^-), && (3.19a)\\ E_x(z = 0^+) &= E_x(z = 0^-), && (3.19b)\\ E_y(z = 0^+) &= E_y(z = 0^-). && (3.19c) \end{aligned}$$

In order to solve this problem we will employ a slightly modified form of the well known method of images. Since $\nabla \wedge \boldsymbol{E} = 0$ everywhere, the electric field can be written in terms of a scalar potential. So, $\boldsymbol{E} = -\nabla\phi$. Consider the region $z > 0$. Let us assume that the scalar potential in this region is the same as that obtained when the whole of space is filled with the dielectric ϵ_1 and, in addition to the real charge q at position A, there is a second charge q' at the image position A' (see diagram). If this is the case then the potential at some point P in the region $z > 0$ is given by

$$\phi(z > 0) = \frac{1}{4\pi\epsilon_0\epsilon_1}\left(\frac{q}{R_1} + \frac{q'}{R_2}\right), \tag{3.20}$$

where $R_1 = \sqrt{\rho^2 + (d - z)^2}$ and $R_2 = \sqrt{\rho^2 + (d + z)^2}$, when written in terms of cylindrical polar coordinates (ρ, φ, z). Note that the potential (3.20) clearly is a solution of Eq. (3.16) in the region $z > 0$. It gives $\nabla\cdot\boldsymbol{E} = 0$, with the appropriate singularity at the position of the point charge q.

Consider the region $z < 0$. Let us assume that the scalar potential in this region is the same as that obtained when the whole of space is filled with the dielectric ϵ_2 and a charge q'' is located at the point A. If this is the case then the potential in this region is given by

$$\phi(z < 0) = \frac{1}{4\pi\epsilon_0\epsilon_2}\frac{q''}{R_1}. \tag{3.21}$$

The above potential is clearly a solution of Eq. (3.17) in the region $z < 0$. It gives $\nabla\cdot\boldsymbol{E} = 0$, with no singularities.

It now remains to choose q' and q'' in such a manner that the boundary conditions (3.19) are satisfied. The boundary conditions (3.19b) and (3.19c) are obviously satisfied if the scalar potential is continuous at the interface between the two dielectric media:

$$\phi(z = 0^+) = \phi(z = 0^-). \tag{3.22}$$

The boundary condition (3.19a) implies a jump in the normal derivative of the scalar potential across the interface:

$$\epsilon_1\,\frac{\partial\phi(z = 0^+)}{\partial z} = \epsilon_2\,\frac{\partial\phi(z = 0^-)}{\partial z}. \tag{3.23}$$

The first matching condition yields

$$\frac{q + q'}{\epsilon_1} = \frac{q''}{\epsilon_2}, \tag{3.24}$$

whereas the second yields

$$q - q' = q''. \tag{3.25}$$

Here, use has been made of

$$\frac{\partial}{\partial z}\left(\frac{1}{R_1}\right)_{z=0} = -\frac{\partial}{\partial z}\left(\frac{1}{R_2}\right)_{z=0} = \frac{d}{(\rho^2 + d^2)^{3/2}}. \tag{3.26}$$

Equations (3.24) and (3.25) imply that

$$q' = -\left(\frac{\epsilon_2 - \epsilon_1}{\epsilon_2 + \epsilon_1}\right) q, \tag{3.27a}$$

$$q'' = \left(\frac{2\epsilon_2}{\epsilon_2 + \epsilon_1}\right) q. \tag{3.27b}$$

The polarization charge density is given by $\rho_b = -\nabla \cdot \boldsymbol{P}$, However, inside either dielectric $\boldsymbol{P} = \epsilon_0 \chi_e \boldsymbol{E}$, so $\nabla \cdot \boldsymbol{P} = \epsilon_0 \chi_e \nabla \cdot \boldsymbol{E} = 0$, except at the point charge q. Thus, there is zero polarization charge density in either dielectric medium. At the interface χ_e takes a discontinuous jump,

$$\Delta \chi_e = \epsilon_1 - \epsilon_2. \tag{3.28}$$

This implies that there is a polarization charge sheet on the interface between the two dielectric media. In fact,

$$\sigma_{\rm pol} = -(\boldsymbol{P}_2 - \boldsymbol{P}_1) \cdot \boldsymbol{n}_{21}, \tag{3.29}$$

where $\boldsymbol{n}_{21}$ is a unit normal to the interface pointing from medium 1 to medium 2 (*i.e.*, along the positive z-axis). Since

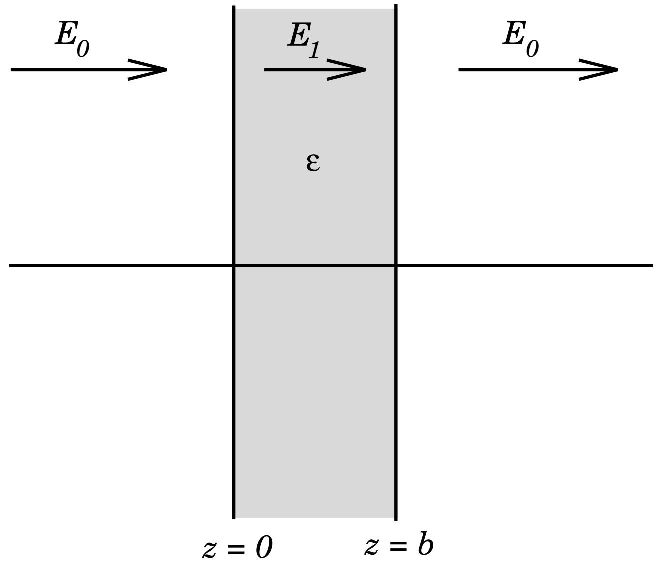
$$\boldsymbol{P}_i = \epsilon_0(\epsilon_i - 1)\boldsymbol{E} = -\epsilon_0(\epsilon_i - 1)\nabla\phi \tag{3.30}$$

in either medium, it is easy to demonstrate that

$$\sigma_{\rm pol} = -\frac{q}{2\pi} \frac{\epsilon_2 - \epsilon_1}{\epsilon_1(\epsilon_2 + \epsilon_1)} \frac{d}{(\rho^2 + d^2)^{3/2}}. \tag{3.31}$$

In the limit $\epsilon_2 \gg \epsilon_1$, the dielectric ϵ_2 behaves like a conducting medium (*i.e.*, $\boldsymbol{E} \rightarrow 0$ in the region $z < 0$), and the polarization surface charge density on the interface approaches that obtained in the case when the plane $z = 0$ coincides with a conducting surface.

The above method can clearly be generalized to deal with problems involving many point charges in the presence of many different dielectric media whose interfaces form parallel planes.



3.4 Boundary value problems with dielectrics - II

Consider a plane slab of dielectric ϵ lying between $z = 0$ and $z = b$. Suppose that this slab is placed in a uniform z-directed electric field of strength E_0. What is the field strength E_1 inside the slab?

Since there are no free charges and this is a one-dimensional problem, it is clear from Eq. (3.5) that the electric displacement D is the same in both the dielectric slab and the vacuum region which surrounds it. In the vacuum region $D = \epsilon_0 E_0$, whereas $D = \epsilon_0 \epsilon\, E_1$ in the dielectric. It follows that

$$E_1 = \frac{E_0}{\epsilon}. \tag{3.32}$$

In other words, the electric field inside the slab is reduced by polarization charges. As before, there is zero polarization charge density inside the dielectric. However, there is a uniform polarization charge sheet on both surfaces of the slab. It is easily demonstrated that

$$\sigma_{\rm pol}(z = b) = -\sigma_{\rm pol}(z = 0) = \epsilon_0\, E_0\, \frac{\epsilon - 1}{\epsilon}. \tag{3.33}$$

In the limit $\epsilon \gg 1$, the slab acts like a conductor and $E_1 \to 0$.

Let us now generalize this result. Consider a dielectric medium whose dielectric constant ϵ varies with z. The medium is assumed to be of finite extent and is surrounded by a vacuum, so that $\epsilon(z) \rightarrow 1$ as $|z| \rightarrow \infty$. Suppose that this dielectric is placed in a uniform z-directed electric field E_0. What is the field $E(z)$ inside the dielectric?

We know that the electric displacement inside the dielectric is given by $D(z) = \epsilon_0\, \epsilon(z)\, E(z)$. We also know from Eq. (3.5) that, since there are no free charges and this is a one-dimensional problem,

$$\frac{dD(z)}{dz} = \epsilon_0 \,\frac{d[\epsilon(z)E(z)]}{dz} = 0. \tag{3.34}$$

Furthermore, $E(z) \rightarrow E_0$ as $|z| \rightarrow \infty$. It follows that

$$E(z) = \frac{E_0}{\epsilon(z)}. \tag{3.35}$$

Thus, the electric field is inversely proportional to the dielectric constant inside the dielectric medium. The polarization charge density inside the dielectric is given by

$$\rho_b = \epsilon_0 \,\frac{dE(z)}{dz} = \epsilon_0 \, E_0 \frac{d}{dz}\left[\frac{1}{\epsilon(z)}\right]. \tag{3.36}$$

3.5 Boundary value problems with dielectrics - III

Suppose that a dielectric sphere of radius a and dielectric constant ϵ is placed in a z-directed electric field of strength E_0 (in the absence of the sphere). What is the electric field inside and around the sphere?

Since this is a static problem we can write $\boldsymbol{E} = -\nabla\phi$. There are no free charges, so Eqs. (3.5) and (3.9) imply that

$$\nabla^2\phi = 0 \tag{3.37}$$

everywhere. The boundary conditions (3.14) and (3.15) reduce to

$$\epsilon \left.\frac{\partial\phi}{\partial r}\right|_{r=a^-} \quad = \quad \left.\frac{\partial\phi}{\partial r}\right|_{r=a^+}, \tag{3.38a}$$

$$\left.\frac{\partial\phi}{\partial\theta}\right|_{r=a^-} = \left.\frac{\partial\phi}{\partial\theta}\right|_{r=a^+}. \tag{3.38b}$$

Furthermore,

$$\phi(r,\theta,\varphi) \to -E_0\, r\, \cos\theta \tag{3.39}$$

as $r \to 0$. Here, (r,θ,φ) are spherical polar coordinates centred on the sphere.

Let us search for an axisymmetric solution, $\phi = \phi(r,\theta)$. Since the solutions to Poisson's equation are *unique*, we know that if we can find such a solution which satisfies all of the boundary conditions then we can be sure that this is the correct solution. Equation (3.37) reduces to

$$\frac{1}{r}\frac{\partial^2(r\phi)}{\partial r^2} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\phi}{\partial\theta}\right) = 0. \tag{3.40}$$

Straightforward separation of the variables yields

$$\phi(r,\theta) = \sum_{l=0}^{\infty}(A_l\, r^l + B_l\, r^{-(l+1)})P_l(\cos\theta), \tag{3.41}$$

where l is a non-negative integer, the A_l and B_l are arbitrary constants, and $P_l(x)$ is a solution to Legendre's equation,

$$\frac{d}{dx}\left[(1-x^2)\,\frac{dP_l}{dx}\right] + l(l+1)\,P_l = 0, \tag{3.42}$$

which is single-valued, finite, and continuous in the interval $-1 \leq x \leq +1$. It can be demonstrated that Eq. (3.42) only possesses such solutions when l takes an integer value. The $P_l(x)$ are known as *Legendre polynomials* (since they are polynomials of order l in x), and are specified by *Rodrigues' formula*

$$P_l(x) = \frac{1}{2^l\, l!}\frac{d^l}{dx^l}(x^2-1)^l. \tag{3.43}$$

Since Eq. (3.42) is a Sturm-Liouville type equation, and the Legendre polynomials satisfy Sturm-Liouville type boundary conditions at $x = \pm 1$, it immediately follows that the $P_l(\cos\theta)$ are orthogonal functions which form *a complete set* in θ-space. The orthogonality relation can be written

$$\int_{-1}^{1} P_{l'}(x)\,P_l(x)\,dx = \frac{2}{2l+1}\,\delta_{ll'}. \tag{3.44}$$

The Legendre polynomials form a complete set of angular functions, and it is easily demonstrated that the r^l and the $r^{-(l+1)}$ form a complete set of radial functions. It follows that Eq. (3.41), with the A_l and B_l unspecified, represents a completely general axisymmetric solution of Eq. (3.37) which is well behaved in θ-space. We now need to find values of the A_l and B_l which are consistent with the boundary conditions.

Let us divide space into the regions $r \leq a$ and $r > a$. In the former region

$$\phi(r,\theta) = \sum_{l=0}^{\infty} A_l\, r^l\, P_l(\cos\theta), \tag{3.45}$$

where we have rejected the $r^{-(l+1)}$ radial solutions because they diverge unphysically as $r \to 0$. In the latter region

$$\phi(r,\theta) = \sum_{l=0}^{\infty} (B_l\, r^l + C_l\, r^{-(l+1)})\,P_l(\cos\theta). \tag{3.46}$$

However, it is clear from the boundary condition (3.39), and Eq. (3.43), that the only non-vanishing B_l is $B_1 = -E_0$. This follows since $P_1(\cos\theta) = \cos\theta$. The boundary condition (3.38b) (which integrates to give $\phi(r=a^-) = \phi(r=a^+)$ for a potential which is well behaved in θ-space) gives

$$A_1 = -E_0 + \frac{C_1}{a^3}, \tag{3.47}$$

and

$$A_l = \frac{C_l}{a^{2l+1}} \tag{3.48}$$

for $l \neq 1$. Note that it is appropriate to match the coefficients of the $P_l(\cos\theta)$ since these functions are *orthogonal.* The boundary condition (3.38a) yields

$$\epsilon A_1 = -E_0 - 2\,\frac{C_1}{a^3}, \tag{3.49}$$

and

$$\epsilon\, l\, A_l = -(l+1)\,\frac{C_l}{a^{2l+1}} \tag{3.50}$$

for $l \neq 1$. Equations (3.48) and (3.50) give $A_l = C_l = 0$ for $l \neq 1$. Equations (3.47) and (3.49) reduce to

$$A_1 \;=\; -\left(\frac{3}{2+\epsilon}\right) E_0, \tag{3.51a}$$

$$C_1 \;=\; \left(\frac{\epsilon-1}{\epsilon+2}\right) a^3 E_0. \tag{3.51b}$$

The solution is therefore

$$\phi = -\left(\frac{3}{2+\epsilon}\right) E_0\, r\cos\theta \tag{3.52}$$

for $r \leq a$, and

$$\phi = -E_0\, r\cos\theta + \left(\frac{\epsilon-1}{\epsilon+2}\right) E_0\,\frac{a^3}{r^2}\cos\theta \tag{3.53}$$

for $r > a$.

Equation (3.52) is the potential of a uniform z-directed electric field of strength

$$E_1 = \frac{3}{2+\epsilon}\, E_0. \tag{3.54}$$

Note that $E_1 < E_0$ provided that $\epsilon > 1$. Thus, the electric field strength is reduced inside the dielectric sphere due to partial shielding by polarization charges. Outside the sphere the potential is equivalent to that of the applied field E_0 plus the field of a point electric dipole, located at the origin and pointing in the z-direction, whose dipole moment is

$$p = 4\pi\epsilon_0 \left(\frac{\epsilon-1}{\epsilon+2}\right) a^3 E_0. \tag{3.55}$$

This dipole moment can be interpreted as the volume integral of the polarization $\boldsymbol{P}$ over the sphere. The polarization is

$$\boldsymbol{P} = \epsilon_0(\epsilon-1)\, E_1\, \hat{\boldsymbol{z}} = 3\epsilon_0 \left(\frac{\epsilon-1}{\epsilon+2}\right) E_0\, \hat{\boldsymbol{z}}. \tag{3.56}$$

Since the polarization is uniform there is zero polarization charge density inside the sphere. However, there is a polarization charge sheet on the surface of the sphere whose density is given by $\sigma_{\rm pol} = \boldsymbol{P}\cdot\hat{\boldsymbol{r}}$ (see Eq. (3.29)). It follows that

$$\sigma_{\rm pol} = 3\epsilon_0 \left(\frac{\epsilon-1}{\epsilon+2}\right) E_0 \cos\theta. \tag{3.57}$$

The problem of a dielectric cavity of radius a in a dielectric medium with dielectric constant ϵ and with an applied electric field E_0 parallel to the z-axis can be treated in much the same manner as that of a dielectric sphere. In fact, it is easily demonstrated that the results for the cavity can be obtained from those for the sphere by making the transformation $\epsilon \rightarrow 1/\epsilon$. Thus, the field inside the cavity is uniform, parallel to the z-axis, and of magnitude

$$E_1 = \frac{3\epsilon}{2\epsilon+1} E_0. \tag{3.58}$$

Note that $E_1 > E_0$ provided that $\epsilon > 1$. The field outside the cavity is the original field plus that of a z-directed dipole, located at the origin, whose dipole moment is

$$p = -4\pi\epsilon_0 \left(\frac{\epsilon-1}{2\epsilon+1}\right) a^3 E_0. \tag{3.59}$$

Here, the negative sign implies that the dipole points in the opposite direction to the external field.

3.6 The energy density within a dielectric medium

Consider a system of free charges embedded in a dielectric medium. The increase in the total energy when a small amount of free charge $\delta\rho_f$ is added to the system is given by

$$\delta U = \int \phi\, \delta\rho_f\, d^3\boldsymbol{r}, \tag{3.60}$$

where the integral is taken over all space and $\phi(\boldsymbol{r})$ is the electrostatic potential. Here, it is assumed that the original charges and the dielectric are held fixed, so that no mechanical work is performed. It follows from Eq. (3.5) that

$$\delta U = \int \phi \, \nabla \cdot \delta \boldsymbol{D} \, d^3\boldsymbol{r}, \tag{3.61}$$

where $\delta\boldsymbol{D}$ is the change in the electric displacement associated with the charge increment. Now the above equation can also be written

$$\delta U = \int \nabla \cdot (\phi \, \delta \boldsymbol{D}) \, d^3\boldsymbol{r} - \int \nabla\phi \cdot \delta\boldsymbol{D} \, d^3\boldsymbol{r}, \tag{3.62}$$

giving

$$\delta U = \int \phi \, \delta \boldsymbol{D} \cdot d\boldsymbol{S} - \int \nabla\phi \cdot \delta\boldsymbol{D} \, d^3\boldsymbol{r}, \tag{3.63}$$

where use has been made of Gauss's theorem. If the dielectric medium is of finite spatial extent then we can neglect the surface term to give

$$\delta U = -\int \nabla\phi \cdot \delta\boldsymbol{D} \, d^3\boldsymbol{r} = \int \boldsymbol{E} \cdot \delta\boldsymbol{D} \, d^3\boldsymbol{r}. \tag{3.64}$$

This energy increment cannot be integrated unless $\boldsymbol{E}$ is a known function of $\boldsymbol{D}$. Let us adopt the conventional approach and assume that $\boldsymbol{D} = \epsilon_0 \epsilon \boldsymbol{E}$, where the dielectric constant ϵ is independent of the electric field. The change in energy associated with taking the displacement field from zero to $\boldsymbol{D}(\boldsymbol{r})$ at all points in space is given by

$$U = \int_{\boldsymbol{0}}^{\boldsymbol{D}} \delta U = \int_{\boldsymbol{0}}^{\boldsymbol{D}} \int \boldsymbol{E} \cdot \delta\boldsymbol{D} \, d^3\boldsymbol{r}, \tag{3.65}$$

or

$$U = \int \int_0^E \frac{\epsilon_0 \epsilon \, \delta(E^2)}{2} \, d^3\boldsymbol{r} = \frac{1}{2} \int \epsilon_0 \epsilon E^2 \, d^3\boldsymbol{r}, \tag{3.66}$$

which reduces to

$$U = \frac{1}{2} \int \boldsymbol{E} \cdot \boldsymbol{D} \, d^3\boldsymbol{r}. \tag{3.67}$$

Thus, the electrostatic energy density inside a dielectric is given by

$$W = \frac{\boldsymbol{E} \cdot \boldsymbol{D}}{2}. \tag{3.68}$$

This is a standard result which is often quoted in textbooks. Nevertheless, it is important to realize that the above formula is only valid in dielectric media in which the electric displacement $\boldsymbol{D}$ varies *linearly* with the electric field $\boldsymbol{E}$.

3.7 The force density within a dielectric medium

Equation (3.67) was derived by considering a virtual process in which true charges are added to a system of charges and dielectrics which are held fixed, so that no mechanical work is done against physical displacements. Let us now consider a different virtual process in which the physical coordinates of the charges and dielectric are given a virtual displacement $\delta\boldsymbol{r}$ at each point in space, but no free charges are added to the system. Since we are dealing with a conservative system, the energy expression (3.67) can still be employed, despite the fact that it was derived in terms of another virtual process. The variation in the total electrostatic energy δU when the system undergoes a virtual displacement $\delta\boldsymbol{r}$ is related to the electrostatic force density $\boldsymbol{f}$ acting within the dielectric medium via

$$\delta U = -\int \boldsymbol{f}\cdot\delta\boldsymbol{r}\, d^3\boldsymbol{r}. \tag{3.69}$$

If the medium is moving with a velocity field $\boldsymbol{u}$ then the rate at which electrostatic energy is drained from the $\boldsymbol{E}$ and $\boldsymbol{D}$ fields is given by

$$\frac{dU}{dt} = -\int \boldsymbol{f}\cdot\boldsymbol{u}\, d^3\boldsymbol{r}. \tag{3.70}$$

Let us now consider the energy increment due to both a change $\delta\rho_f$ in the free charge distribution and a change $\delta\epsilon$ in the dielectric constant, caused by the displacements. From Eq. (3.67)

$$\delta U = \frac{1}{2\epsilon_0}\int \left[D^2\delta(1/\epsilon) + 2\boldsymbol{D}\cdot\delta\boldsymbol{D}/\epsilon\right]\, d^3\boldsymbol{r}, \tag{3.71}$$

or

$$\delta U = -\frac{\epsilon_0}{2}\int E^2\,\delta\epsilon\, d^3\boldsymbol{r} + \int \boldsymbol{E}\cdot\delta\boldsymbol{D}\, d^3\boldsymbol{r}. \tag{3.72}$$

Here, the first term represents the energy increment due to the change in dielectric constant associated with the virtual displacements, whereas the second term corresponds to the energy increment caused by displacements of the free charges. The second term can be written

$$\int \boldsymbol{E}\cdot\delta\boldsymbol{D}\, d^3\boldsymbol{r} = -\int \nabla\phi\cdot\delta\boldsymbol{D}\, d^3\boldsymbol{r} = \int \phi\,\nabla\cdot\delta\boldsymbol{D}\, d^3\boldsymbol{r} = \int \phi\,\delta\rho_f\, d^3\boldsymbol{r}, \tag{3.73}$$

where surface terms have been neglected. Thus, Eq. (3.72) implies that

$$\frac{dU}{dt} = \int \left(\phi\,\frac{\partial\rho_f}{\partial t} - \frac{\epsilon_0}{2}\,E^2\,\frac{\partial\epsilon}{\partial t}\right)\,d^3\boldsymbol{r}. \tag{3.74}$$

In order to arrive at an expression for the force density $\boldsymbol{f}$ we need to express the time derivatives $\partial\rho/\partial t$ and $\partial\epsilon/\partial t$ in terms of the velocity field $\boldsymbol{u}$. This can be achieved by adopting a dielectric equation of state; *i.e.*, a relation which gives the dependence of the dielectric constant ϵ on the mass density ρ_m. Let us assume that $\epsilon(\rho_m)$ is a known function. It follows that

$$\frac{D\epsilon}{Dt} = \frac{d\epsilon}{d\rho_m}\frac{D\rho_m}{Dt}, \tag{3.75}$$

where

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \boldsymbol{u}\cdot\nabla \tag{3.76}$$

is the total time derivative (*i.e.*, the time derivative in a frame of reference which is locally co-moving with the dielectric.) The hydrodynamic equation of continuity of the dielectric is

$$\frac{\partial\rho_m}{\partial t} + \nabla\cdot(\rho_m\boldsymbol{u}) = 0, \tag{3.77}$$

which implies that

$$\frac{D\rho_m}{Dt} = -\rho_m\nabla\cdot\boldsymbol{u}. \tag{3.78}$$

It follows that

$$\frac{\partial\epsilon}{\partial t} = -\frac{d\epsilon}{d\rho_m}\,\rho_m\nabla\cdot\boldsymbol{u} - \boldsymbol{u}\cdot\nabla\epsilon. \tag{3.79}$$

The conservation equation for the free charges is written

$$\frac{\partial\rho_f}{\partial t} + \nabla\cdot(\rho_f\boldsymbol{u}) = 0. \tag{3.80}$$

Thus, we can express Eq. (3.74) in the form

$$\frac{dU}{dt} = \int \left[-\phi\,\nabla\cdot(\rho_f\boldsymbol{u}) + \frac{\epsilon_0}{2}\,E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\,\nabla\cdot\boldsymbol{u} + \left(\frac{\epsilon_0}{2}\,E^2\,\nabla\epsilon\right)\cdot\boldsymbol{u}\right]\,d^3\boldsymbol{r}. \tag{3.81}$$

Integrating the first term by parts and neglecting any surface contributions, we obtain

$$-\int \phi\,\nabla\cdot(\rho_f\,\boldsymbol{u})\,d^3\boldsymbol{r} = \int \rho_f\,\nabla\phi\cdot\boldsymbol{u}\,d^3\boldsymbol{r}. \tag{3.82}$$

Likewise,

$$\int \frac{\epsilon_0}{2}\,E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\,\nabla\cdot\boldsymbol{u}\,d^3\boldsymbol{r} = -\int \frac{\epsilon_0}{2}\,\nabla\left(E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\right)\cdot\boldsymbol{u}\,d^3\boldsymbol{r}. \tag{3.83}$$

Thus, Eq. (3.81) becomes

$$\frac{dU}{dt} = \int\left[-\rho_f\,\boldsymbol{E} + \frac{\epsilon_0}{2}\,E^2\,\nabla\epsilon - \frac{\epsilon_0}{2}\,\nabla\left(E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\right)\right]\cdot\boldsymbol{u}\,d^3\boldsymbol{r}. \tag{3.84}$$

Comparing with Eq. (3.70), we see that the force density inside the dielectric is given by

$$\boldsymbol{f} = \rho_f\,\boldsymbol{E} - \frac{\epsilon_0}{2}\,E^2\,\nabla\epsilon + \frac{\epsilon_0}{2}\,\nabla\left(E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\right). \tag{3.85}$$

The first term in the above equation is the standard electrostatic force density. The second term represents a force which appears whenever an inhomogeneous dielectric is placed in an electric field. The last term, known as the *electrostriction* term, gives a force acting on a dielectric in an inhomogeneous electric field. Note that the magnitude of the electrostriction force depends explicitly on the dielectric equation of state of the material, through $d\epsilon/d\rho_m$. The electrostriction term gives zero net force acting on any finite region of dielectric if we can integrate over a large enough portion of the dielectric that its extremities lie in a field free region. For this reason the term is frequently omitted, since in the calculation of the total forces acting on dielectric bodies it usually does not contribute. Note, however, that if the electrostriction term is omitted an incorrect pressure variation within the dielectric is obtained, even though the total force is given correctly.

3.8 The Clausius-Mossotti relation

Let us now investigate what a dielectric equation of state actually looks like. Suppose that a dielectric medium is made up of identical molecules which develop a dipole moment

$$\boldsymbol{p} = \alpha\epsilon_0 \boldsymbol{E} \tag{3.86}$$

when placed in an electric field $\boldsymbol{E}$. The constant α is called the *molecular polarizability.* If N is the number density of such molecules then the polarization of the medium is

$$\boldsymbol{P} = N\boldsymbol{p} = N\alpha\epsilon_0 \boldsymbol{E}, \tag{3.87}$$

or

$$\boldsymbol{P} = \frac{N_A\,\rho_m\alpha}{M}\,\epsilon_0 \boldsymbol{E}, \tag{3.88}$$

where ρ_m is the mass density, N_A is Avogadro's number, and M is the molecular weight. But, how does the electric field experienced by an individual molecule relate to the average electric field in the medium? This is not a trivial question since we expect the electric field to vary strongly (on atomic length-scales) inside the dielectric.

Suppose that the dielectric is polarized with a mean electric field $\boldsymbol{E}_0$ which is uniform (on macroscopic length-scales) and directed along the z-axis. Consider one of the molecules which constitute the dielectric. Let us draw a sphere of radius a about this particular molecule. This is intended to represent the boundary between the microscopic and the macroscopic range of phenomena affecting the molecule. We shall treat the dielectric outside the sphere as a continuous medium and the dielectric inside the sphere as a collection of polarized molecules. According to Eq. (3.29) there is a polarization surface charge of magnitude

$$\sigma_{\rm pol} = -P\cos\theta \tag{3.89}$$

on the inside of the sphere, where (r, θ, φ) are spherical polar coordinates, and $\boldsymbol{P} = P\,\hat{\boldsymbol{z}} = \epsilon_0(\epsilon - 1)E_0\,\hat{\boldsymbol{z}}$ is the uniform polarization of the dielectric. The magnitude of E_z at the molecule due to the surface charge is

$$E_z = -\frac{1}{4\pi\epsilon_0}\int \frac{\sigma_{\rm pol}\cos\theta}{a^2}\,dS, \tag{3.90}$$

where $dS = 2\pi a^2\,\sin\theta\,d\theta$ is a surface element of the sphere. It follows that

$$E_z = \frac{P}{2\epsilon_0}\int_0^\pi \cos^2\theta\,\sin\theta\,d\theta = \frac{P}{3\epsilon_0}. \tag{3.91}$$

It is easily demonstrated that $E_\theta = E_\varphi = 0$ at the molecule. Thus, the field at the molecule due to the surface charges on the sphere is

$$\boldsymbol{E} = \frac{\boldsymbol{P}}{3\epsilon_0}. \tag{3.92}$$

The field due to the individual molecules within the sphere is obtained by summing over the dipole fields of these molecules. The electric field at a distance $\boldsymbol{r}$ from a dipole $\boldsymbol{p}$ is

$$\boldsymbol{E} = -\frac{1}{4\pi\epsilon_0}\left[\frac{\boldsymbol{p}}{r^3} - \frac{3(\boldsymbol{p}\cdot\boldsymbol{r})\boldsymbol{r}}{r^5}\right]. \tag{3.93}$$

It is assumed that the dipole moment of each molecule within the sphere is the same, and also that the molecules are evenly distributed throughout the sphere. This being the case, the value of E_z at the molecule due to all of the other molecules within in the sphere,

$$E_z = -\frac{1}{4\pi\epsilon_0}\sum_{\text{mols}}\left[\frac{p_z}{r^3} - \frac{3(p_x\,xz + p_y\,yz + p_z\,z^2)}{r^5}\right], \tag{3.94}$$

is zero, since

$$\sum_{\text{mols}} x^2 = \sum_{\text{mols}} y^2 = \sum_{\text{mols}} z^2 = \frac{1}{3}\sum_{\text{mols}} r^2 \tag{3.95}$$

and

$$\sum_{\text{mols}} xy = \sum_{\text{mols}} yz = \sum_{\text{mols}} zx = 0. \tag{3.96}$$

It is easily seen that $E_\theta = E_\varphi = 0$. Hence, the electric field at the molecule due to the other molecules within the sphere vanishes.

It is clear that the net electric field seen by an individual molecule is

$$\boldsymbol{E} = \boldsymbol{E}_0 + \frac{\boldsymbol{P}}{3\epsilon_0}. \tag{3.97}$$

This is *larger* than the average electric field $\boldsymbol{E}_0$ in the dielectric. The above analysis indicates that this effect is ascribable to the long range (rather than the short range) interactions of the molecule with the other molecules in the medium. Making use of Eq. (3.88) and the definition $\boldsymbol{P} = \epsilon_0(\epsilon - 1)\boldsymbol{E}_0$, we obtain

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{N_A\,\rho_m \alpha}{3M}. \tag{3.98}$$

This is called the *Clausius-Mossotti* relation. This formula is found to work pretty well for a wide class of dielectric liquids and gases. The Clausius-Mossotti relation yields

$$\frac{d\epsilon}{d\rho_m} = \frac{(\epsilon - 1)(\epsilon + 2)}{3\rho_m}. \tag{3.99}$$

3.9 Dielectric liquids in electrostatic fields

Consider the behaviour of an uncharged dielectric liquid placed in an electrostatic field. If p is the pressure in the liquid when in equilibrium with the electrostatic force density $\boldsymbol{f}$, then force balance requires that

$$\nabla p = \boldsymbol{f}. \tag{3.100}$$

It follows from Eq. (3.85) that

$$\nabla p = -\frac{\epsilon_0}{2}\,E^2\,\nabla\epsilon + \frac{\epsilon_0}{2}\,\nabla\!\left(E^2\,\frac{d\epsilon}{d\rho_m}\,\rho_m\right) = \frac{\epsilon_0\rho_m}{2}\,\nabla\!\left(E^2\,\frac{d\epsilon}{d\rho_m}\right). \tag{3.101}$$

We can integrate this equation to give

$$\int_{p_1}^{p_2}\frac{dp}{\rho_m} = \frac{\epsilon_0}{2}\left(\left[E^2\,\frac{d\epsilon}{d\rho_m}\right]_2 - \left[E^2\,\frac{d\epsilon}{d\rho_m}\right]_1\right), \tag{3.102}$$

where 1 and 2 refer to two general points in the liquid. Here, it is assumed that the liquid possesses an equation of state, so that $p = p(\rho_m)$. If the liquid is essentially incompressible ($\rho_m \simeq$ constant) then

$$p_2 - p_1 = \frac{\epsilon_0\rho_m}{2}\left[E^2\,\frac{d\epsilon}{d\rho_m}\right]_1^2. \tag{3.103}$$

Finally, if the liquid obeys the Clausius-Mossotti relation then

$$p_2 - p_1 = \left[\frac{\epsilon_0 E^2}{2}\frac{(\epsilon-1)(\epsilon+2)}{3}\right]_1^2. \tag{3.104}$$

According to Eqs. (3.54) and (3.104), if a sphere of dielectric liquid is placed in a uniform electric field $\boldsymbol{E}_0$ then the pressure inside the liquid takes the constant value

$$p = \frac{3}{2}\,\epsilon_0 {E_0}^2\,\frac{\epsilon-1}{\epsilon+2}. \tag{3.105}$$

It is clear that the electrostatic forces acting on the dielectric are all concentrated at the edge of the sphere and are directed radially inwards; *i.e.*, the dielectric is *compressed* by the external electric field. This is a somewhat surprising result since the electrostatic forces acting on a rigid conducting sphere are concentrated at the edge of the sphere but are directed radially outwards. We might expect these two cases to give the same result in the limit $\epsilon \to \infty$. The reason that this does not occur is because a dielectric liquid is slightly compressible and is, therefore, subject to an electrostriction force. There is no electrostriction force for the case of a completely rigid body. In fact, the force density inside a rigid dielectric (for which $\nabla\cdot\boldsymbol{u} = 0$) is given by Eq. (3.85) with the third term (the electrostriction term) missing. It is easily seen that the force exerted by an electric field on a rigid dielectric is directed outwards and approaches that exerted on a rigid conductor in the limit $\epsilon \to 0$.

As is well known, when a pair of charged (parallel plane) capacitor plates are dipped into a dielectric liquid the liquid is drawn up between the plates to some extent. Let us examine this effect. We can, without loss of generality, assume that the transition from dielectric to vacuum takes place in a continuous manner. Consider the electrostatic pressure difference between a point A lying just above the surface of the liquid in between the plates and a point B lying just above the surface of the liquid well away from the capacitor where $E = 0$. The pressure difference is given by

$$p_A - p_B = -\int_A^B \boldsymbol{f}\cdot d\boldsymbol{l}. \tag{3.106}$$

Note, however, that the Clausius-Mossotti relation yields $d\epsilon/d\rho_m = 0$ at both A and B, since $\epsilon = 1$ in a vacuum (see Eq. (3.99)). Thus, it is clear from Eq. (3.85) that the electrostriction term makes no contribution to the line integral (3.106). It follows that

$$p_A - p_B = \frac{\epsilon_0}{2} \int_A^B E^2 \nabla\epsilon\cdot d\boldsymbol{l}. \tag{3.107}$$

The only contribution to this integral comes from the vacuum/dielectric interface in the vicinity of point A (since ϵ is constant inside the liquid, and $E = 0$ in the vicinity of point B). Suppose that the electric field at point A has normal and tangential (to the surface) components E_n and E_t, respectively. Making use of the boundary conditions that E_t and ϵE_n are constant across a vacuum/dielectric interface, we obtain

$$p_A - p_B = \frac{\epsilon_0}{2}\left[{E_t}^2(\epsilon - 1) + \epsilon^2 {E_n}^2(\epsilon) \int_1^\epsilon \frac{d\epsilon}{\epsilon^2}\right], \tag{3.108}$$

giving

$$p_A - p_B = \frac{\epsilon_0\,(\epsilon - 1)}{2}\left[{E_t}^2 + \frac{{E_n}^2}{\epsilon}\right]. \tag{3.109}$$

This electrostatic pressure difference can be equated to the hydrostatic pressure difference $\rho_m\,g\,h$ to determine the height h that the liquid rises between the plates. At first sight, the above analysis appears to suggest that the dielectric liquid is drawn upward by a surface force acting on the vacuum/dielectric interface in the region between the plates. In fact, this is far from being the case. A brief examination of Eq. (3.104) shows that this surface force is actually directed downwards. According to Eq. (3.85), the force which causes the liquid to rise between the plates is a volume force which develops in the region of non-uniform electric field at the base of the capacitor, where the field splays out between the plates. Thus, although we can determine the height to which the fluid rises between the plates without reference to the electrostriction force, it is, somewhat paradoxically, this force which is actually responsible for supporting the liquid against gravity.

Let us consider another paradox concerning the electrostatic forces exerted in a dielectric medium. Suppose that we have two charges embedded in a uniform dielectric ϵ. The electric field generated by each charge is the same as that in vacuum, except that it is reduced by a factor ϵ. Therefore, we expect that the force exerted by one charge on another is the same as that in vacuum, except that it is also reduced by a factor ϵ. Let us examine how this reduction in force comes about. Consider a simple example. Suppose that we take a parallel plate capacitor and insert a block of solid dielectric between the plates. Suppose, further, that there is a small vacuum gap between the faces of the block and each of the capacitor plates. Let $\pm\sigma$ be the surface charge densities on each of the capacitor plates, and let $\pm\sigma_p$ be the polarization charge densities which develop on the outer faces of the intervening dielectric block. The two layers of polarization charge produce equal and opposite electric fields on each plate, and their effects therefore cancel each other. Thus, from the point of view of electrical interaction alone there would appear to be no change in the force exerted by one capacitor plate on the other when a dielectric slab is placed between them (assuming that σ remains constant during this process). That is, the force per unit area (which is attractive) remains

$$f_s = \frac{\sigma^2}{2\epsilon_0}. \tag{3.110}$$

However, in experiments in which a capacitor is submerged in a dielectric liquid the force per unit area exerted by one plate on another is observed to decrease to

$$f_s = \frac{\sigma^2}{2\epsilon_0\epsilon}. \tag{3.111}$$

This apparent paradox can be explained by taking into account the difference in liquid pressure in the field filled space between the plates and the field free region outside the capacitor. This pressure difference is balanced by internal elastic forces in the case of the solid dielectric discussed earlier, but is transmitted to the plates in the case of the liquid. We can compute the pressure difference between a point A on the inside surface of one of the capacitor plates and a point B on the outside surface of the same plate using Eq. (3.107). If we neglect end effects then the electric field is normal to the plates in the region between the plates and is zero everywhere else. Thus, the only contribution to the line integral (3.107) comes from the plate/dielectric interface in the vicinity of point A. Using Eq. (3.109), we find that

$$p_A - p_B = \frac{\epsilon_0}{2}\left(1 - \frac{1}{\epsilon}\right) E^2 = \frac{\sigma^2}{2\epsilon_0}\left(1 - \frac{1}{\epsilon}\right), \tag{3.112}$$

where E is the normal field strength between the plates in the absence of dielectric. The sum of this pressure force and the purely electrical force (3.110) yields a net attractive force per unit area

$$f_s = \frac{\sigma^2}{2\epsilon_0\epsilon} \tag{3.113}$$

acting between the plates. Thus, any decrease in the forces exerted by charges on one another when they are immersed or embedded in some dielectric medium can only be understood in terms of mechanical forces transmitted between these charges by the medium itself.

3.10 Magnetization

All matter is built up out of atoms, and each atom consists of electrons in motion. The currents associated with this motion are termed *atomic currents*. Each atomic current is a tiny closed circuit of atomic dimensions, and may therefore be appropriately described as a magnetic dipole. If the atomic currents of a given atom all flow in the same plane then the atomic dipole moment is directed normal to the plane (in the sense given by the right-hand rule) and its magnitude is the product of the total circulating current and the area of the current loop. More generally, if $\boldsymbol{j}(\boldsymbol{r})$ is the atomic current density at the point $\boldsymbol{r}$ then the magnetic moment of the atom is

$$\boldsymbol{m} = \frac{1}{2}\int \boldsymbol{r}\wedge\boldsymbol{j}\,d^3\boldsymbol{r}, \tag{3.114}$$

where the integral is over the volume of the atom. If there are N such atoms or molecules per unit volume then the *magnetization* $\boldsymbol{M}$ (*i.e.*, the magnetic dipole moment per unit volume) is given by $\boldsymbol{M} = N\boldsymbol{m}$. More generally,

$$\boldsymbol{M}(\boldsymbol{r}) = \sum_i N_i\langle\boldsymbol{m}_i\rangle, \tag{3.115}$$

where $\langle\boldsymbol{m}_i\rangle$ is the average magnetic dipole moment of the ith type of molecule in the vicinity of point $\boldsymbol{r}$, and N_i is the average number of such molecules per unit volume at $\boldsymbol{r}$.

Consider a general medium which is made up of molecules which are polarizable and possess a net magnetic moment. It is easily demonstrated that any circulation in the magnetization field $\boldsymbol{M}(\boldsymbol{r})$ gives rise to an effective current density $\boldsymbol{j}_m$ in the medium. In fact,

$$\boldsymbol{j}_m = \nabla \wedge \boldsymbol{M}. \tag{3.116}$$

This current density is called the *magnetization current density*, and is usually distinguished from the *true current density* $\boldsymbol{j}_t$, which represents the convection of free charges in the medium. In fact, there is a third type of current called a *polarization current*, which is due to the apparent convection of bound charges. It is easily demonstrated that the polarization current density $\boldsymbol{j}_p$ is given by

$$\boldsymbol{j}_p = \frac{\partial \boldsymbol{P}}{\partial t}. \tag{3.117}$$

Thus, the total current density $\boldsymbol{j}$ in the medium is given by

$$\boldsymbol{j} = \boldsymbol{j}_t + \nabla \wedge \boldsymbol{M} + \frac{\partial \boldsymbol{P}}{\partial t}. \tag{3.118}$$

It must be emphasized that all terms on the right-hand side of this equation represent real physical currents, although only the first term is due to the motion of real charges (over more than atomic dimensions).

The Ampère-Maxwell equation takes the form

$$\nabla \wedge \boldsymbol{B} = \mu_0 \, \boldsymbol{j} + \mu_0 \epsilon_0 \, \frac{\partial \boldsymbol{E}}{\partial t}, \tag{3.119}$$

which can also be written

$$\nabla \wedge \boldsymbol{B} = \mu_0 \, \boldsymbol{j}_t + \mu_0 \nabla \wedge \boldsymbol{M} + \mu_0 \frac{\partial \boldsymbol{D}}{\partial t}, \tag{3.120}$$

where use has been made of the definition $\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}$. The above expression can be rearranged to give

$$\nabla \wedge \boldsymbol{H} = \boldsymbol{j}_t + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{3.121}$$

where

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M} \tag{3.122}$$

is termed the *magnetic intensity*, and has the same dimensions as $\boldsymbol{M}$ (*i.e.*, magnetic dipole moment per unit volume). In a steady-state situation, Stokes's theorem tell us that

$$\oint_C \boldsymbol{H}\cdot d\boldsymbol{l} = \int_S \boldsymbol{j}_t\cdot d\boldsymbol{S}. \tag{3.123}$$

In other words, the line integral of $\boldsymbol{H}$ around some closed curve is equal to the flux of true current through any surface attached to that curve. Unlike the magnetic field $\boldsymbol{B}$ (which specifies the force $e\,\boldsymbol{v}\wedge\boldsymbol{B}$ acting on a charge e moving with velocity $\boldsymbol{v}$) or the magnetization $\boldsymbol{M}$ (the magnetic dipole moment per unit volume), the magnetic intensity $\boldsymbol{H}$ has no clear physical meaning. The only reason for introducing it is that it enables us to calculate fields in the presence of magnetic materials without first having to know the distribution of magnetization currents. However, this is only possible if we possess a constitutive relation connecting $\boldsymbol{B}$ and $\boldsymbol{H}$.

3.11 Magnetic susceptibility and permeability

In a large class of materials there exists an approximately linear relationship between $\boldsymbol{M}$ and $\boldsymbol{H}$. If the material is isotropic then

$$\boldsymbol{M} = \chi_m \boldsymbol{H}, \tag{3.124}$$

where χ_m is called the magnetic susceptibility. If χ_m is positive the material is called *paramagnetic*, and the magnetic field is strengthened by the presence of the material. If χ_m is negative then the material is *diamagnetic* and the magnetic field is weakened in the presence of the material. The magnetic susceptibilities of paramagnetic and diamagnetic materials are generally extremely small. A few sample values are given in Table 1.[10]

A linear relationship between $\boldsymbol{M}$ and $\boldsymbol{H}$ also implies a linear relationship between $\boldsymbol{B}$ and $\boldsymbol{H}$. In fact, we can write

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{3.125}$$

[10] Data obtained from the *Handbook of Chemistry and Physics*, Chemical Rubber Company Press, Baca Raton, FL

Material	χ_m
Aluminium	2.3×10^{-5}
Copper	-0.98×10^{-5}
Diamond	-2.2×10^{-5}
Tungsten	6.8×10^{-5}
Hydrogen (1 atm)	-0.21×10^{-8}
Oxygen (1 atm)	209.0×10^{-8}
Nitrogen (1 atm)	-0.50×10^{-8}

Table 1: Magnetic susceptibilities of some paramagnetic and diamagnetic materials at room temperature

where

$$\mu = \mu_0(1 + \chi_m) \tag{3.126}$$

is termed the magnetic *permeability* of the material in question. (Likewise, μ_0 is termed the *permeability of free space*.) It is clear from Table 1 that the permeabilities of common diamagnetic and paramagnetic materials do not differ substantially from that of free space. In fact, to all intents and purposes the magnetic properties of such materials can be safely neglected (*i.e.*, $\mu = \mu_0$).

3.12 Ferromagnetism

There is, however, a third class of magnetic materials called *ferromagnetic* materials. Such materials are characterized by a possible permanent magnetization, and generally have a profound effect on magnetic fields (*i.e.*, $\mu/\mu_0 \gg 1$). Unfortunately, ferromagnetic materials do *not* exhibit a linear dependence between $\boldsymbol{M}$ and $\boldsymbol{H}$ or $\boldsymbol{B}$ and $\boldsymbol{H}$, so that we cannot employ Eqs. (3.124) and (3.125) with constant values of χ_m and μ. It is still expedient to use Eq. (3.125) as the definition of μ, with $\mu = \mu(\boldsymbol{H})$, however this practice can lead to difficulties under certain circumstances. The permeability of a ferromagnetic material, as defined by Eq. (3.125), can vary through the entire range of possible values from zero to infinity, and may be either positive or negative. The most sensible approach is to consider each problem involving ferromagnetic materials separately, try to

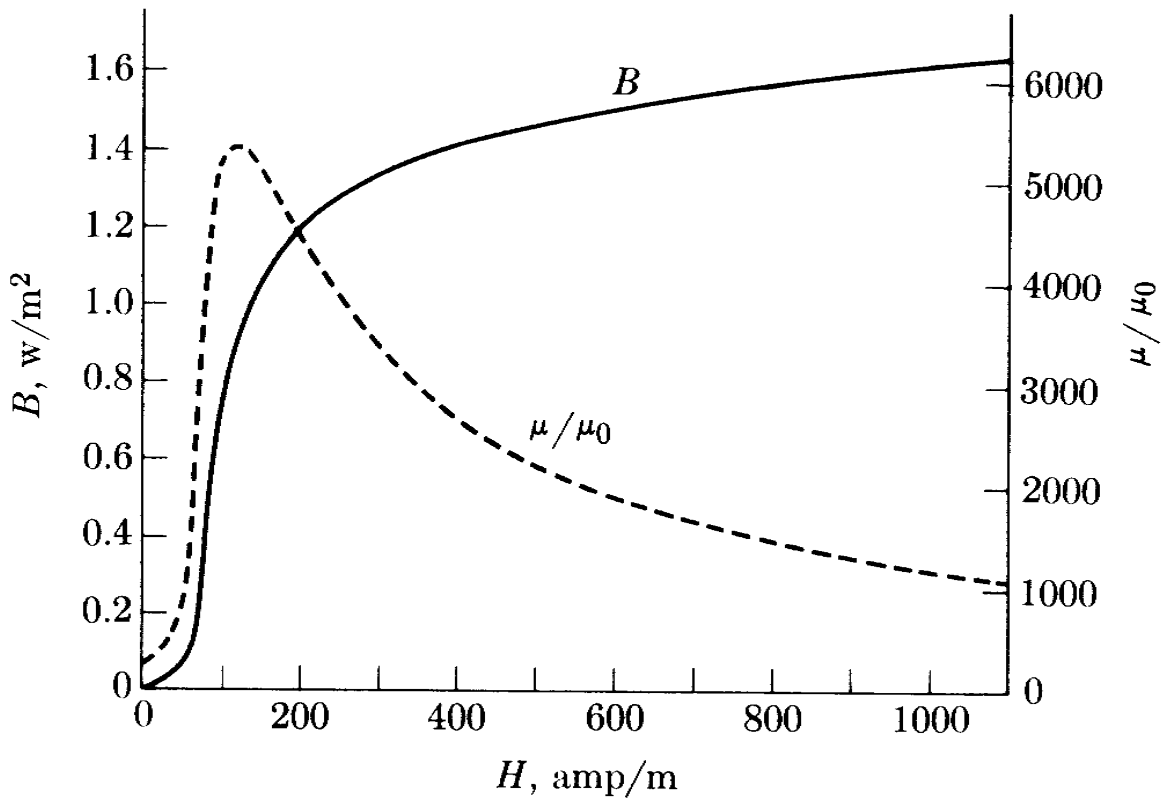


Figure 2: *Magnetization curve and relative permeability of commercial iron (annealed)*

determine which region of the $\boldsymbol{B}$-$\boldsymbol{H}$ diagram is important for the particular case in hand, and then make approximations appropriate to this region.

First, let us consider an unmagnetized sample of ferromagnetic material. If the magnetic intensity, which is initially zero, is increased *monotonically*, then the $\boldsymbol{B}$-$\boldsymbol{H}$ relationship traces out a curve such as that shown in Fig. 2. This is called a *magnetization curve*. It is evident that the permeabilities μ derived from the curve (according to the rule $\mu = B/H$) are always positive, and show a wide range of values. The maximum permeability occurs at the "knee" of the curve. In some materials this maximum permeability is as large as $10^5\,\mu_0$. The reason for the knee in the curve is that the magnetization $\boldsymbol{M}$ reaches a maximum value in the material, so that

$$\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}) \tag{3.127}$$

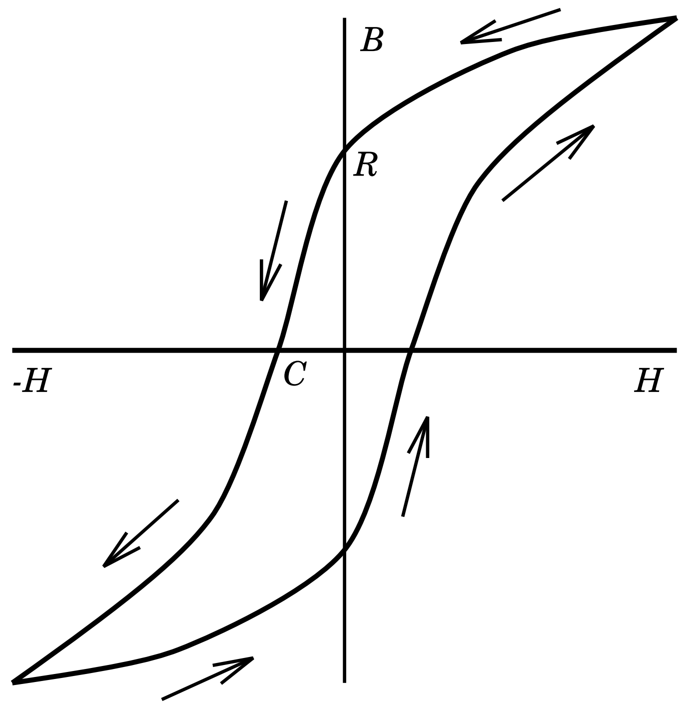


Figure 3: *Typical hysteresis loop of a ferromagnetic material*

continues to increase at large $\boldsymbol{H}$ only because of the $\mu_0 \boldsymbol{H}$ term. The maximum value of $\boldsymbol{M}$ is called the *saturation magnetization* of the material.

Next, consider a ferromagnetic sample magnetized by the above procedure. If the magnetic intensity $\boldsymbol{H}$ is decreased, the $\boldsymbol{B}$-$\boldsymbol{H}$ relation does not follow back down the curve of Fig. 2, but instead moves along a new curve, shown in Fig. 3, to the point R. The magnetization, once established, does not disappear with the removal of $\boldsymbol{H}$. In fact, it takes a reversed magnetic intensity to reduce the magnetization to zero. If $\boldsymbol{H}$ continues to build up in the reversed direction, then $\boldsymbol{M}$ (and hence $\boldsymbol{B}$) becomes increasingly negative. Finally, when $\boldsymbol{H}$ increases again the operating point follows the lower curve of Fig. 3. Thus, the $\boldsymbol{B}$-$\boldsymbol{H}$ curve for increasing $\boldsymbol{H}$ is quite different to that for decreasing $\boldsymbol{H}$. This phenomenon is known as *hysteresis.*

The curve of Fig. 3 is called the hysteresis loop of the material in question. The value of $\boldsymbol{B}$ at the point R is called the retentivity or *remanence.* The magnitude of $\boldsymbol{H}$ at the point C is called the *coercivity.* It is evident that μ is negative in the second and fourth quadrants of the diagram and positive in the first and third quadrants. The shape of the hysteresis loop depends not only on the nature of the ferromagnetic material but also on the maximum value of $\boldsymbol{H}$ to which the material is subjected. However, once this maximum value, $\boldsymbol{H}_{\rm max}$, becomes sufficient to produce saturation in the material the hysteresis loop does not change shape with any further increase in $\boldsymbol{H}_{\rm max}$.

Ferromagnetic materials are used either to channel magnetic flux (*e.g.*, around transformer circuits) or as sources of magnetic field (permanent magnets). For use as a permanent magnet, the material is first magnetized by placing it in a strong magnetic field. However, once the magnet is removed from the external field it is subject to a demagnetizing $\boldsymbol{H}$. Thus, it is vitally important that a permanent magnet should possess both a large remanence and a large coercivity. As will become clear later on, it is generally a good idea for the ferromagnetic materials used to channel magnetic flux around transformer circuits to possess small remanences and small coercivities.

3.13 Boundary conditions for $\boldsymbol{B}$ and $\boldsymbol{H}$

What are the matching conditions for $\boldsymbol{B}$ and $\boldsymbol{H}$ at the boundary between two media? The governing equations for a steady state situation are

$$\nabla\cdot\boldsymbol{B} = 0, \tag{3.128}$$

and

$$\nabla\wedge\boldsymbol{H} = \boldsymbol{j}_t. \tag{3.129}$$

Integrating Eq. (3.128) over a Gaussian pill-box enclosing part of the boundary surface between the two media gives

$$(\boldsymbol{B}_2 - \boldsymbol{B}_1)\cdot\boldsymbol{n}_{21} = 0, \tag{3.130}$$

where $\boldsymbol{n}_{21}$ is the unit normal to this surface directed from medium 1 to medium 2. Integrating Eq. (3.129) around a small loop which straddles the boundary surface yields

$$(\boldsymbol{H}_2 - \boldsymbol{H}_1)\wedge\boldsymbol{n}_{21} = 0, \tag{3.131}$$

assuming that there is no true current sheet flowing in this surface. In general, there is a magnetization current sheet flowing in the boundary surface whose density is given by

$$\boldsymbol{J}_m = \boldsymbol{n}_{21} \wedge (\boldsymbol{M}_2 - \boldsymbol{M}_1), \tag{3.132}$$

where $\boldsymbol{M}_1$ is the magnetization in medium 1 at the boundary, *etc.* It is clear that the normal component of the magnetic field and the tangential component of the magnetic intensity are both continuous across any boundary between magnetic materials.

3.14 Permanent ferromagnets

Let us consider the magnetic field generated by a distribution of permanent ferromagnets. Let us suppose that the magnets in question are sufficiently "hard" that their magnetization is essentially independent of the applied field for moderate field strengths. Such magnets can be treated as if they contain a fixed, specified magnetization $\boldsymbol{M}(\boldsymbol{r})$.

Let us assume that there are no true currents in the problem, so that $\boldsymbol{j}_t = 0$. Let us also assume that we are dealing with a steady state situation. Under these circumstances Eq. (3.121) reduces to

$$\nabla \wedge \boldsymbol{H} = 0. \tag{3.133}$$

It follows that we can write

$$\boldsymbol{H} = -\nabla \phi_m, \tag{3.134}$$

where ϕ_m is called the *magnetic scalar potential.* Now, we know that

$$\nabla\cdot\boldsymbol{B} = \mu_0 \nabla\cdot(\boldsymbol{H} + \boldsymbol{M}) = 0. \tag{3.135}$$

Equations (3.134) and (3.135) combine to give

$$\nabla^2 \phi_m = -\rho_m, \tag{3.136}$$

where

$$\rho_m = -\nabla\cdot\boldsymbol{M}. \tag{3.137}$$

Thus, the *magnetostatic* field $\boldsymbol{H}$ is determined by Poisson's equation. We can think of ρ_m as an *effective magnetic charge density.* Of course, this magnetic charge has no physical reality. We have only introduced it in order to make the problem of the steady magnetic field generated by a set of permanent magnets look formally the same as that of the steady electric field generated by a distribution of charges.

The unique solution of Poisson's equation, subject to sensible boundary conditions at infinity, is well known:

$$\phi_m(\boldsymbol{r}) = \frac{1}{4\pi}\int \frac{\rho_m(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.138}$$

This yields

$$\phi_m(\boldsymbol{r}) = -\frac{1}{4\pi}\int \frac{\nabla'\cdot\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.139}$$

If the magnetization field $\boldsymbol{M}(\boldsymbol{r})$ is well behaved and localized we can integrate by parts to obtain

$$\phi_m(\boldsymbol{r}) = \frac{1}{4\pi}\int \boldsymbol{M}(\boldsymbol{r}')\cdot\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)\,d^3\boldsymbol{r}'. \tag{3.140}$$

Now

$$\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) = -\nabla\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right), \tag{3.141}$$

so our expression for the magnetic potential can be written

$$\phi_m(\boldsymbol{r}) = -\frac{1}{4\pi}\nabla\cdot\int \frac{\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.142}$$

Far from the region of non-vanishing magnetization the potential reduces to

$$\phi_m(\boldsymbol{r}) \simeq -\nabla\left(\frac{1}{4\pi r}\right)\cdot\int \boldsymbol{M}(\boldsymbol{r}')\,d^3\boldsymbol{r}' \simeq \frac{\boldsymbol{m}\cdot\boldsymbol{r}}{4\pi r^3}, \tag{3.143}$$

where $\boldsymbol{m} = \int \boldsymbol{M}\,d^3\boldsymbol{r}$ is the total magnetic moment of the distribution. This is the scalar potential of a dipole. Thus, an arbitrary localized distribution of magnetization asymptotically produces a dipole magnetic field whose strength is determined by the net magnetic moment of the distribution.

It is often a good approximation to treat the magnetization field $\boldsymbol{M}(\boldsymbol{r})$ as a discontinuous quantity. In other words, $\boldsymbol{M}(\boldsymbol{r})$ is specified throughout the "hard" ferromagnets in question, and suddenly falls to zero at the boundaries of these magnets. Integrating Eq. (3.137) over a Gaussian pill-box which straddles one of these boundaries leads to the conclusion that there is an *effective magnetic surface charge density*,

$$\sigma_m = \boldsymbol{n}\cdot\boldsymbol{M}, \tag{3.144}$$

on the surface of the ferromagnets, where $\boldsymbol{M}$ is the surface magnetization, and $\boldsymbol{n}$ is a unit outward directed normal to the surface. Under these circumstances Eq. (3.139) yields

$$\phi_m(\boldsymbol{r}) = -\frac{1}{4\pi}\int_V \frac{\nabla'\cdot\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}' + \frac{1}{4\pi}\int_S \frac{\boldsymbol{M}(\boldsymbol{r}')\cdot d\boldsymbol{S}'}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{3.145}$$

where V represents the volume occupied by the magnets and S is the bounding surface to V. Here, $d\boldsymbol{S}$ is an outward directed volume element to S. It is clear that Eq. (3.145) consists of a volume integral involving the volume magnetic charges $\rho_m = -\nabla\cdot\boldsymbol{M}$ and a surface integral involving the surface magnetic charges $\sigma_m = \boldsymbol{n}\cdot\boldsymbol{M}$. If the magnetization is uniform throughout the volume V then the first term in the above expression vanishes and only the surface integral makes a contribution.

We can also write $\boldsymbol{B} = \nabla\wedge\boldsymbol{A}$ in order to satisfy $\nabla\cdot\boldsymbol{B} = 0$ automatically. It follows from Eqs. (3.121) and (3.122) that

$$\nabla\wedge\boldsymbol{H} = \nabla\wedge(\boldsymbol{B}/\mu_0 - \boldsymbol{M}) = 0, \tag{3.146}$$

which gives

$$\nabla^2\boldsymbol{A} = -\mu_0\,\boldsymbol{j}_m, \tag{3.147}$$

since $\boldsymbol{j}_m = \nabla\wedge\boldsymbol{M}$. The unique solution to Eq. (3.147), subject to sensible boundary conditions at infinity, is very well known:

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}_m(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.148}$$

Thus,

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int \frac{\nabla' \wedge \boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \, d^3\boldsymbol{r}'. \tag{3.149}$$

If the magnetization field is discontinuous it is necessary to add a surface integral to the above expression. It is straightforward to show that

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int_V \frac{\nabla' \wedge \boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \, d^3\boldsymbol{r}' + \frac{\mu_0}{4\pi} \int_S \frac{\boldsymbol{M}(\boldsymbol{r}') \wedge d\boldsymbol{S}'}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{3.150}$$

It is clear that the above expression consists of a volume integral involving the volume magnetization currents $\boldsymbol{j}_m = \nabla \wedge \boldsymbol{M}$ and a surface integral involving the surface magnetization currents $\boldsymbol{J}_m = \boldsymbol{M} \wedge \boldsymbol{n}$ (see Eq. (3.132)). If the magnetization field is uniform throughout V then only the surface integral makes a contribution.

3.15 A uniformly magnetized sphere

Consider a sphere of radius a, with a uniform permanent magnetization $\boldsymbol{M} = M_0 \, \hat{\boldsymbol{z}}$, surrounded by a vacuum region. The simplest way of solving this problem is in terms of the scalar magnetic potential introduced in Eq. (3.134). From Eqs. (3.136) and (3.137), it is clear that ϕ_m satisfies Laplace's equation,

$$\nabla^2 \phi_m = 0, \tag{3.151}$$

since there is zero volume magnetic charge density in a vacuum or a uniformly magnetized magnetic medium. However, according to Eq. (3.144), there is a magnetic surface charge density,

$$\sigma_m = \hat{\boldsymbol{r}} \cdot \boldsymbol{M} = M_0 \cos\theta, \tag{3.152}$$

on the surface of the sphere. One of the matching conditions at the surface of the sphere is that the tangential component of $\boldsymbol{H}$ must be continuous. It follows from Eq. (3.134) that the scalar magnetic potential must be continuous at $r = a$, so that

$$\phi_m(r = a_+) = \phi_m(r = a_-). \tag{3.153}$$

Integrating Eq. (3.136) over a Gaussian pill-box straddling the surface of the sphere yields

$$\left[\frac{\partial\phi_m}{\partial r}\right]_{r=a-}^{r=a+} = -\sigma_m = -M_0\cos\theta. \tag{3.154}$$

In other words, the magnetic charge sheet on the surface of the sphere gives rise to a discontinuity in the radial gradient of the magnetic scalar potential at $r = a$.

The most general axisymmetric solution to Eq. (3.151) which satisfies physical boundary conditions at $r = a$ and $r = \infty$ is

$$\phi_m(r,\theta) = \sum_{l=0}^{\infty} A_l\, r^l\, P_l(\cos\theta) \tag{3.155}$$

for $r < a$, and

$$\phi_m(r,\theta) = \sum_{l=0}^{\infty} B_l\, r^{-(l+1)}\, P_l(\cos\theta) \tag{3.156}$$

for $r \geq a$. The boundary condition (3.153) yields

$$B_l = A_l\, a^{2l+1} \tag{3.157}$$

for all l. The boundary condition (3.154) gives

$$-\frac{(l+1)\,B_l}{a^{l+2}} - l A_l\, a^{l-1} = -M_0\,\delta_{l1} \tag{3.158}$$

for all l, since $P_l(\cos\theta) = \cos\theta$. It follows that

$$A_l = B_l = 0 \tag{3.159}$$

for $l \neq 1$, and

$$A_1 = \frac{M_0}{3}, \tag{3.160a}$$

$$B_1 = \frac{M_0 a^3}{3}. \tag{3.160b}$$

Thus,

$$\phi_m(r,\theta) = \frac{M_0 a^2}{3}\,\frac{r}{a^2}\cos\theta \tag{3.161}$$

for $r < a$, and

$$\phi_m(r,\theta) = \frac{M_0 a^2}{3}\,\frac{a}{r^2}\cos\theta \tag{3.162}$$

for $r \geq a$. Since there is a uniqueness theorem associated with Poisson's equation, we can be sure that this axisymmetric potential is the only solution to the problem which satisfies physical boundary conditions at $r = 0$ and infinity.

In the vacuum region outside the sphere

$$\boldsymbol{B} = \mu_0 \boldsymbol{H} = -\mu_0 \nabla\phi_m. \tag{3.163}$$

It is easily demonstrated from Eq. (3.162) that

$$\boldsymbol{B}(r > a) = \frac{\mu_0}{4\pi}\left[-\frac{\boldsymbol{m}}{r^3} + \frac{3(\boldsymbol{m}\cdot\boldsymbol{r})\,\boldsymbol{r}}{r^5}\right], \tag{3.164}$$

where

$$\boldsymbol{m} = \frac{4}{3}\,\pi a^3 \boldsymbol{M}. \tag{3.165}$$

This, of course, is the magnetic field of a magnetic dipole $\boldsymbol{m}$. Not surprisingly, the net dipole moment of the sphere is equal to the integral of the magnetization $\boldsymbol{M}$ (which is the dipole moment per unit volume) over the volume of the sphere.

Inside the sphere we have $\boldsymbol{H} = -\nabla\phi_m$ and $\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M})$, giving

$$\boldsymbol{H} = -\frac{\boldsymbol{M}}{3}, \tag{3.166}$$

and

$$\boldsymbol{B} = \frac{2}{3}\,\mu_0 \boldsymbol{M}. \tag{3.167}$$

Thus, both the $\boldsymbol{H}$ and $\boldsymbol{B}$ fields are uniform inside the sphere. Note that the magnetic intensity is oppositely directed to the magnetization. In other words, the $\boldsymbol{H}$ field acts to *demagnetize* the sphere. How successful it is at achieving

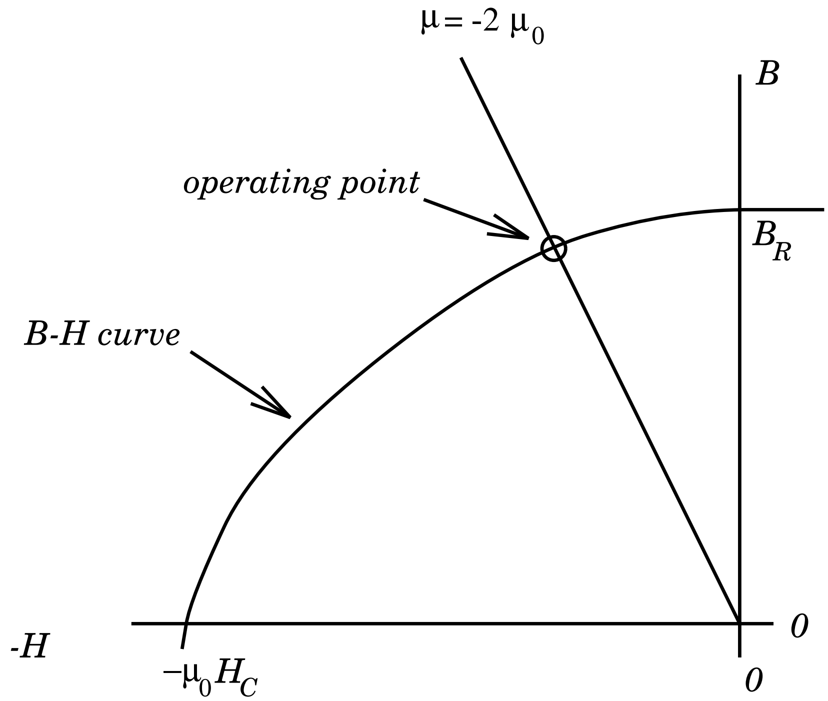


Figure 4: *Schematic demagnetization curve for a permanent magnet*

this depends on the shape of the hysteresis curve in the negative H and positive B quadrant. This curve is sometimes called the *demagnetization curve* of the magnetic material which makes up the sphere. Figure 4 shows a schematic demagnetization curve. The curve is characterized by two quantities: the retentivity B_R (*i.e.*, the residual magnetic field strength at zero magnetic intensity) and the coercivity $\mu_0 H_c$ (*i.e.*, the negative magnetic intensity required to demagnetize the material: this quantity is conventionally multiplied by μ_0 to give it the units of magnetic field strength). The operating point (*i.e.*, the values of B and $\mu_0 H$ inside the sphere) is obtained from the intersection of the demagnetization curve and the curve $B = \mu H$. It is clear from Eqs. (3.166) and (3.167) that

$$\mu = -2\,\mu_0 \tag{3.168}$$

for a uniformly magnetized sphere in the absence of external fields. The magnetization inside the sphere is easily calculated once the operating point has been determined. In fact, $M_0 = B - \mu_0 H$. It is clear from Fig. 4 that for a magnetic material to be a good permanent magnet it must possess both a large retentivity *and* a large coercivity. A material with a large retentivity but a small coercivity is unable to retain a significant magnetization in the absence of a strong external magnetizing field.

3.16 A soft iron sphere in a uniform magnetic field

The opposite extreme to a "hard" ferromagnetic material, which can maintain a large remnant magnetization in the absence of external fields, is a "soft" ferromagnetic material, for which the remnant magnetization is relatively small. Let us consider a somewhat idealized situation in which the remnant magnetization is negligible. In this situation there is no hysteresis, so the $\boldsymbol{B}$-$\boldsymbol{H}$ relation for the material reduces to

$$\boldsymbol{B} = \mu(B)\,\boldsymbol{H}, \tag{3.169}$$

where $\mu(B)$ is a single valued function. The most commonly occurring "soft" ferromagnetic material is soft iron (*i.e.*, annealed, low impurity iron).

Consider a sphere of soft iron placed in an initially uniform external field $\boldsymbol{B}_0 = B_0\,\hat{\boldsymbol{z}}$. The $\mu_0\boldsymbol{H}$ and $\boldsymbol{B}$ fields inside the sphere are most easily obtained by taking the solutions (3.166) and (3.167) (which are still valid in this case), and superimposing on them the uniform field $\boldsymbol{B}_0$. We are justified in doing this because the equations which govern magnetostatic problems are *linear*. Thus, inside the sphere we have

$$\mu_0\boldsymbol{H} = \boldsymbol{B}_0 - \frac{1}{3}\,\mu_0\boldsymbol{M}, \tag{3.170a}$$

$$\boldsymbol{B} = \boldsymbol{B}_0 + \frac{2}{3}\,\mu_0\boldsymbol{M}. \tag{3.170b}$$

Combining Eqs. (3.169) and (3.170) yields

$$\mu_0\boldsymbol{M} = 3\left(\frac{\mu-\mu_0}{\mu+2\mu_0}\right)\boldsymbol{B}_0, \tag{3.171}$$

with

$$\boldsymbol{B} = \left(\frac{3\mu}{\mu+2\mu_0}\right)\boldsymbol{B}_0, \tag{3.172}$$

where, in general, $\mu = \mu(B)$. Clearly, for a highly permeable material (*i.e.*, $\mu/\mu_0 \gg 1$, which is certainly the case for soft iron) the magnetic field strength inside the sphere is approximately three times that of the externally applied field. In other words, the magnetic field is amplified inside the sphere.

The amplification of the magnetic field by a factor three in the high permeability limit is specific to a sphere. It can be shown that for elongated objects (*e.g.*, rods), aligned along the direction of the external field, the amplification factor can be considerably larger than this.

It is important to realize that the magnetization inside a ferromagnetic material cannot increase without limit. The maximum possible value of $\boldsymbol{M}$ is called the saturation magnetization, and is usually denoted $\boldsymbol{M}_s$. Most ferromagnetic materials saturate when they are placed in external magnetic fields whose strengths are greater than, or of order, one tesla. Suppose that our soft iron sphere first attains the saturation magnetization when the unperturbed external magnetic field strength is B_s. It follows from (3.170b) and (3.171) (with $\mu \gg \mu_0$) that

$$B = B_0 + 2B_s \tag{3.173}$$

inside the sphere, for $B_0 > B_s$. In this case, the field amplification factor is

$$\frac{B}{B_0} = 1 + 2\,\frac{B_s}{B_0}. \tag{3.174}$$

Thus, for $B_0 \gg B_s$ the amplification factor approaches unity. We conclude that if a ferromagnetic material is placed in an external field which greatly exceeds that required to cause saturation then the material effectively loses its magnetic properties, so that $\mu \simeq \mu_0$. Clearly, it is very important to avoid saturating the soft magnets used to channel magnetic flux around transformer circuits. This sets an upper limit on the magnetic field strengths which can occur in such circuits.

3.17 Magnetic shielding

There are many situations, particularly in experimental physics, where it is desirable to shield a certain region from magnetic fields. This can be achieved by surrounding the region in question by a material of high permeability. It is vitally important that a material used as a magnetic shield does not develop a permanent magnetization in the presence of external fields, otherwise the material itself may become a source of magnetic fields. The most effective commercially available magnetic shielding material is called *Mumetal*, and is an alloy of 5% Copper, 2% Chromium, 77% Nickel, and 16% Iron. The maximum permeability of Mumetal is about $10^5\,\mu_0$. This material also possesses a particularly low retentivity and coercivity. Unfortunately, Mumetal is *extremely* expensive. Let us investigate how much of this material is actually required to shield a given region from an external magnetic field.

Consider a spherical shell of magnetic shielding, made up of material of permeability μ, placed in a formerly uniform magnetic field $\boldsymbol{B}_0 = B_0\,\hat{\boldsymbol{z}}$. Suppose that the inner radius of the shell is a and the outer radius is b. Since there are no free currents in the problem, we can write $\boldsymbol{H} = -\nabla\phi_m$. Furthermore, since $\boldsymbol{B} = \mu\boldsymbol{H}$ and $\nabla\cdot\boldsymbol{B} = 0$, it is clear that the magnetic scalar potential satisfies Laplace's equation, $\nabla^2\phi_m = 0$, throughout all space. The boundary conditions are that the potential must be well behaved at $r = 0$ and $r\rightarrow\infty$, and also that the tangential and the normal components of $\boldsymbol{H}$ and $\boldsymbol{B}$, respectively, must be continuous at $r = a$ and $r = b$. The boundary conditions on $\boldsymbol{H}$ merely imply that the scalar potential ϕ_m must be continuous at $r = a$ and $r = b$. The boundary conditions on $\boldsymbol{B}$ yield

$$\mu_0\,\frac{\partial\phi_m(r=a-)}{\partial r} = \mu\,\frac{\partial\phi_m(r=a+)}{\partial r}, \tag{3.175a}$$

$$\mu_0\,\frac{\partial\phi_m(r=b+)}{\partial r} = \mu\,\frac{\partial\phi_m(r=b-)}{\partial r}. \tag{3.175b}$$

Let us try the following test solution for the magnetic potential:

$$\phi_m = -\frac{B_0}{\mu_0}\,r\cos\theta + \frac{\alpha}{r^2}\cos\theta \tag{3.176}$$

for $r > b$,

$$\phi_m = \left(\beta\,r + \frac{\gamma}{r^2}\right)\cos\theta \tag{3.177}$$

for $b \geq r \geq a$, and

$$\phi_m = \delta\, r \cos\theta \tag{3.178}$$

for $r < a$. This potential is certainly a solution of Laplace's equation throughout space. It yields the uniform magnetic field $\boldsymbol{B}_0$ as $r \rightarrow \infty$, and satisfies physical boundary conditions at $r = 0$ and infinity. Since there is a uniqueness theorem associated with Poisson's equation, we can be certain that this potential is the correct solution to the problem provided that the arbitrary constants α, β, *etc.* can be adjusted in such a manner that the boundary conditions at $r = a$ and $r = b$ are also satisfied.

The continuity of ϕ_m at $r = a$ and $r = b$ requires that

$$\beta\, a + \frac{\gamma}{a^2} = \delta\, a, \tag{3.179}$$

and

$$\beta\, b + \frac{\gamma}{b^2} = -\frac{B_0}{\mu_0}\, b + \frac{\alpha}{b^2}. \tag{3.180}$$

The boundary conditions (3.175) yield

$$\mu_0\, \delta = \mu \left(\beta - \frac{2\gamma}{a^3}\right), \tag{3.181}$$

and

$$\mu_0 \left(-\frac{B_0}{\mu_0} - \frac{2\alpha}{b^3}\right) = \mu \left(\beta - \frac{2\gamma}{b^3}\right). \tag{3.182}$$

It follows that

$$\mu_0\, \alpha = \left[\frac{(2\mu + \mu_0)(\mu - \mu_0)}{(2\mu + \mu_0)(\mu + 2\mu_0) - 2(a^3/b^3)(\mu - \mu_0)^2}\right] (b^3 - a^3) B_0, \tag{3.183a}$$

$$\mu_0\, \beta = -\left[\frac{3(2\mu + \mu_0)\mu_0}{(2\mu + \mu_0)(\mu + 2\mu_0) - 2(a^3/b^3)(\mu - \mu_0)^2}\right] B_0, \tag{3.183b}$$

$$\mu_0\, \gamma = -\left[\frac{3(\mu - \mu_0)\mu_0}{(2\mu + \mu_0)(\mu + 2\mu_0) - 2(a^3/b^3)(\mu - \mu_0)^2}\right] a^3 B_0, \tag{3.183c}$$

$$\mu_0\, \delta = -\left[\frac{9\mu\mu_0}{(2\mu + \mu_0)(\mu + 2\mu_0) - 2(a^3/b^3)(\mu - \mu_0)^2}\right] B_0. \tag{3.183d}$$

Consider the limit of a thin, high permeability shell for which $b = a + d$, $d/a \ll 1$, and $\mu/\mu_0 \gg 1$. In this limit, the field inside the shell is given by

$$\boldsymbol{B} \simeq \frac{3}{2}\frac{\mu_0}{\mu}\frac{a}{d}\,\boldsymbol{B}_0. \tag{3.184}$$

Thus, if $\mu \simeq 10^5\mu_0$ for Mumetal, then we can reduce the magnetic field strength inside the shell by almost a factor of 1000 using a shell whose thickness is only 1/100 th of its radius. Clearly, a little Mumetal goes a long way! Note, however, that as the external field strength, B_0, is increased, the Mumetal shell eventually saturates, and μ/μ_0 gradually falls to unity. Thus, extremely strong magnetic fields (typically, $B_0 \gtrsim 1$ tesla) are hardly shielded at all by Mumetal, or similar magnetic materials.

3.18 Magnetic energy

Consider an electrical conductor. Suppose that a battery with an electromotive field $\boldsymbol{E}'$ is feeding energy into this conductor. The energy is either dissipated as heat or is used to generate a magnetic field. Ohm's law inside the conductor gives

$$\boldsymbol{j}_t = \sigma(\boldsymbol{E} + \boldsymbol{E}'), \tag{3.185}$$

where $\boldsymbol{j}_t$ is the true current density, σ is the conductivity, and $\boldsymbol{E}$ is the inductive electric field. Taking the scalar product with $\boldsymbol{j}_t$, we obtain

$$\boldsymbol{E}'\cdot\boldsymbol{j}_t = \frac{{j_t}^2}{\sigma} - \boldsymbol{E}\cdot\boldsymbol{j}_t. \tag{3.186}$$

The left-hand side of this equation represents the rate at which the battery does work on the conductor. The first term on the right-hand side is the rate of Joule heating inside the conductor. We tentatively identify the remaining term with the rate at which energy is fed into the magnetic field. If all fields are quasi-stationary (*i.e.*, slowly varying) then the displacement current can be neglected, and the Ampère-Maxwell equation reduces to $\nabla \wedge \boldsymbol{H} = \boldsymbol{j}_t$. Substituting this expression into Eq. (3.186) and integrating over all space, we get

$$\int \boldsymbol{E}'\cdot(\nabla\wedge\boldsymbol{H})\,d^3\boldsymbol{r} = \int \frac{(\nabla\wedge\boldsymbol{H})^2}{\sigma}\,d^3\boldsymbol{r} - \int \boldsymbol{E}\cdot(\nabla\wedge\boldsymbol{H})\,d^3\boldsymbol{r}. \tag{3.187}$$

The last term can be integrated by parts using the relation

$$\nabla\cdot(\boldsymbol{E}\wedge\boldsymbol{H}) = \boldsymbol{H}\cdot(\nabla\wedge\boldsymbol{E}) - \boldsymbol{E}\cdot(\nabla\wedge\boldsymbol{H}). \tag{3.188}$$

The divergence theorem plus the Faraday-Maxwell equation yield

$$\int \boldsymbol{E}\cdot(\nabla\wedge\boldsymbol{H})\,d^3\boldsymbol{r} = -\int \boldsymbol{H}\cdot\frac{\partial\boldsymbol{B}}{\partial t}\,d^3\boldsymbol{r} - \int(\boldsymbol{E}\wedge\boldsymbol{H})\cdot d\boldsymbol{S}. \tag{3.189}$$

Since $\boldsymbol{E}\wedge\boldsymbol{H}$ falls off at least as fast as $1/r^5$ in electrostatic and quasi-stationary magnetic fields ($1/r^2$ comes from electric monopole fields, and $1/r^3$ from magnetic dipole fields), the surface integral in the above expression can be neglected. Of course, this is not the case for radiation fields, for which $\boldsymbol{E}$ and $\boldsymbol{H}$ fall off like $1/r$. Thus, the constraint of "quasi-stationarity" effectively means that the fields vary sufficiently slowly that any radiation fields can be neglected.

The total power expended by the battery can now be written

$$\int \boldsymbol{E}'\cdot(\nabla\wedge\boldsymbol{H})\,d^3\boldsymbol{r} = \int \frac{(\nabla\wedge\boldsymbol{H})^2}{\sigma}\,d^3\boldsymbol{r} + \int \boldsymbol{H}\cdot\frac{\partial\boldsymbol{B}}{\partial t}\,d^3\boldsymbol{r}. \tag{3.190}$$

The first term on the right-hand side has already been identified as the energy loss rate due to Joule heating. The last term is obviously the rate at which energy is fed into the magnetic field. The variation δU in the magnetic field energy can therefore be written

$$\delta U = \int \boldsymbol{H}\cdot\delta\boldsymbol{B}\,d^3\boldsymbol{r}. \tag{3.191}$$

This result is analogous to the result (3.64) for the variation in the energy of an electrostatic field.

In order to make Eq. (3.191) integrable, we must assume a functional relationship between $\boldsymbol{H}$ and $\boldsymbol{B}$. For a medium which magnetizes linearly the integration can be carried out in much the same manner as Eq. (3.67), to give

$$U = \frac{1}{2}\int \boldsymbol{H}\cdot\boldsymbol{B}\,d^3\boldsymbol{r}. \tag{3.192}$$

Thus, the magnetostatic energy density inside a linear magnetic material is given by

$$W = \frac{\boldsymbol{H}\cdot\boldsymbol{B}}{2}. \tag{3.193}$$

Unfortunately, most interesting magnetic materials, such as ferromagnets, exhibit a nonlinear relationship between $\boldsymbol{H}$ and $\boldsymbol{B}$. For such materials, Eq. (3.191) can only be integrated between definite states, and the result, in general, depends on the past history of the sample. For ferromagnets, the integral of Eq. (3.191) has a finite, non-zero value when $\boldsymbol{B}$ is integrated around a complete magnetization cycle. This cyclic energy loss is given by

$$\Delta U = \int \oint \boldsymbol{H}\cdot d\boldsymbol{B}\, d^3\boldsymbol{r}. \tag{3.194}$$

In other words, the energy expended per unit volume when a magnetic material is carried through a magnetization cycle is equal to the area of its hysteresis loop as plotted in a graph of B against H. Thus, it is particularly important to ensure that the magnetic materials used to form transformer cores possess hysteresis loops with comparatively small areas, otherwise the transformers are likely to be extremely inefficient.

4 Electromagnetic wave propagation in dielectrics

4.1 Introduction

It is easily demonstrated that the fields associated with an electromagnetic wave propagating through a uniform dielectric medium of dielectric constant ϵ satisfy

$$\left(\frac{\epsilon}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\boldsymbol{E} = 0, \tag{4.1}$$

and

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{4.2}$$

The plane wave solutions to these equations are well known:

$$\boldsymbol{E} = \boldsymbol{E}_0\, \mathrm{e}^{\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \tag{4.3a}$$

$$\boldsymbol{B} = \boldsymbol{B}_0\, \mathrm{e}^{\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \tag{4.3b}$$

where $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ are constant vectors, with

$$\frac{\omega^2}{k^2} = \frac{c^2}{\epsilon}, \tag{4.4}$$

and

$$\boldsymbol{B}_0 = \frac{\boldsymbol{k}\wedge\boldsymbol{E}_0}{\omega}. \tag{4.5}$$

The phase velocity of the wave is given by

$$v = \frac{\omega}{k} = \frac{c}{n}, \tag{4.6}$$

where

$$n = \sqrt{\epsilon} \tag{4.7}$$

is called the *refractive index* of the medium. It is clear that an electromagnetic wave propagates with a phase velocity which is slower than the velocity of light in a conventional (*i.e.*, ϵ real and greater than unity) dielectric medium.

In some dielectric media ϵ is complex. This leads, from Eq. (4.4), to a complex wave vector $\boldsymbol{k}$. For a wave propagating in the x-direction we obtain

$$\boldsymbol{E} = \boldsymbol{E}_0 \exp[\,\mathrm{i}\,(\mathrm{Re}(k)\,x - \omega t)]\exp[-\mathrm{Im}(k)\,x]. \tag{4.8}$$

Thus, a complex dielectric constant leads to the attenuation (or amplification) of the wave as it propagates through the medium in question.

Up to now, we have tacitly assumed that ϵ is the same for waves of all frequencies. In practice, this is not the case. In dielectric media ϵ is, in general, complex, and varies (in some cases, strongly) with the wave frequency, ω. Thus, waves of different frequencies propagate through a dielectric medium with different phase velocities. This phenomenon is known as *dispersion*. Moreover, there may exist frequency bands in which the waves are attenuated (*i.e.*, absorbed). All of this makes the problem of determining the behaviour of a wave packet as it propagates through a dielectric medium far from straightforward. Recall, that the solution to this problem for a wave packet traveling through a vacuum is fairly trivial. The packet propagates at the velocity c without changing its shape. What is the equivalent result for the case of a dielectric medium? This is an important question, since nearly all of our information regarding the universe is obtained from the study of electromagnetic waves emitted by distant objects. All of these waves have to propagate through dispersive media (*e.g.*, the interstellar medium, the ionosphere, the atmosphere) before reaching us. It is, therefore, vitally important that we understand which aspects of these wave signals are predominantly determined by the wave sources, and which are strongly modified by the dispersive media through which they have propagated in order to reach us.

The study of wave propagation through dispersive media was pioneered by two scientists, Arnold Sommerfeld and Léon Brillouin, during the first half of this century. In the following discussion, we shall stick as close as possible to Sommerfeld and Brillouin's original analysis.

4.2 The form of the dielectric constant

Let us investigate an electromagnetic wave propagating through a transparent, isotropic, non-conducting, medium. The electric displacement inside the medium is given by

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{4.9}$$

where $\boldsymbol{P}$ is the electric polarization. Since electrons are much lighter than ions (or atomic nuclei), we would expect the former to displace further than the latter under the influence of an electric field. Thus, to a first approximation the polarization $\boldsymbol{P}$ is determined by the electron response to the wave. Suppose that the electrons displace a distance $\boldsymbol{s}$ from their rest positions in the presence of the wave. It follows that

$$\boldsymbol{P} = -N e\, \boldsymbol{s}, \tag{4.10}$$

where N is the number density of electrons.

Let us assume that the electrons are bound "quasi-elastically" to their rest positions, so that they seek to return to these positions when displaced from them by a field $\boldsymbol{E}$. It follows that $\boldsymbol{s}$ satisfies the differential equation of the form

$$m\, \ddot{\boldsymbol{s}} + f\, \boldsymbol{s} = -e\boldsymbol{E}, \tag{4.11}$$

where m is the electron mass, $-f\boldsymbol{s}$ is the restoring force, and $\dot{}$ denotes a partial derivative with respect to time. The above equation can also be written

$$\ddot{\boldsymbol{s}} + g\, \omega_0\, \dot{\boldsymbol{s}} + {\omega_0}^2 \boldsymbol{s} = -\frac{e}{m}\, \boldsymbol{E}, \tag{4.12}$$

where

$${\omega_0}^2 = \frac{f}{m} \tag{4.13}$$

is the characteristic oscillation frequency of the electrons. In almost all dielectric media this frequency lies in the far *ultraviolet* region of the electromagnetic spectrum. In Eq. (4.12) we have added a phenomenological damping term $g\, \omega_0\, \dot{\boldsymbol{s}}$, in order to take into account the fact that an electron excited by an impulsive electric field does not oscillate for ever. In general, however, electrons in dielectric media can be regarded as high-Q oscillators, which is another way of saying that the dimensionless damping constant g is typically much less than unity. Thus, an electron "rings" for a long time after being excited by an impulse.

Let us assume that the electrons oscillate in sympathy with the wave at the wave frequency ω. It follows from Eq. (4.12) that

$$\boldsymbol{s} = -\frac{(e/m)\,\boldsymbol{E}}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.14}$$

Note that we have neglected the response of the electrons to the magnetic component of the the wave. It is easily demonstrated that this is a good approximation provided that the electrons do not oscillate with relativistic velocities (*i.e.*, provided that the amplitude of the wave is sufficiently small). Thus, Eq. (4.10) yields

$$\boldsymbol{P} = \frac{(N e^2/m)\,\boldsymbol{E}}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.15}$$

Since, by definition,

$$\boldsymbol{D} = \epsilon_0 \epsilon\,\boldsymbol{E} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{4.16}$$

it follows that

$$\epsilon(\omega) \equiv n^2(\omega) = 1 + \frac{(N e^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.17}$$

Thus, the index of refraction is frequency dependent. Since ω_0 typically lies in the ultraviolet region of the spectrum (and since $g \ll 1$), it is clear that the denominator ${\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0 \simeq {\omega_0}^2 - \omega^2$ is positive in the entire visible spectrum, and is larger at the red end than at the blue end. This implies that *blue light is refracted more than red light.* This is normal dispersion. Incidentally, an expression, like the above, which specifies the dispersion of waves propagating through some dielectric medium is usually called a *dispersion relation.*

Let us now suppose that there are N molecules per unit volume with Z electrons per molecule, and that instead of a single oscillation frequency for all electrons, there are f_i electrons per molecule with oscillation frequency ω_i and damping constant g_i. It is easily demonstrated that

$$n^2(\omega) = 1 + \frac{N e^2}{\epsilon_0 m} \sum_i \frac{f_i}{{\omega_i}^2 - \omega^2 - \mathrm{i}\,g_i\,\omega\,\omega_i}, \tag{4.18}$$

where the *oscillator strengths* f_i satisfy the sum rule,

$$\sum_i f_i = Z. \tag{4.19}$$

A more exact quantum mechanical treatment of the response of an atom, or molecule, to a low amplitude electromagnetic wave also leads to a dispersion relation of the form (4.18), except that the quantities f_i, ω_i, and g_i can, in principle, be calculated from first principles. In practice, this is too difficult except for the very simplest cases.

Since the damping constants g_i are generally small compared to unity, it follows from Eq. (4.18) that $n(\omega)$ is a predominately real quantity at most wave frequencies. The factor $({\omega_i}^2 - \omega^2)^{-1}$ is positive for $\omega < \omega_i$ and negative for $\omega > \omega_i$. Thus, at low frequencies, below the smallest ω_i, all of the terms in the sum in (4.18) are positive, and $n(\omega)$ is consequently greater than unity. As ω is raised so that it passes successive ω_i values, more and more negative terms occur in the sum, until eventually the whole sum is negative and $n(\omega)$ is less than unity. Thus, at very high frequencies electromagnetic waves propagate through dielectric media with phase velocities which exceed the velocity of light in a vacuum. For $\omega \simeq \omega_i$, Eq. (4.18) predicts a rather violent variation of the refractive index with frequency. Let us examine this phenomenon more closely.

4.3 Anomalous dispersion and resonant absorption

When ω is approximately equal to ω_i the dispersion relation (4.18) reduces to

$$n^2 = {n_i}^2 + \frac{N e^2 f_i/\epsilon_0 m}{{\omega_i}^2 - \omega^2 - \mathrm{i}\, g_i\, \omega\, \omega_i}, \tag{4.20}$$

where n_i is the average contribution in the vicinity of $\omega = \omega_i$ of all other resonances (also included in n_i is the contribution 1 of the vacuum displacement current, which was previously written down separately). The refractive index is clearly complex. For a wave propagating in the x-direction

$$\boldsymbol{E} = \boldsymbol{E}_0 \exp[\mathrm{i}\,(\omega/c)(\mathrm{Re}(n)\, x - c\, t)] \exp[-(\omega/c)\,\mathrm{Im}(n)\, x]. \tag{4.21}$$

Thus, the phase velocity of the wave is determined by the real part of the refractive index via

$$v = \frac{c}{\mathrm{Re}(n)}. \tag{4.22}$$

Note that a positive imaginary component of the refractive index leads to the attenuation of the wave as it propagates.

Let

$$\begin{aligned} a^2 &= \frac{N e^2 f_i}{\epsilon_0 m\, {\omega_i}^2}, &\quad (4.23a)\\ x &= \frac{\omega^2 - {\omega_i}^2}{{\omega_i}^2}, &\quad (4.23b)\\ y &= \frac{\mathrm{Re}(n)^2 - \mathrm{Im}(n)^2}{a^2}, &\quad (4.23c)\\ z &= \frac{2\,\mathrm{Re}(n)\,\mathrm{Im}(n)}{a^2}, &\quad (4.23d) \end{aligned}$$

where a, x, y, z are all dimensionless quantities. It follows from Eq. (4.20) that

$$\begin{aligned} y &= \frac{{n_i}^2}{a^2} - \frac{x}{x^2 + {g_i}^2(1+x)}, &\quad (4.24a)\\ z &= \frac{g_i\sqrt{1+x}}{x^2 + {g_i}^2(1+x)}. &\quad (4.24b) \end{aligned}$$

Let us adopt the physical ordering $g_i \ll 1$. The extrema of the function y occur at $x = \pm g_i$. It is easily demonstrated that

$$\begin{aligned} y_{\rm min} &= y(x = g_i) = \frac{{n_i}^2}{a^2} - \frac{1}{2g_i}, &\quad (4.24c)\\ y_{\rm max} &= y(x = -g_i) = \frac{{n_i}^2}{a^2} + \frac{1}{2g_i}. &\quad (4.24d) \end{aligned}$$

The maximum value of the function z occurs at $x = 0$. In fact,

$$z_{\rm max} = \frac{1}{g_i}. \quad (4.25)$$

Note that

$$z(x = \pm g_i) = \frac{1}{2g_i}. \quad (4.26)$$

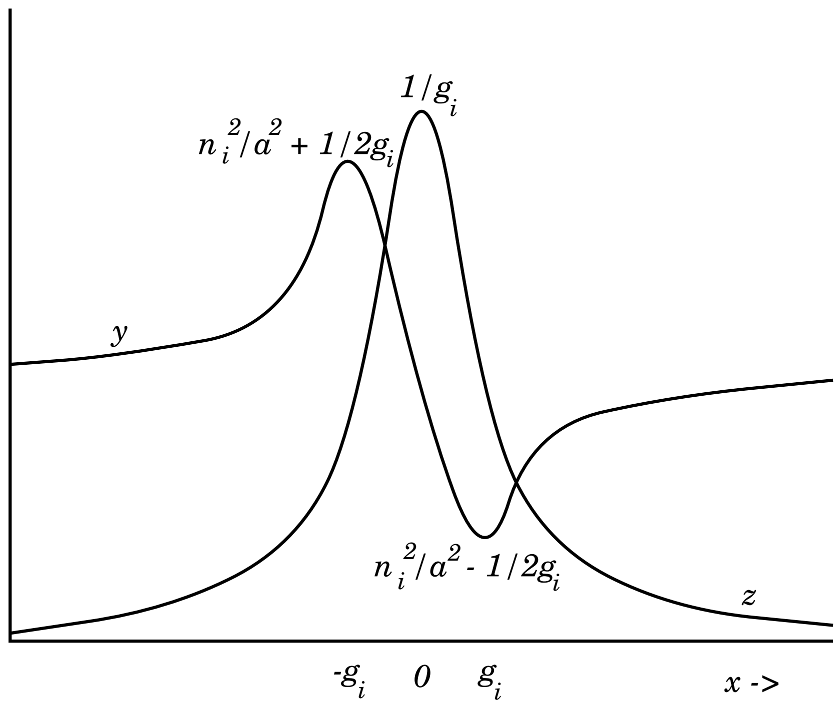


Figure 5: Sketch of the variation of the functions y and z with x

Figure 5 shows a sketch of the variation of the functions y and z with x. These curves are also indicative of the variation of $\mathrm{Re}(n)$ and $\mathrm{Im}(n)$, respectively, with frequency ω in the vicinity of the resonant frequency ω_i. Recall that normal dispersion is associated with an increase in $\mathrm{Re}(n)$ with increasing ω. The reverse situation is termed *anomalous dispersion*. It is clear from the figure that normal dispersion occurs everywhere except in the immediate neighbourhood of the resonant frequency ω_i. It is also clear that the imaginary part of the refractive index is only appreciable in those regions of the electromagnetic spectrum where anomalous dispersion takes place. A positive imaginary component of the refractive index implies that the wave is absorbed as it propagates through the medium, so the regions of the spectrum where $\mathrm{Im}(n)$ is appreciable are called regions of *resonant absorption*. Anomalous dispersion and resonant absorption take place in the vicinity of the ith resonance when $|\omega - \omega_i| \lesssim O(g_i)$. Since the damping constants g_i are, in practice, very small compared to unity, the regions of the spectrum in which resonant absorption takes place are strongly localized in the vicinity of the various resonant frequencies.

The dispersion relation (4.18) only takes electron resonances into account. Of course, there are also resonances associated with displacements of the ions (or atomic nuclei). The off-resonance contributions to the right-hand side of Eq. (4.18) from the ions are smaller than those from the electrons by a factor of order m/M (where M is a typical ion mass). Nevertheless, the ion contributions are important because they give rise to anomalous dispersion and resonant absorption close to the ion resonant frequencies. The ion resonances associated with the stretching and bending of molecular bonds typically lie in the infrared region of the electromagnetic spectrum. Those associated with molecular rotation (these resonances only affect the dispersion relation if the molecule is polar) occur in the microwave region of the spectrum. Thus, both air and water exhibit strong resonant absorption of electromagnetic waves in both the ultraviolet and infrared regions of the spectrum. In the first case this is due to electron resonances, and in the second to ion resonances. The visible region of the spectrum exists as a narrow window lying between these two regions in which there is comparatively little attenuation of electromagnetic waves.

4.4 Wave propagation through a conducting medium

In the limit $\omega \to 0$, there is a significant difference in the response of a dielectric medium, depending on whether the lowest resonant frequency is zero or non-zero. For insulators the lowest resonant frequency is different from zero. In this case, the low frequency refractive index is predominately real, and is also greater than unity. Suppose, however, that some fraction f_0 of the electrons are "free," in the sense of having $\omega_0 = 0$. In this situation, the low frequency dielectric constant takes the form

$$\epsilon(\omega) \equiv n^2(\omega) = {n_0}^2 + \mathrm{i}\,\frac{N e^2}{\epsilon_0 m}\frac{f_0}{\omega\,(\gamma_0 - \mathrm{i}\,\omega)}, \tag{4.27}$$

where n_0 is the contribution to the refractive index from all of the other resonances, and $\gamma_0 = \lim_{\omega_0\to 0} g_0\,\omega_0$. Note that for a conducting medium the contribution to the refractive index from the free electrons is singular at $\omega = 0$. This singular behaviour can be explained as follows. Consider the Ampère-Maxwell equation

$$\nabla \wedge \boldsymbol{B} = \mu_0 \left(\boldsymbol{j}_t + \frac{\partial \boldsymbol{D}}{\partial t} \right). \tag{4.28}$$

Let us assume that the medium in question obeys Ohm's law, $\boldsymbol{j}_t = \sigma \boldsymbol{E}$, and has a "normal" dielectric constant ${n_0}^2$. Here, σ is the conductivity. Assuming an $\exp(-\mathrm{i}\,\omega t)$ time dependence of all field quantities the above equation yields

$$\frac{\nabla \wedge \boldsymbol{B}}{\mu_0} = -\mathrm{i}\,\epsilon_0\,\omega \left({n_0}^2 + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega} \right) \boldsymbol{E}. \tag{4.29}$$

Suppose, however, that we do not explicitly use Ohm's law but, instead, attribute all of the properties of the medium to the dielectric constant. In this case, the effective dielectric constant of the medium is equivalent to the term in round brackets on the right-hand side of the above equation. Thus,

$$\epsilon(\omega) \equiv n^2(\omega) = {n_0}^2 + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega}. \tag{4.30}$$

A comparison of this term with Eq. (4.27) yields the following expression for the conductivity

$$\sigma = \frac{f_0 N e^2}{m(\gamma_0 - \mathrm{i}\,\omega)}. \tag{4.31}$$

Thus, at low frequencies conductors possess predominately real conductivities (*i.e.*, the current remains in phase with the electric field). However, at higher frequencies the conductivity becomes complex. At these sorts of frequencies there is little meaningful distinction between a conductor and an insulator, since the "conductivity" contribution to $\epsilon(\omega)$ appears as a resonant amplitude just like the other contributions. For a good conductor, such as Copper, the conductivity remains predominately real for all frequencies up to and including those in the microwave region of the electromagnetic spectrum.

The conventional way in which to represent the complex refractive index of a conducting medium (in the low frequency limit) is to write it in terms of a real "normal" dielectric constant, $\epsilon = {n_0}^2$, and a real conductivity, σ. Thus, from Eq. (4.30)

$$n^2(\omega) = \epsilon + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega}. \tag{4.32}$$

For a poor conductor ($\sigma/\epsilon\,\epsilon_0\,\omega \ll 1$) we find

$$k = n\,\frac{\omega}{c} \simeq \sqrt{\epsilon}\,\frac{\omega}{c} + {\rm i}\,\frac{\sigma}{2\sqrt{\epsilon}\,\epsilon_0\,c}. \tag{4.33}$$

In this limit ${\rm Re}(k) \gg {\rm Im}(k)$, and the attenuation of the wave, which is governed by ${\rm Im}(k)$ [see Eq. (4.8)], is independent of the frequency. Thus, for a poor conductor the wave is basically the same as a wave propagating through a conventional dielectric with dielectric constant ϵ, except that the wave attenuates gradually over a distance of very many wavelengths. For a good conductor ($\sigma/\epsilon\,\epsilon_0\,\omega \gg 1$)

$$k \simeq e^{{\rm i}\,\pi/4}\sqrt{\mu_0\,\sigma\,\omega}. \tag{4.34}$$

It follows from Eq. (4.5) that

$$\frac{cB_0}{E_0} = \frac{kc}{\omega} = e^{{\rm i}\,\pi/4}\sqrt{\frac{\sigma}{\epsilon_0\,\omega}}. \tag{4.35}$$

Thus, the phase of the magnetic field *lags* that of the electric field by 45°. Moreover, the magnitude of cB_0 is much larger than that of E_0 (since $\sigma/\epsilon_0\,\omega \gg \epsilon \gtrsim 1$). It follows that the field energy is almost entirely magnetic in nature. It is clear that an electromagnetic wave propagating through a good conductor has markedly different properties to a wave propagating through a conventional dielectric. For a wave propagating in the x-direction, the amplitudes of the electric and magnetic fields attenuate like $\exp(-x/d)$, where

$$d = \sqrt{\frac{2}{\mu_0\,\sigma\,\omega}}. \tag{4.36}$$

This quantity is known as the *skin depth*. It is clear that an electromagnetic wave incident on a conducting medium will not penetrate more than a few skin depths into that medium.

4.5 The high frequency limit

Consider the behaviour of the dispersion relation (4.18) in the high frequency limit $\omega \gg \omega_i$ (for all i). In this limit, the relation simplifies considerably to give

$$n^2(\omega) = 1 - \frac{{\omega_p}^2}{\omega^2}, \tag{4.37}$$

where the quantity

$$\omega_p = \sqrt{\frac{N Z e^2}{\epsilon_0 m}} \tag{4.38}$$

is called the *plasma frequency.* The wave-number in the high frequency limit is given by

$$k = n\,\frac{\omega}{c} = \frac{\sqrt{\omega^2 - {\omega_p}^2}}{c}. \tag{4.39}$$

This expression is only valid in dielectrics when $\omega \gg \omega_p$. Thus, the refractive index is real and slightly less than unity, giving waves which propagate without attenuation with a phase velocity slightly larger than the velocity of light in vacuum. However, in certain ionized media (in particular, in tenuous plasmas such as occur in the ionosphere) the electrons are free and the damping is negligible. In this case, Eqs. (4.37) and (4.39) are valid even when $\omega < \omega_p$. It is clear that a wave can only propagate through a tenuous plasma if its frequency exceeds the plasma frequency (in which case it has a real wave-number). If wave frequency is less than the plasma frequency then the wave-number is purely imaginary, according to Eq. (4.39), and the wave is therefore attenuated. This accounts for the fact that long-wave and medium-wave radio signals can be received even when the transmitter lies over the horizon. The frequency of these waves is less than the plasma frequency of the ionosphere, which reflects them, so they are trapped between the ionosphere and the surface of the Earth (which is also a good reflector of radio waves), and can, in certain cases, travel many times around the Earth before being attenuated. Unfortunately, this scheme does not work very well for medium-wave signals at night. The problem is that the plasma frequency of the ionosphere is proportional to the square root of the number density of free ionospheric electrons. These free electrons are generated through the ionization of neutral molecules by ultraviolet radiation from the Sun. Of course, there is no radiation from the Sun at night so the density of free electrons starts to drop as the electrons gradually recombine with ions in the ionosphere. Eventually, the plasma frequency of the ionosphere falls below the frequency of medium-wave radio signals allowing them to be transmitted through the ionosphere into outer space. The ionosphere appears almost completely transparent to high frequency signals such as TV and FM radio signals. Thus, this type of signal is not reflected by the ionosphere. Consequently, to receive such signals it is necessary to be in the line of sight of the relevant transmitter.

4.6 Faraday rotation

The electromagnetic force acting on an electron is given by

$$\boldsymbol{f} = -e\,(\boldsymbol{E} + \boldsymbol{v}\wedge\boldsymbol{B}). \tag{4.40}$$

If the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields in question are due to an electromagnetic wave propagating through a dielectric medium then

$$|B| = \frac{n}{c}\,|E|. \tag{4.41}$$

It follows that the ratio of the magnetic to the electric forces acting on the electron is $n\,v/c$. In other words, the magnetic force is completely negligible unless the wave amplitude is sufficiently high that the electron moves relativistically in response to the wave. This state of affairs is rare, but can occur when intense laser beams are made to propagate through plasmas.

Suppose, however, that the dielectric medium contains an externally generated magnetic field $\boldsymbol{B}$. This can easily be made much stronger than the optical magnetic field. In this case, it is possible for a magnetic field to affect the propagation of low amplitude electromagnetic waves. The electron equation of motion (4.11) generalizes to

$$m\,\ddot{\boldsymbol{s}} + f\,\boldsymbol{s} = -e(\boldsymbol{E} + \dot{\boldsymbol{s}}\wedge\boldsymbol{B}), \tag{4.42}$$

where any damping of the motion has been neglected. Suppose that the direction of $\boldsymbol{B}$ is in the positive z-direction, and that the wave propagates in the same direction. With these assumptions the $\boldsymbol{E}$ and $\boldsymbol{s}$ vectors lie in the x-y plane. The above equation reduces to

$$({\omega_0}^2 - \omega^2)\,s_x - \mathrm{i}\,\omega\Omega\,s_y \;=\; -\frac{e}{m}\,E_x, \tag{4.43a}$$

$$({\omega_0}^2 - \omega^2)\,s_y + \mathrm{i}\,\omega\Omega\,s_x \;=\; -\frac{e}{m}\,E_y, \tag{4.43b}$$

provided that all perturbed quantities have an $\exp(-\mathrm{i}\,\omega t)$ time dependence. Here,

$$\Omega = \frac{eB}{m} \tag{4.44}$$

is the electron cyclotron frequency. Let

$$E_\pm = E_x \pm \mathrm{i}\,E_y, \tag{4.45}$$

and

$$s_\pm = s_x \pm \mathrm{i}\,s_y. \tag{4.46}$$

Note that

$$E_x = \frac{1}{2}\,(E_+ + E_-), \tag{4.47a}$$

$$E_y = \frac{1}{2\,\mathrm{i}}\,(E_+ - E_-). \tag{4.47b}$$

Equations (4.43) reduce to

$$({\omega_0}^2 - \omega^2 - \omega\,\Omega)\,s_+ = -\frac{e}{m}\,E_+, \tag{4.48a}$$

$$({\omega_0}^2 - \omega^2 + \omega\,\Omega)\,s_- = -\frac{e}{m}\,E_-. \tag{4.48b}$$

Defining $P_\pm = P_x \pm \mathrm{i}\,P_y$, it follows from Eq. (4.10) that

$$P_\pm = \frac{(N e^2/m)\,E_\pm}{{\omega_0}^2 - \omega^2 \mp \omega\,\Omega}. \tag{4.49}$$

Finally, from Eq. (4.15), we can write

$$\epsilon_\pm \equiv n_\pm^{\,2} = 1 + \frac{P_\pm}{\epsilon_0\,E_\pm}, \tag{4.50}$$

giving

$$n_\pm^{\,2}(\omega) = 1 + \frac{(N e^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 \mp \omega\,\Omega}. \tag{4.51}$$

According to the dispersion relation (4.51), the refractive index of a magnetized dielectric medium can take one of two possible values, which presumably correspond to two different types of wave propagating along the z-axis. The first wave has the refractive index n_+ and an associated electric field [see Eqs. (4.45)]

$$E_x = E_0 \cos[(\omega/c)(n_+ z - ct)], \tag{4.52a}$$

$$E_y = E_0 \sin[(\omega/c)(n_+ z - ct)]. \tag{4.52b}$$

This corresponds to a *left-handed circularly polarized wave* propagating in the z-direction with the phase velocity c/n_+. The second wave has the refractive index n_- and an associated electric field

$$E_x = E_0 \cos[(\omega/c)(n_- z - ct)], \tag{4.53a}$$

$$E_y = -E_0 \sin[(\omega/c)(n_- z - ct)]. \tag{4.53b}$$

This corresponds to a *right-handed circularly polarized wave* propagating in the z-direction with the phase velocity c/n_-. It is clear from Eq. (4.51) that $n_+ > n_-$. Thus, we conclude that in the presence of a z-directed magnetic field, a z-directed left-handed circularly polarized wave propagates with a phase velocity which is slightly *less* than that of the corresponding right-handed wave. It should be remarked that the refractive index is always real (in the absence of damping), so the magnetic field gives rise to no net absorption of electromagnetic radiation. This is not surprising since the magnetic field *does no work* on charged particles, and can therefore transfer no energy to the particles from any waves propagating through the medium.

We have seen that right-handed and left-handed circularly polarized waves propagate with different phase velocities through a magnetized dielectric medium. But, what does this imply for the propagation of a plane polarized wave? Let us superimpose the left-handed wave whose electric field is given by Eqs. (4.52) on the right-handed wave whose electric field is given by Eqs. (4.53). In the absence of a magnetic field $n_+ = n_- = n$, and we obtain

$$E_x = 2E_0 \cos[(\omega/c)(nz - ct)], \tag{4.54a}$$

$$E_y = 0. \tag{4.54b}$$

This, of course, is the field of a plane polarized wave (polarized along the x-direction) propagating along the z-axis with the phase velocity c/n. In the presence of a magnetic field we obtain

$$E_x = 2E_0 \cos[(\omega/c)(nz - ct)] \cos[(\omega/2c)(n_+ - n_-)z], \tag{4.55a}$$

$$E_y = 2E_0 \cos[(\omega/c)(nz - ct)] \sin[(\omega/2c)(n_+ - n_-)z], \tag{4.55b}$$

where

$$n = \frac{1}{2}(n_+ + n_-) \tag{4.56}$$

is the mean index of refraction. Equations (4.55) can be recognized as the field of a plane polarized wave whose angle of polarization with respect to the x-axis,

$$\chi = \tan^{-1}(E_y/E_x), \tag{4.57}$$

rotates as the wave propagates along the z-axis with the phase velocity c/n. In fact, the angle of polarization is given by

$$\chi = \frac{\omega}{2c}(n_+ - n_-)\, z, \tag{4.58}$$

which clearly increases linearly with the distance traveled by the wave along the direction of the magnetic field. This rotation of the plane of polarization of a linearly polarized wave propagating through a magnetized dielectric medium is known as *Faraday rotation* (since it was discovered by Michael Faraday in 1845).

Assuming that the cyclotron frequency Ω is relatively small compared to the wave frequency ω, and also that ω does not lie close to the resonant frequency ω_0, it is easily demonstrated that

$$n \simeq 1 + \frac{(Ne^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2}, \tag{4.59}$$

and

$$n_+ - n_- \simeq \frac{Ne^2}{\epsilon_0 m\, n} \frac{\omega\Omega}{({\omega_0}^2 - \omega^2)^2}. \tag{4.60}$$

It follows that the rate at which the plane of polarization of an electromagnetic wave rotates with the distance traveled by the wave is given by

$$\frac{d\chi}{dl} = \frac{\kappa(\omega)\, NB_{\parallel}}{n(\omega)}, \tag{4.61}$$

where $B_{\|}$ is the component of the magnetic field along the direction of propagation of the wave, and

$$\kappa(\omega) = \frac{e^3}{2\epsilon_0 m^2 c} \frac{\omega^2}{(\omega_0^2 - \omega^2)^2}. \tag{4.62}$$

If the medium in question is a tenuous plasma then $n \simeq 1$ and $\omega_0 = 0$. Thus,

$$\frac{d\chi}{dl} \simeq \frac{e^3}{2\epsilon_0 m^2 c} \frac{N B_{\|}}{\omega^2} \tag{4.63}$$

Clearly, the rate at which the plane of polarization rotates is proportional to the product of the electron number density and the parallel magnetic field strength. Moreover, the plane of rotation rotates faster for low frequency waves than for high frequency waves. The total angle by which the plane of polarization is twisted after passing through a magnetized plasma is given by

$$\Delta\chi \simeq \frac{e^3}{2\epsilon_0 m^2 c\, \omega^2} \int N(l) B_{\|}(l)\, dl, \tag{4.64}$$

provided that N and $B_{\|}$ vary on length-scales which are large compared to the wavelength of the radiation. This formula is regularly employed in radio astronomy to infer the magnetic field-strength in interstellar space.

4.7 Wave propagation through a magnetized plasma

For a plasma ($\omega_0 = 0$) the dispersion relation (4.51) reduces to

$$n_{\pm}^{\,2}(\omega) = 1 - \frac{{\omega_p}^2}{\omega(\omega \mp \Omega)}. \tag{4.65}$$

The upper sign corresponds to a left-handed circularly polarized wave and the lower sign to a right-handed polarized wave. Of course, Eq. (4.65) is only valid for wave propagation along the direction of the magnetic field. Wave propagation through the Earth's ionosphere is well described by the above dispersion relation. There are wide frequency intervals where one of $n_+^{\,2}$ or $n_-^{\,2}$ is positive and the other negative. At such frequencies one state of circular polarization cannot propagate through the plasma. Consequently, a wave of that polarization incident on the plasma is totally reflected. The other state of polarization is partially transmitted.

The behaviour of $n_-^{\,2}(\omega)$ at low frequencies is responsible for a strange phenomenon known to radio hams as "whistlers." As the frequency tends to zero, Eq. (4.65) yields

$$n_-^{\,2} \simeq \frac{{\omega_p}^2}{\omega\,\Omega}. \tag{4.66}$$

At this sort of frequency n_+^2 is negative, so only right-hand polarized waves can propagate. The wave-number of such waves is given by

$$k_- = n_-\,\frac{\omega}{c} \simeq \frac{\omega_p}{c}\sqrt{\frac{\omega}{\Omega}}. \tag{4.67}$$

Energy transport is governed by the *group velocity* (see later)

$$v_g(\omega) = \frac{d\omega}{dk_-} \simeq 2c\,\frac{\sqrt{\omega\,\Omega}}{\omega_p}. \tag{4.68}$$

Thus, low frequency waves transmit energy *slower* than high frequency waves. A lightning strike in one hemisphere of the Earth generates a wide spectrum of radiation, some of which propagates along the dipolar field lines of the Earth's magnetic field in a manner described approximately by the dispersion relation (4.68). The high frequency components of the signal return to the surface of the Earth before the low frequency components (since they travel faster along the magnetic field). This gives rise to a radio signal which begins at a high frequency and then "whistles" down to lower frequencies.

4.8 The propagation of electromagnetic radiation through a dispersive medium

Let us now investigate the propagation of electromagnetic radiation through a dispersive medium by studying a simple one-dimensional problem. Suppose that our dispersive medium extends from $x = 0$, where it interfaces with a vacuum, to $x = \infty$. Suppose further that a wave is incident *normally* on the medium, so that the field quantities only depend on x and t. The wave is specified as a given function of t at $x = 0$. Since we are not interested in the reflected wave, let this function, $f(t)$, say, give the wave amplitude *just inside* the surface of the dispersive medium. Suppose that the wave arrives at this surface at $t = 0$, and that

$$f(t) = \begin{cases} 0 & \text{for } t < 0, \\ \sin\left(\frac{2\pi t}{\tau}\right) & \text{for } t \geq 0. \end{cases} \tag{4.69}$$

How does the wave subsequently develop in the region $x > 0$? In order to answer this question we must first of all decompose $f(t)$ into harmonic components of the form $\exp(-\mathrm{i}\,\omega t)$ (*i.e.*, Fourier harmonics). Unfortunately, if we attempt this using only real frequencies, ω, we encounter convergence difficulties, since $f(t)$ does not vanish at $t = \infty$. For the moment, we can circumvent these difficulties by only considering *finite* (in time) wave forms. In other words, we now imagine that $f(t) = 0$ for $t < 0$ *and* $t > T$. Such a wave form can be thought of as the superposition of two infinite (in time) wave forms, the first beginning at $t = 0$ and the second at $t = T$ with the opposite phase, so that the two cancel for all time $t > T$.

According to standard Fourier transform theory

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} f(t')\,\mathrm{e}^{-\mathrm{i}\,\omega(t-t')}\,dt'. \tag{4.70}$$

Since $f(t)$ is a real function of t which is zero for $t < 0$ and $t > T$, we can write

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \int_{0}^{T} f(t')\,\cos[\omega(t-t')\,]\,dt'. \tag{4.71}$$

Finally, it follows from symmetry (in ω) that

$$f(t) = \frac{1}{\pi}\int_{0}^{\infty} d\omega \int_{0}^{T} f(t')\,\cos[\omega(t-t')\,]\,dt'. \tag{4.72}$$

Equation (4.69) yields

$$f(t) = \frac{1}{\pi}\int_{0}^{\infty} d\omega \int_{0}^{T} \sin\!\left(\frac{2\pi t'}{\tau}\right)\,\cos[\omega(t-t')\,]\,dt', \tag{4.73}$$

or

$$f(t) \;=\; \frac{1}{2\pi}\int_0^\infty d\omega \left\{\frac{\cos[2\pi t'/\tau+\omega(t-t')\,]}{\omega-2\pi/\tau} - \frac{\cos[2\pi t'/\tau-\omega(t-t')\,]}{\omega+2\pi/\tau}\right\}_{t'=0}^{t'=T}. \tag{4.74}$$

Let us assume, for the sake of simplicity, that

$$T = N\tau, \tag{4.75}$$

where N is a positive integer. This ensures that $f(t)$ is continuous at $t = T$. Equation (4.74) reduces to

$$f(t) = \frac{2}{\tau}\int_0^\infty \frac{d\omega}{\omega^2-(2\pi/\tau)^2}\left(\cos[\omega(t-T)\,]-\cos\omega t\,\right). \tag{4.76}$$

This expression can be written

$$f(t) = \frac{1}{\tau}\int_{-\infty}^\infty \frac{d\omega}{\omega^2-(2\pi/\tau)^2}\left(\cos[\omega(t-T)\,]-\cos\omega t\,\right), \tag{4.77}$$

or

$$f(t) = \frac{1}{2\pi}\,\mathrm{Re}\!\int_{-\infty}^\infty \frac{d\omega}{\omega-2\pi/\tau}\left(\mathrm{e}^{-\mathrm{i}\,\omega(t-T)}-\mathrm{e}^{-\mathrm{i}\,\omega t}\right). \tag{4.78}$$

It is not entirely obvious that Eq. (4.78) is equivalent to Eq. (4.77). However, we can easily prove that this is the case by taking Eq. (4.78) and using the standard definition of a real part (*i.e.*, half the sum of the quantity in question and its complex conjugate) to give

$$\begin{aligned} f(t) \;&=\; \frac{1}{4\pi}\int_{-\infty}^\infty \frac{d\omega}{\omega-2\pi/\tau}\left(\mathrm{e}^{-\mathrm{i}\,\omega(t-T)}-\mathrm{e}^{-\mathrm{i}\,\omega t}\right)\\ &\quad+\frac{1}{4\pi}\int_{-\infty}^\infty \frac{d\omega}{\omega-2\pi/\tau}\left(\mathrm{e}^{+\mathrm{i}\,\omega(t-T)}-\mathrm{e}^{+\mathrm{i}\,\omega t}\right). \end{aligned} \tag{4.79}$$

Replacing the dummy integration variable ω by $-\omega$ in the second integral and then making use of symmetry, it is easily seen that the above expression reduces to Eq. (4.77).

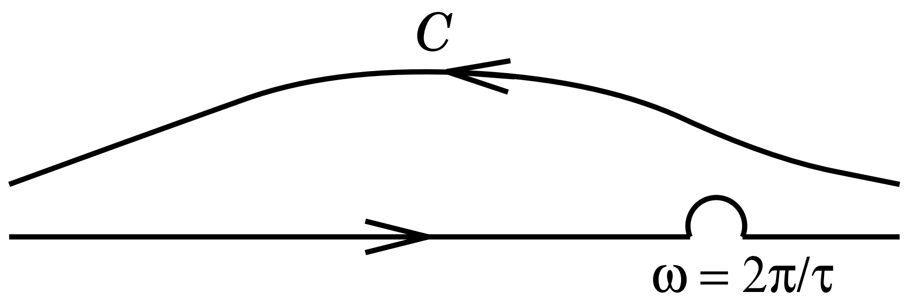


Figure 6: Sketch of the integration contours used to evaluate Eqs. (4.78) and (4.81)

Equation (4.77) can be written

$$f(t) = \frac{2}{\tau}\int_{-\infty}^{\infty} d\omega\, \sin[\omega(t-T/2)]\,\frac{\sin(\omega T/2)}{\omega^2-(2\pi/\tau)^2}. \tag{4.80}$$

Note that the integrand is finite at $\omega = 2\pi/\tau$, since at this point the vanishing of the denominator is compensated for by the simultaneous vanishing of the numerator. It follows that the integrand in Eq. (4.78) is also not infinite at $\omega = 2\pi/\tau$, as long as we do not separate the two exponentials. Thus, we can replace the integration along the real axis through this point by a small semi-circle in the upper half of the complex plane. Once this has been done, we can deform the path still further and can integrate the two exponentials in Eq. (4.78) separately:

$$f(t) = \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{-\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega-2\pi/\tau} - \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{-\mathrm{i}\,\omega(t-T)}\,\frac{d\omega}{\omega-2\pi/\tau} \tag{4.81}$$

The contour C is sketched in Fig. 6. Note that it runs from $+\infty$ to $-\infty$, which accounts for the change of sign between Eqs. (4.78) and (4.81).

We have already noted that a finite wave form which is zero for $t<0$ and $t>T$ can be through of as the superposition of two out of phase infinite wave forms, one starting at $t=0$ and the other at $t=T$. It is plausible, therefore, that the first term in the above expression corresponds to the infinite wave form starting at $t=0$, and the second to the infinite wave form starting at $t=T$. If this is the case then the signal (4.69), which starts at $t=0$ and ends at $t=\infty$, can be written in the form

$$f(t) = \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{-\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega - 2\pi/\tau}. \tag{4.82}$$

Let us test this proposition. In order to do this we must replace the original path of integration C by two equivalent paths.

First, consider $t < 0$. In this case, $-\mathrm{i}\,\omega t$ has a negative real part in the upper half plane which increases indefinitely with increasing distance from the axis. Thus, we can replace the original path of integration by the path A (see Fig. 7). The integral clearly vanishes along this path if we let A approach infinity in the upper half plane. Consequently,

$$f(t) = 0 \tag{4.83}$$

for $t < 0$.

Next, consider $t > 0$. Now, $-\mathrm{i}\,\omega t$ has a negative real part in the lower half plane, so that the exponential vanishes at infinity in this half plane. If we attempt to deform C to infinity in the lower half plane, the path of integration "catches" on the singularity of the integrand at $\omega = 2\pi/\tau$ (see Fig. 7). The path of integration B therefore consists of three parts: the part at infinity, B_1, where the integral vanishes due to the exponential factor $\mathrm{e}^{-\mathrm{i}\,\omega t}$; B_2, the two parts leading to infinity which cancel each other and thus contribute nothing to the integral; the path B_3 around the singularity. This latter contribution can easily be evaluated using the Cauchy residue theorem:

$$B_3 = \frac{1}{2\pi}\,\mathrm{Re}\,(2\pi\mathrm{i}\;\mathrm{e}^{-2\pi\mathrm{i}\,t/\tau}) = \sin\left(\frac{2\pi t}{\tau}\right). \tag{4.84}$$

Thus, it is proven that the expression (4.82) actually describes a wave form beginning at $t = 0$ whose subsequent motion is specified by Eq. (4.69).

Equation (4.82) can immediately be generalized to give the wave motion in the region $x > 0$:

$$f(x,t) = \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega - 2\pi/\tau}. \tag{4.85}$$

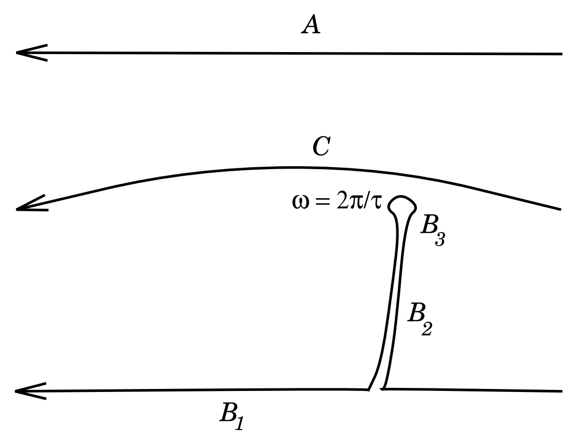


Figure 7: Sketch of the integration contours used to evaluate Eq. (4.82)

This follows from standard wave theory, because we know that an unterminated wave motion at $x = 0$ of the form $\mathrm{e}^{-\mathrm{i}\,\omega t}$ takes the form $\mathrm{e}^{\,\mathrm{i}\,(kx-\omega t)}$ after moving a distance x in the dispersive medium, provided that k and ω are related by the appropriate dispersion relation. For a medium consisting of a single resonant species this dispersion relation is written (see Eq. (4.17))

$$k^2 = \frac{\omega^2}{c^2}\left(1 + \frac{(N e^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}\right). \tag{4.86}$$

4.9 Propagation of the wave front in a dispersive medium

It is helpful to define

$$s = t - \frac{x}{c}. \tag{4.87}$$

Let us consider the two cases $s < 0$ and $s > 0$ separately.

Suppose that $s < 0$. In this case we distort the path C, used to evaluate the integral (4.85), into the path A shown in Fig. 8. This is only a sensible thing to do if the real part of $\mathrm{i}\,(kx - \omega t)$ is negative at infinity in the upper half plane. It is clear from the dispersion relation (4.86) that $k = \omega/c$ in the limit $|\omega| \to \infty$. Thus,

$$\mathrm{i}\,(kx - \omega t) = -\mathrm{i}\,\omega(t - x/c) = -\mathrm{i}\,\omega s. \tag{4.88}$$

It follows that $\mathrm{i}\,(kx-\omega t)$ possesses a large negative real part along path A provided that $s<0$. Thus, Eq. (4.85) yields

$$f(x,t)=0 \tag{4.89}$$

for $s<0$. In other words, *it is impossible for the wave front to propagate through the dispersive medium with a velocity greater than the velocity of light in a vacuum.*

Suppose that $s>0$. In this case we distort the path C into the *lower* half plane, since $\mathrm{i}\,(kx-\omega t)=-\mathrm{i}\,\omega s$ has a negative real part at infinity in this region. In doing this, the path becomes stuck not only at the singularity of the denominator when $\omega=2\pi/\tau$, but also at the branch points of the expression for k. After a little algebra, the dispersion relation (4.86) yields

$$k=\frac{\omega}{c}\sqrt{\frac{\omega_{1+}-\omega}{\omega_{0+}-\omega}}\sqrt{\frac{\omega_{1-}-\omega}{\omega_{0-}-\omega}}, \tag{4.90}$$

where

$$\omega_{0\pm}=-\mathrm{i}\,\rho\pm\sqrt{{\omega_0}^2-\rho^2}, \tag{4.91}$$

and

$$\omega_{1\pm}=-\mathrm{i}\,\rho\pm\sqrt{{\omega_0}^2+{\omega_p}^2-\rho^2}. \tag{4.92}$$

Here,

$$\omega_p=\sqrt{Ne^2/\epsilon_0 m} \tag{4.93}$$

is the plasma frequency, and

$$\rho=\frac{g\,\omega_0}{2}\ll\omega_0 \tag{4.94}$$

parameterizes the damping. In order to prevent multiple roots of Eq. (4.90) it is necessary to place branch cuts between ω_{0+} and ω_{1+} and also between ω_{0-} and ω_{1-} (see Fig. 8).

The path of integration B is conveniently split into the parts B_1 through B_5. The contribution from B_1 is negligible since the exponential in Eq. (4.85) is vanishingly small on this part of the integration path. Likewise, the contribution

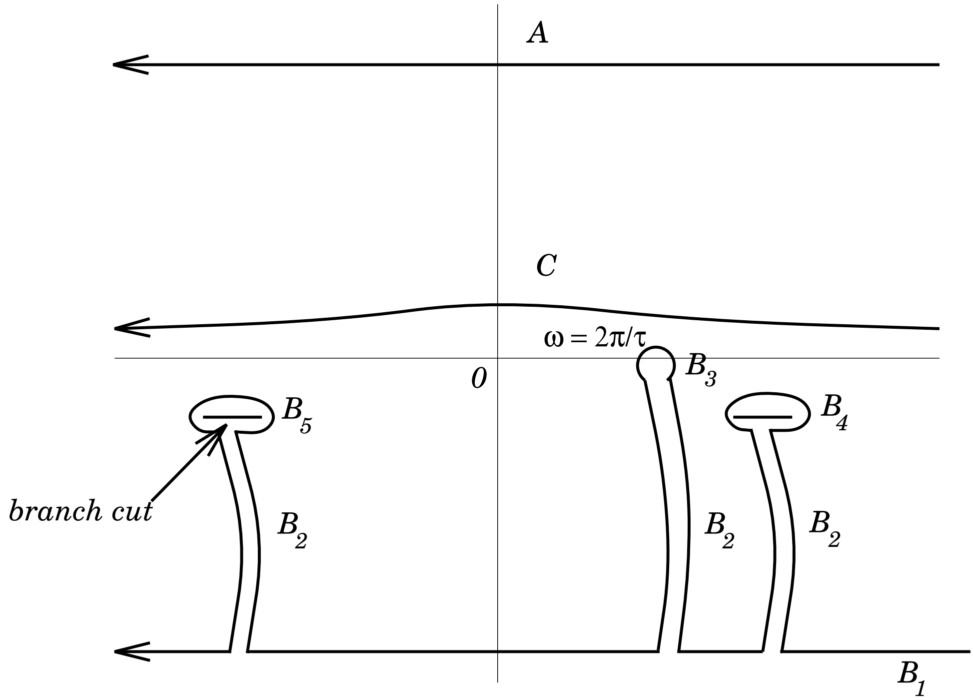


Figure 8: Sketch of the integration contours used to evaluate Eq. (4.85)

from B_2 is zero since its two sections always cancel. The contribution from B_3 follows from the residue theorem:

$$B_3 = \frac{1}{2\pi}\,\mathrm{Re}\,(2\pi\mathrm{i}\,\mathrm{e}^{\mathrm{i}\,[k_\tau x - 2\pi t/\tau]}). \tag{4.95}$$

Here, k_τ denotes the value of k obtained from the dispersion relation (4.86) in the limit $\omega \to 2\pi/\tau$. Thus,

$$B_3 = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x}\sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right). \tag{4.96}$$

In general, the contributions from B_4 and B_5 cannot be simplified further. For the moment we denote them as

$$B_4 = \frac{1}{2\pi}\,\mathrm{Re}\,\oint_{B_4}\mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega - 2\pi/\tau}, \tag{4.97}$$

and

$$B_5 = \frac{1}{2\pi}\,\mathrm{Re}\,\oint_{B_5}\mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega - 2\pi/\tau}, \tag{4.98}$$

where the paths of integration circle the appropriate branch cuts. In all, we have

$$f(x,t) = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x} \sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) + B_4 + B_5 \tag{4.99}$$

for $s > 0$.

Let us now look at the special case $s = 0$. For this value of s we can change the original path of integration to one at infinity in either the upper or the lower half plane, since the integrand vanishes in each case, through no longer exponentially, but rather as $1/\omega^2$. We can see this from Eq. (4.82), which can be written in the form

$$f(t) = \frac{1}{4\pi}\left(\int_C \mathrm{e}^{-\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega - 2\pi/\tau} + \int_C \mathrm{e}^{+\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega - 2\pi/\tau}\right). \tag{4.100}$$

Substitution of ω for $-\omega$ in the second integral yields

$$f(t) = \frac{1}{\tau}\int \mathrm{e}^{-\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega^2 - (2\pi/\tau)^2}. \tag{4.101}$$

Now, applying dispersion theory, we get from the above equation, just as we got Eq. (4.85) from Eq. (4.82),

$$f(x,t) = \frac{1}{\tau}\int \mathrm{e}^{\mathrm{i}\,(kx - \omega t)}\,\frac{d\omega}{\omega^2 - (2\pi/\tau)^2}. \tag{4.102}$$

Clearly, the integrand vanishes as $\mathrm{e}^{-\mathrm{i}\,\omega s}/\omega^2$ as ω becomes very large. Thus, it vanishes as $1/\omega^2$ for $s = 0$. Since we can calculate $f(x,t)$ by using either path A or path B, we can see that

$$f(x,t) = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x} \sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) + B_4 + B_5 = 0 \tag{4.103}$$

for $s = 0$. Thus, there is continuity in the transition from the region $s < 0$ to the region $s > 0$.

We are now in a position to make some meaningful statements about the behaviour of the signal at depth x inside the dispersive medium. Prior to the time $t = x/c$ there is no motion. Even if the phase velocity is superluminal, no electromagnetic signal can arrive earlier than one propagating with the velocity of light in vacuum c. The wave motion for $t > x/c$ is conveniently divided into two parts: *free oscillations* and *forced oscillations*. The former are given by $B_4 + B_5$, and the latter by

$$e^{-\mathrm{Im}(k_\tau)\,x}\sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) = e^{-\mathrm{Im}(k_\tau)\,x}\sin\left(\frac{2\pi}{\tau}\left[t - \frac{x}{v_p}\right]\right), \tag{4.104}$$

where

$$v_p = \frac{2\pi}{\tau\,\mathrm{Re}(k_\tau)} \tag{4.105}$$

is termed the *phase velocity*. The forced oscillations have the same sine wave characteristics and oscillation frequency as the incident wave. However, the wave amplitude is diminished by the damping coefficient, although, as we have seen, this is generally a negligible effect unless the frequency of the incident wave closely matches one of the resonant frequencies of the dispersive medium. The phase velocity v_p determines the velocity with which a point of constant phase (*e.g.*, a peak or trough) of the forced oscillation signal propagates into the medium. However, *the phase velocity has no effect on the velocity with which the forced oscillation wave front propagates into the medium.* This latter velocity is equivalent to the velocity of light in vacuum c . The phase velocity v_p can be either greater or less than c, in which case peaks and troughs either catch up with or fall further behind the wave front. Of course, peaks can never overtake the wave front.

It is clear from Eqs. (4.91), (4.92), (4.97), and (4.98) that the free oscillations oscillate with real frequencies which are somewhere between the resonant frequency ω_0 and the plasma frequency ω_p. Furthermore, the free oscillations are *damped* in time like $\exp(-\rho\,t)$. The free oscillations, like the forced oscillations, begin at time $t = x/c$. At $t = x/c$ the free and forced oscillations just cancel (see Eq. (4.103)). As t increases both the free and forced oscillations set in, but the former rapidly damp away, leaving only the forced oscillations. Thus, the free oscillations can be regarded as some sort of *transient* response of the medium to the incident wave, whereas the forced oscillations determine the time asymptotic response. The real frequency of the forced oscillations is that imposed externally by the incident wave, whereas the real frequency of the free oscillations is determined by the nature of the dispersive medium, quite independently of the frequency of the incident wave.

One slightly surprising result of the above analysis is the prediction that the wave front of the signal propagates into the dispersive medium with the velocity of light in vacuum, irrespective of the dispersive properties of the medium. Actually, this is a fairly obvious result. As is well described by Feynman in his famous *Lectures on Physics*, when an electromagnetic wave propagates through a dispersive medium, the electrons and ions which make up that medium oscillate in sympathy with the incident wave and in doing so emit radiation. Both the radiation from the electrons and ions and the incident radiation travel at the velocity c. However, when these two radiation signals are superposed the net effect is as if the incident signal propagates through the dispersive medium with a phase velocity which is different from c. Consider the wave front of the incident signal, which clearly propagates into the medium with the velocity c. Prior to the arrival of this wave front the electrons and ions are at rest, since no information regarding the arrival of the incident wave at the surface of medium can propagate faster than c. After the arrival of the wave front the electrons and ions are set into motion and emit radiation which can affect the apparent phase velocity of radiation which arrives somewhat later. But this radiation certainly cannot affect the propagation velocity of the wave front itself, which has already passed by the time the electrons and ions are set into motion (because of the finite inertia of the electrons and ions).

4.10 The Sommerfeld precursor

Let us consider the situation immediately after the arrival of the signal; *i.e.*, when s is small and positive. Let us start from Eq. (4.102), which can be written in the form

$$f(x,t) = \frac{1}{\tau}\int_C \mathrm{e}^{\,\mathrm{i}\,([k-\omega/c]x-\omega s)}\,\frac{d\omega}{\omega^2-(2\pi/\tau)^2}. \tag{4.106}$$

We can deform the original path of integration C into a large semi-circle of radius R in the upper half-plane, plus the segments of the real axis, as shown in Fig. 9. Because of the denominator $\omega^2 - (2\pi/\tau)^2$, the integrand tends to zero as $1/\omega^2$ on the real axis. We may add the path in the lower half plane which is shown as a dotted line in the figure, for if the radius of the semi-circular portion of this lower path is increased to infinity, the integrand vanishes exponentially because

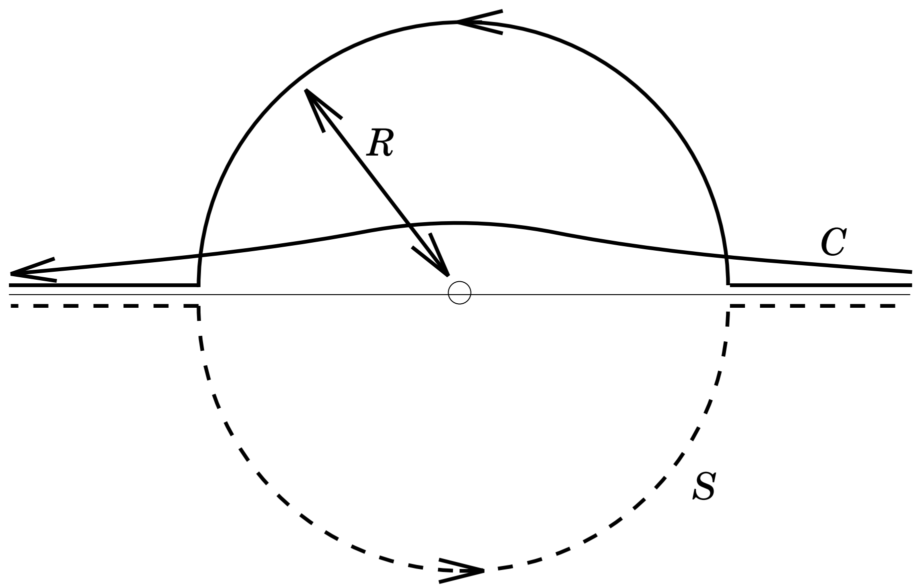


Figure 9: Sketch of the integration contour used to evaluate Eq. (4.107)

$s > 0$. Therefore, we may replace our original path of integration by the entire circle S. Thus,

$$f(x,t) = \frac{1}{\tau}\oint_S \mathrm{e}^{\,\mathrm{i}\,([k-\omega/c]x-\omega s)}\,\frac{d\omega}{\omega^2-(2\pi/\tau)^2} \tag{4.107}$$

in the limit that the radius of the circle R tends to infinity.

The dispersion relation (4.86) yields

$$k - \frac{\omega}{c} \simeq \frac{\omega}{c}\left(\sqrt{1-\frac{{\omega_p}^2}{\omega^2}}-1\right) \simeq -\frac{{\omega_p}^2}{2\,c\,\omega} \tag{4.108}$$

in the limit $|\omega| \to \infty$. Using the abbreviation

$$\xi = \frac{{\omega_p}^2}{2\,c}\,x, \tag{4.109}$$

and henceforth neglecting $2\pi/\tau$ with respect to ω, we obtain from Eq. (4.107)

$$f(x,t) = f_1(\xi,t) \simeq \frac{1}{\tau}\oint_S \exp\left[\mathrm{i}\left(-\frac{\xi}{\omega}-\omega\,s\right)\right]\frac{d\omega}{\omega^2}. \tag{4.110}$$

This expression can also be written

$$f_1(\xi,t) = \frac{1}{\tau}\oint_S \exp\left[-\mathrm{i}\sqrt{\xi\, s}\left(\frac{1}{\omega}\sqrt{\frac{\xi}{s}} + \omega\sqrt{\frac{s}{\xi}}\right)\right]\frac{d\omega}{\omega^2}. \tag{4.111}$$

Let

$$\omega\sqrt{\frac{s}{\xi}} = \mathrm{e}^{\mathrm{i}\,u}. \tag{4.112}$$

It follows that

$$\frac{d\omega}{\omega} = \mathrm{i}\, du, \tag{4.113}$$

giving

$$\frac{d\omega}{\omega^2} = \mathrm{i}\sqrt{\frac{s}{\xi}}\,\mathrm{e}^{-\mathrm{i}\,u}\, du. \tag{4.114}$$

Substituting the angular variable u for ω as the integration variable in Eq. (4.111) yields

$$f_1(\xi,t) = \frac{\mathrm{i}}{\tau}\sqrt{\frac{s}{\xi}}\int_0^{2\pi}\exp(-2\mathrm{i}\sqrt{\xi s}\,\cos u)\,\mathrm{e}^{-\mathrm{i}\,u}\, du. \tag{4.115}$$

Here, we have taken $\sqrt{\xi/s}$ as the radius of the circular integration path in the ω-plane. This is indeed a large radius, since $s \ll 1$. From symmetry, Eq. (4.115) simplifies to

$$f_1(\xi,t) = \frac{\mathrm{i}}{\tau}\sqrt{\frac{s}{\xi}}\int_0^{2\pi}\exp(-2\mathrm{i}\sqrt{\xi s}\,\cos u)\,\cos u\, du. \tag{4.116}$$

The following mathematical identity is very well-known[11]

$$J_n(z) = \frac{\mathrm{i}^{-n}}{2\pi}\int_0^{2\pi}\mathrm{e}^{\mathrm{i}\,z\,\cos\theta}\cos(n\theta)\, d\theta, \tag{4.117}$$

where $J_n(z)$ is Bessel function of order n. It follows from Eq. (4.115) that

$$f_1(\xi,t) = \frac{2\pi}{\tau}\sqrt{\frac{s}{\xi}}\, J_1(2\sqrt{\xi s}). \tag{4.118}$$

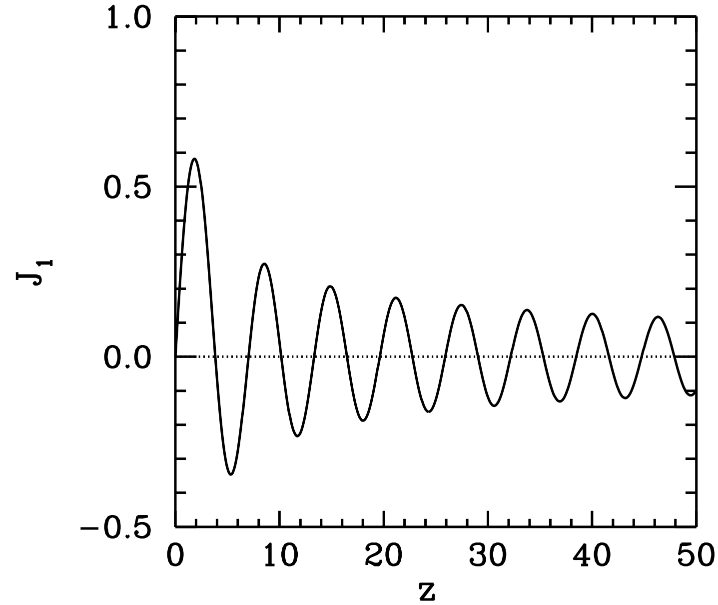


Figure 10: The Bessel function $J_1(z)$

Here, we have made use of the fact that $J_1(-z) = -J_1(z)$.

The properties of Bessel functions are well-known and are listed in many standard references on mathematical functions (see, for instance, Abramowitz and Stegun). In the small argument limit $z \ll 1$ we find that

$$J_1(z) = \frac{z}{2} + O(z^3). \tag{4.119}$$

On the other hand, in the large argument limit $z \gg 1$ we obtain

$$J_1(z) = \sqrt{\frac{2}{\pi z}}\cos(z - 3\pi/4) + O(z^{-3/2}). \tag{4.120}$$

The behaviour of $J_1(z)$ is further illustrated in Fig. 10.

We are now in a position to make some quantitative statements regarding the signal which first arrives at depth x in the dispersive medium. This signal propagates at the velocity of light in vacuum and is called the *Sommerfeld*

[11] M. Abramowitz, and I.A. Stegun, *Handbook of mathematical functions*, (Dover, New York, 1965), Eq. 9.1.21.

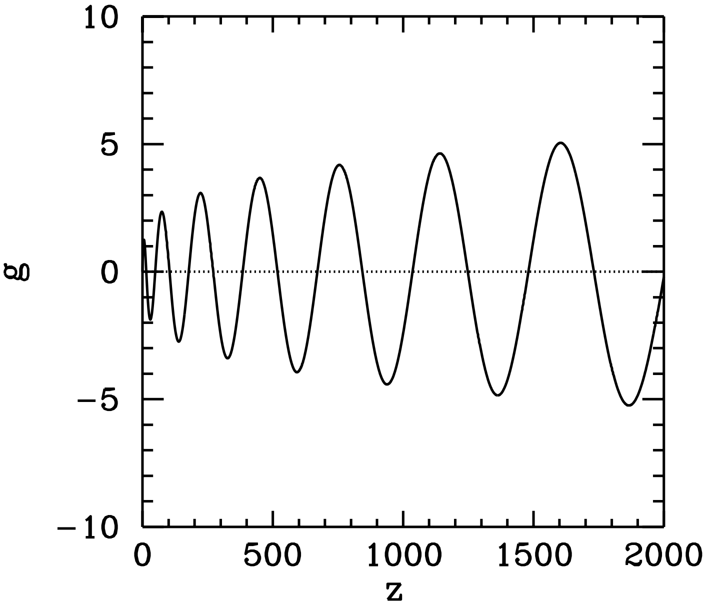


Figure 11: The Sommerfeld precursor

precursor. The first important point to note is that the amplitude of the Sommerfeld precursor is very small compared to that of the incident wave (whose amplitude is normalized to unity). We can easily see this because in deriving Eq. (4.118) we assumed that $|\omega| = \sqrt{\xi/s} \gg 2\pi/\tau$ on the circular integration path S. Since the magnitude of J_1 is always less than, or of order, unity, it is clear that $|f_1| \ll 1$. This is a comforting result, since in a naive treatment of wave propagation through a dielectric medium the wave front propagates at the group velocity v_g (which is usually less than c) and, therefore, no signal should reach depth x in the medium before time x/v_g. We are finding that there is, in fact, a precursor which arrives at $t = x/c$, but that this signal is fairly small. Note from Eq. (4.109) that ξ is proportional to x. Clearly, the amplitude of the Sommerfeld precursor decreases like one over the distance traveled by the wave front through the dispersive medium (since J_1 attains its maximum value when $s \sim 1/\xi$). Thus, the Sommerfeld precursor is likely to become undetectable after the wave has traveled a long distance through the medium.

Equation (4.118) can be written

$$f_1(\xi,t) = \frac{\pi}{\xi\,\tau}\,g(s/s_0), \tag{4.121}$$

where $s_0 = 1/4\,\xi$, and

$$g(z) = \sqrt{z}\,J_1(\sqrt{z}). \tag{4.122}$$

The normalized Sommerfeld precursor $g(z)$ is shown in Fig. 11. It can be seen that both the amplitude and the oscillation period of the precursor gradually increase. The roots of $J_1(z)$ [*i.e.*, the solutions of $J_1(z) = 0$] are spaced at distances of approximately π apart. Thus, the time interval for the mth half period of the precursor is approximately given by

$$\Delta t_m \sim \frac{m\pi^2}{2\xi}. \tag{4.123}$$

Note that the initial period of oscillation,

$$\Delta t_0 \sim \frac{\pi^2}{2\xi}, \tag{4.124}$$

is extremely small compared to the incident period τ. Moreover, the initial period of oscillation is *completely independent* of the frequency of the incident wave. In fact, Δt_0 depends only on the depth x and on the dispersive power of the medium. The period decreases with increasing distance x traveled by the wave front though the medium. So, when visible radiation is incident on some dispersive medium it is quite possible for the first signal detected well inside the medium to lie in the X-ray region of the electromagnetic spectrum.

4.11 The method of stationary phase

Equation (4.102) can be written in the form

$$f(x,t) = \int_C \mathrm{e}^{\,\mathrm{i}\,\phi(\omega)}\,F(\omega)\,d\omega \tag{4.125}$$

where

$$F(\omega) = \frac{1}{\tau} \frac{1}{\omega^2 - (2\pi/\tau)^2}, \tag{4.126}$$

and

$$\phi(\omega) = k(\omega)\, x - \omega t. \tag{4.127}$$

It is clear that $F(\omega)$ is a relatively slowly varying function of ω (except in the immediate vicinity of the singular points $\omega = \pm 2\pi/\tau$), whereas the phase $\phi(\omega)$ is generally large and rapidly varying. The rapid oscillations of $\exp(\mathrm{i}\,\phi)$ over most of the range of integration means that the integrand averages to almost zero. Exceptions to this cancellation rule occur only when $\phi(\omega)$ is *stationary*; *i.e.*, when $\phi(\omega)$ has an extremum. The integral can therefore be estimated by finding places where $\phi(\omega)$ has a vanishing derivative, evaluating (approximately) the integral in the neighbourhood of each of these points, and summing the contributions. This procedure is called the *method of stationary phase*.

Suppose that $\phi(\omega)$ has a vanishing first derivative at $\omega = \omega_s$. In the neighbourhood of this point, $\phi(\omega)$ can be expanded as a Taylor series,

$$\phi(\omega) = \phi_s + \frac{1}{2}\phi_s''(\omega - \omega_s)^2 + \cdots. \tag{4.128}$$

Here, the subscript s is used to indicate ϕ or its second derivative evaluated at $\omega = \omega_s$. Since $F(\omega)$ is slowly varying, the contribution to the integral from this stationary phase point is approximately

$$f_s \simeq F(\omega_s)\, e^{\mathrm{i}\,\phi_s} \int_{\infty}^{-\infty} \mathrm{e}^{(\mathrm{i}/2)\phi_s''(\omega-\omega_s)^2}\, d\omega. \tag{4.129}$$

It is tacitly assumed that the stationary point lies on the real axis in ω-space, so that locally the integral along the contour C is an integral along the real axis in the direction of decreasing ω. The above expression can be written in the form

$$f_s \simeq -F(\omega_s)\, \mathrm{e}^{\mathrm{i}\,\phi_s} \sqrt{\frac{4\pi}{\phi_s''}} \int_0^{\infty} \left[\cos(\pi t^2/2) + \mathrm{i}\,\sin(\pi t^2/2)\right]\, dt, \tag{4.130}$$

where

$$\frac{\pi}{2}\, t^2 = \frac{1}{2}\, \phi_s''\, (\omega - \omega_s)^2. \tag{4.131}$$

The integrals in the above expression are, *Fresnel integrals*[12] and can be shown to take the values

$$\int_0^\infty \cos(\pi t^2/2)\, dt = \int_0^\infty \sin(\pi t^2/2)\, dt = \frac{1}{2}. \tag{4.132}$$

It follows that

$$f_s \simeq -\sqrt{\frac{2\pi\,\mathrm{i}}{\phi_s''}}\, F(\omega_s)\, \mathrm{e}^{\,\mathrm{i}\,\phi_s}. \tag{4.133}$$

It is easily seen that the arc length (in ω-space) of the integration contour which makes a significant contribution to f_s is of order $\Delta\omega/\omega_s \sim 1/\sqrt{k(\omega_s)\,x}$. Thus, the arc length is relatively short provided that the wavelength of the signal is much less than the distance propagated through the dispersive medium. If there is more than one point of stationary phase in the range of integration then the integral is approximated as a sum of terms like the above.

Integrals of the form (4.125) can be calculated exactly using the *method of steepest decent*.[13] The stationary phase approximation (4.133) agrees with the leading term of the method of steepest decent (which is far more difficult to implement than the method of stationary phase) provided that $\phi(\omega)$ is real (*i.e.*, provided that the stationary point lies on the real axis). If ϕ is complex, however, the stationary phase method can yield erroneous results. This suggests that the stationary phase method is likely to break down when the extremum point $\omega = \omega_s$ approaches any poles or branch cuts in the ω-plane (see Fig. 8).

4.12 The group velocity

The point of stationary phase, defined by $\partial\phi/\partial\omega = 0$, satisfies the condition

$$\frac{c}{v_g} = \frac{ct}{x}, \tag{4.134}$$

[12] M. Abramowitz, and I.A. Stegun, *Handbook of mathematical functions*, (Dover, New York, 1965), Sec. 7.3.

[13] Léon Brillouin, *Wave propagation and group velocity*, (Academic press, New York, 1960).

where

$$v_g = \frac{d\omega}{dk} \tag{4.135}$$

is conventionally termed the *group velocity*. Thus, the signal seen at position x and time t is dominated by the frequency range whose group velocity v_g is equal to x/t. In this respect, the signal incident at the surface of the medium ($x = 0$) at time $t = 0$ can be said to propagate through the medium at the group velocity $v_g(\omega)$.

The simple one-resonance dispersion relation (4.86) yields

$$\frac{c}{v_g} \simeq n(\omega)\left[1 + \frac{\omega^2}{{\omega_0}^2 - \omega^2} + \frac{\omega^2}{\omega^2 - {\omega_0}^2 - {\omega_p}^2}\right] \tag{4.136}$$

in the limit $g \to 0$, where

$$n(\omega) = \frac{ck}{\omega} = \sqrt{\frac{{\omega_0}^2 + {\omega_p}^2 - \omega^2}{{\omega_0}^2 - \omega^2}}. \tag{4.137}$$

The variation of c/v_g and the refractive index n with frequency is sketched in Fig. 12. With $g = 0$ the group velocity is less than c for all ω, except for $\omega_0 < \omega < \omega_1 \equiv \sqrt{{\omega_0}^2 + {\omega_p}^2}$, where it is purely imaginary. Note that the refractive index is also complex in this frequency range. The phase velocity $v_p = c/n$ is subluminal for $\omega < \omega_0$, imaginary for $\omega_0 \leq \omega \leq \omega_1$, and superluminal for $\omega > \omega_1$.

The frequency range which contributes to the amplitude at time t is determined graphically by finding the intersection of a horizontal line with ordinate ct/x with the solid curve in Fig. 12. There is no crossing of the two curves for $t < t_0 \equiv x/c$, thus no signal arrives before this time. For times immediately following t_0 the point of stationary phase is seen to be at $\omega \to \infty$. In this large ω limit the point of stationary phase is given by

$$\omega_s \simeq \omega_p \sqrt{\frac{t_0}{2(t - t_0)}}. \tag{4.138}$$

Note that $\omega = -\omega_s$ is also a point of stationary phase. It is easily demonstrated that

$$\phi_s \simeq -2\sqrt{\xi(t - t_0)}, \tag{4.139}$$

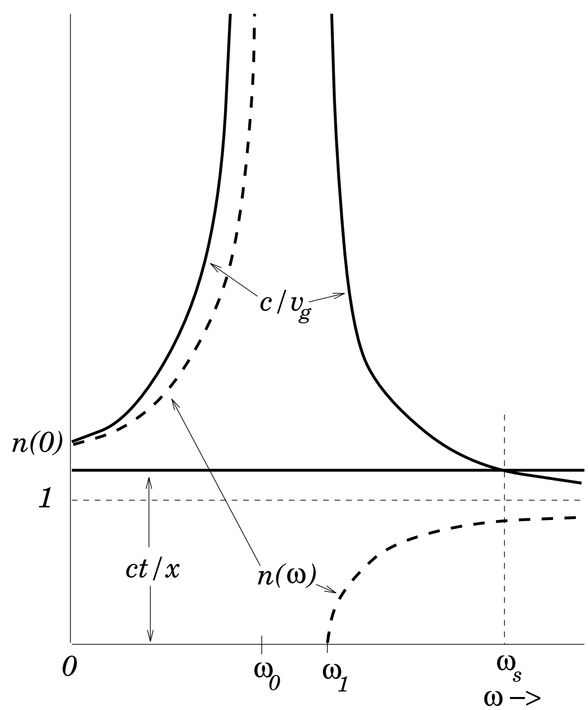


Figure 12: The typical variation of the functions $c/v_g(\omega)$ and $n(\omega)$. Here, $\omega_1 = ({\omega_0}^2 + {\omega_p}^2)^{1/2}$.

and

$$\phi_s'' \simeq -2\,\frac{(t-t_0)^{3/2}}{\xi^{1/2}}, \tag{4.140}$$

with

$$F(\omega_s) \simeq \frac{t-t_0}{\tau\,\xi}. \tag{4.141}$$

Here, ξ is given by Eq. (4.109). The stationary phase approximation (4.133) gives

$$f_s \simeq \sqrt{\frac{\pi\,\xi^{1/2}}{(t-t_0)^{3/2}}}\,\frac{t-t_0}{\tau\,\xi}\,\mathrm{e}^{-2\mathrm{i}\sqrt{\xi(t-t_0)}+3\pi\mathrm{i}/4} + \text{c.c.}, \tag{4.142}$$

where c.c. denotes the complex conjugate of the preceding term (this contribution comes from the second point of stationary phase located at $\omega = -\omega_s$). The above expression reduces to

$$f_s \simeq \frac{2\sqrt{\pi}}{\tau}\,\frac{(t-t_0)^{1/4}}{\xi^{3/4}}\,\cos\!\left[2\sqrt{\xi(t-t_0)} - 3\pi/4\right]. \tag{4.143}$$

It is easily demonstrated that the above formula is the same as the expression (4.118) for the Sommerfeld precursor in the large argument limit $t - t_0 \gg 1/\xi$. Thus, the method of stationary phase yields an expression for the Sommerfeld precursor which is accurate at all times except those immediately following the first arrival of the signal.

4.13 The Brillouin precursor

As time progresses the horizontal line ct/x in Fig. 12 gradually rises and the point of stationary phase moves to ever lower frequencies. In general, however, the amplitude remains relatively small. Only when the elapsed time reaches

$$t_1 = \frac{n(0)\,x}{c} > t_0 \tag{4.144}$$

is there a qualitative change. This time marks the arrival of a second precursor known as the *Brillouin precursor*. The reason for the qualitative change is evident from Fig. 12. At $t = t_1$ the lower region of the c/v_g curve is intersected for the first time, and $\omega = 0$ becomes a point of stationary phase. It is clear that the oscillation frequency of the Brillouin precursor is far less than that of the Sommerfeld precursor. Moreover, it is easily demonstrated that the second derivative of $k(\omega)$ vanishes at $\omega = 0$. This means that $\phi_s'' = 0$. The stationary phase result (4.133) gives an infinite answer in such circumstances. Of course, the amplitude of the Brillouin precursor is not infinite, but it is significantly larger than that of the Sommerfeld precursor.

In order to generalize the result (4.133) to deal with a stationary phase point at $\omega = 0$ it is necessary to expand $\phi(\omega)$ about this point, keeping terms up to ω^3. Thus,

$$\phi(\omega) \simeq \omega(t_1 - t) + \frac{x}{6}\, k_0'''\, \omega^3, \tag{4.145}$$

where

$$k_0''' \equiv \left(\frac{d^3k}{d\omega^3}\right)_{\omega=0} = \frac{3\,{\omega_p}^2}{c\, n(0)\, {\omega_0}^4} \tag{4.146}$$

for the simple dispersion relation (4.86). The amplitude (4.125) is therefore given approximately by

$$f(x,t) \simeq F(0) \int_{\infty}^{-\infty} \mathrm{e}^{\mathrm{i}\,\omega(t_1 - t) + \mathrm{i}\,(x/6) k_0''' \omega^3}\, d\omega. \tag{4.147}$$

This expression reduces to

$$f(x,t) = \frac{\tau}{\sqrt{2}\,\pi^2} \sqrt{\frac{|t - t_1|}{x\, k_0'''}} \int_0^{\infty} \cos\left[\frac{3}{2} z \left(\frac{v^3}{3} \pm v\right)\right] dv, \tag{4.148}$$

where

$$v = \sqrt{\frac{x\, k_0'''}{2\,|t - t_1|}}\, \omega, \tag{4.149}$$

and

$$z = \frac{2\sqrt{2}\,|t - t_1|^{3/2}}{3\sqrt{x\, k_0'''}}. \tag{4.150}$$

The positive (negative) sign in the integrand is taken for $t < t_1$ ($t > t_1$).

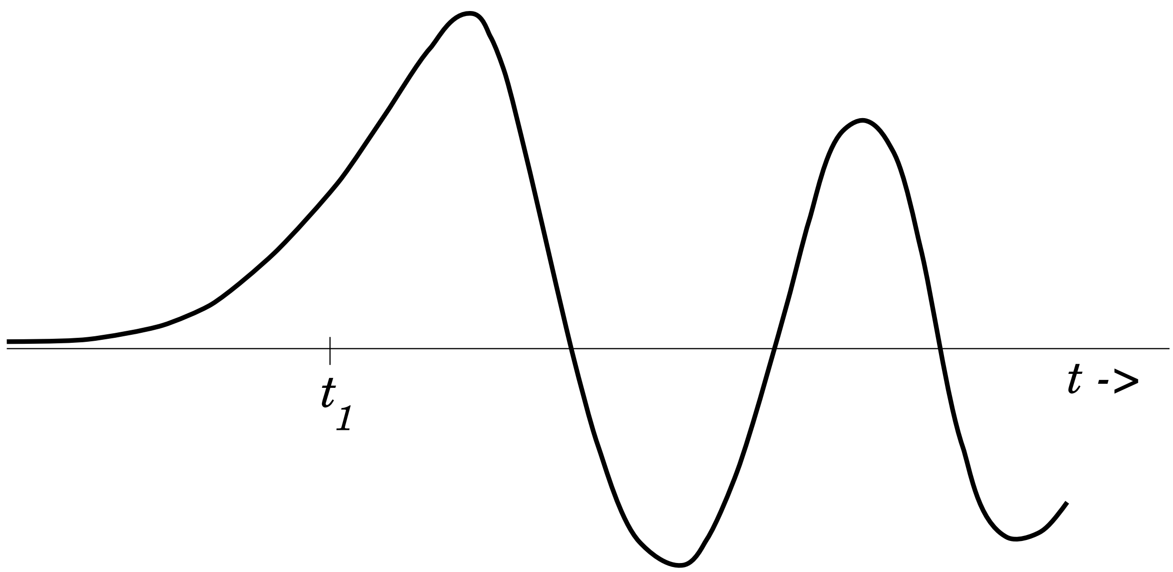


Figure 13: A sketch of the behaviour of the Brillouin precursor as a function of time

The integral in Eq. (4.150) is known as an *Airy integral.* It can be expressed in terms of Bessel functions of order 1/3, as follows:

$$\int_0^\infty \cos\left[\frac{3}{2}z\left(\frac{v^3}{3}+v\right)\right]dv = \frac{1}{\sqrt{3}}\,K_{1/3}(z), \tag{4.151}$$

and

$$\int_0^\infty \cos\left[\frac{3}{2}z\left(\frac{v^3}{3}-v\right)\right]dv = \frac{\pi}{3}\left[J_{1/3}(z)+J_{-1/3}(z)\right]. \tag{4.152}$$

From the well-known properties of Bessel functions the precursor can be seen to have a growing exponential character for times earlier than $t = t_1$, and an oscillating character for $t > t_1$. The amplitude in the neighbourhood of $t = t_1$ is plotted in Fig. 13.

The initial oscillation period of the Brillouin precursor is crudely estimated (from $z \sim 1$) as

$$\Delta t_0 \sim (x\,k_0''')^{1/3}. \tag{4.153}$$

The amplitude of the Brillouin precursor is approximately

$$|f| \sim \frac{\tau}{(x\,k_0''')^{1/3}}. \tag{4.154}$$

Let us adopt the ordering

$$1/\tau \sim \omega_0 \sim \omega_p \ll \xi, \tag{4.155}$$

which corresponds to most physical situations involving the propagation of electromagnetic radiation through dielectric media. It follows from the above results, plus the results of Section 4.10, that

$$(\Delta t_0\,\omega_p)_{\rm brillouin} \sim \left(\frac{\xi}{\omega_p}\right)^{1/3} \gg 1, \tag{4.156}$$

and

$$(\Delta t_0\,\omega_p)_{\rm sommerfeld} \sim \left(\frac{\omega_p}{\xi}\right) \ll 1. \tag{4.157}$$

Furthermore,

$$|f|_{\rm brillouin} \sim \left(\frac{\omega_p}{\xi}\right)^{1/3} \ll 1, \tag{4.158}$$

and

$$|f|_{\rm sommerfeld} \sim \left(\frac{\omega_p}{\xi}\right) \ll |f|_{\rm brillouin}. \tag{4.159}$$

It is clear that the Sommerfeld precursor is a low amplitude, high frequency signal, whereas the Brillouin precursor is a higher amplitude, low frequency signal. Note that the amplitude of the Brillouin precursor, whilst it is significantly higher than that of the Sommerfeld precursor, is still much less than that of the incident wave.

4.14 Signal arrival

Let us try to establish at what time t_2 a signal first arrives at position x inside the dielectric medium whose amplitude is comparable with that of the wave incident at time $t = 0$ on the surface of the medium ($x = 0$). Let us term this event the "arrival" of the signal. It is plausible from the discussion in Section 4.11 regarding the stationary phase approximation that signal arrival corresponds to the situation where the point of stationary phase in ω-space corresponds to a pole of the function $F(\omega)$. In other words, when ω_s approaches the frequency

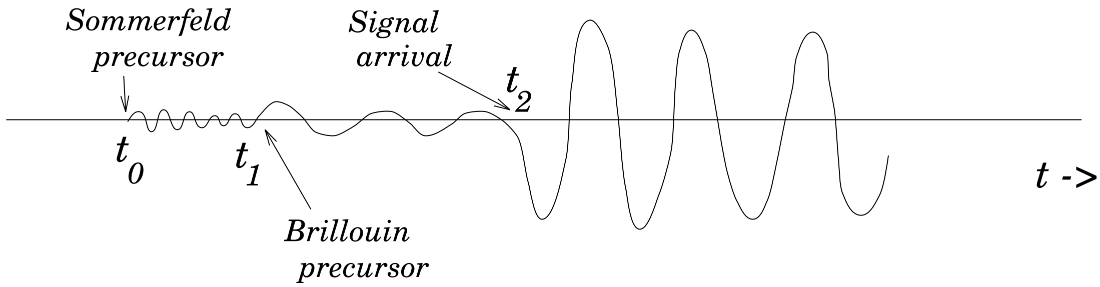


Figure 14: A sketch of the signal amplitude as a function of time as seen inside some dielectric medium subject to an incident wave which starts at some specific time

$2\pi/\tau$ of the incident signal. It is certainly the case that the stationary phase approximation yields a particularly large amplitude signal when $\omega_s \rightarrow 2\pi/\tau$. Unfortunately, as has already been discussed, the method of stationary phase becomes inaccurate under these circumstances. However, calculations involving the more robust method of steepest decent[14] confirm that in most cases the signal amplitude first becomes significant when $\omega_s = 2\pi/\tau$. Thus, the signal arrival time is

$$t_2 = \frac{x}{v_g(2\pi/\tau)}, \tag{4.160}$$

where $v_g(2\pi/\tau)$ is the group velocity calculated using the frequency of the incident signal. It is clear from Fig. 12 that

$$t_0 < t_1 < t_2. \tag{4.161}$$

Thus, the main signal arrives later than the Sommerfeld and Brillouin precursors.

The final picture which emerges from our investigations is summarized in Fig. 14. The main signal arrives at the group velocity corresponding to the frequency of the incident wave. However, it is possible to detect the arrival of the signal before this, given sufficiently accurate detection equipment. In fact, the first information regarding the arrival of the incident wave at the vacuum/dielectric interface propagates at the velocity of light in a vacuum.

[14] Léon Brillouin, *Wave propagation and group velocity*, (Academic press, New York, 1960).

4.15 The propagation of radio waves through the ionosphere

We have studied the *transient* behaviour of an electromagnetic wave incident on a spatially *uniform* dielectric medium in great detail. Let us now consider a quite different, but equally important, problem. What is the time asymptotic *steady-state* behaviour of an electromagnetic wave propagating though a spatially *non-uniform* dielectric medium?

As a specific example, let us consider the propagation of radio waves through the Earth's ionosphere. The refractive index of the ionosphere can be written [see Eq. (4.27)]

$$n^2 = 1 - \frac{{\omega_p}^2}{\omega(\omega + \mathrm{i}\,\nu)}, \tag{4.162}$$

where ν is a real positive constant which parameterizes the damping of electron motion (in fact, ν is the collision frequency of free electrons with other particles in the ionosphere), and

$$\omega_p = \sqrt{\frac{N e^2}{\epsilon_0 m}} \tag{4.163}$$

is the plasma frequency. In the above formula, N is the density of free electrons in the ionosphere and m is the electron mass. We shall assume that the ionosphere is horizontally stratified, so that $N = N(z)$, where the coordinate z measures height above the Earth's surface (*n.b.*, the curvature of the Earth is neglected in the following analysis). The ionosphere actually consists of two main layers; the E-layer, and the F-layer. We shall concentrate on the lower E-layer, which lies about 100 km above the surface of the Earth, and is about 50 km thick. The typical day-time number density of free electrons in the E-layer is $N \sim 3\times 10^{11}\,\mathrm{m}^{-3}$. At night-time, the density of free electrons falls to about half this number. The typical day-time plasma frequency of the E-layer is, therefore, about 5 MHz. The typical collision frequency of free electrons in the E-layer is about 0.05 MHz. According to simplistic theory, any radio wave whose frequency lies below the day-time plasma frequency, 5 MHz, (*i.e.*, any wave whose wavelength exceeds about 60 m) is reflected by the ionosphere during the day. Let us investigate in more detail exactly how this process takes place. Note, incidentally, that for mega-Hertz frequency radio waves $\nu \ll \omega$, so it follows from Eq. (4.162) that n^2 is predominately real (*i.e.*, under most circumstances, the electron collisions can be neglected).

The problem of radio wave propagation through the ionosphere was of great practical importance during the first half of the 20th Century, since at that time long-wave radio waves were the principle means of military communication. Nowadays, the military have far more reliable ways of communicating. Nevertheless, this subject area is still worth studying because the principle tool used to deal with the problem of wave propagation through a non-uniform medium, the so-called W.K.B. approximation, is of great theoretical importance. In particular, the W.K.B. approximation is very widely used in quantum mechanics (in fact, there is a great similarity between the problem of wave propagation through a non-uniform medium and the problem of solving Schrödinger's equation in the presence of a non-uniform potential).

Maxwell's equations for a wave propagating through a non-uniform, unmagnetized, dielectric medium are:

$$\nabla\cdot\boldsymbol{E} = 0, \tag{4.164a}$$

$$\nabla\cdot c\boldsymbol{B} = 0, \tag{4.164b}$$

$$\nabla\wedge\boldsymbol{E} = \mathrm{i}\,k\,c\boldsymbol{B}, \tag{4.164c}$$

$$\nabla\wedge c\boldsymbol{B} = -\mathrm{i}\,k\,n^2\,\boldsymbol{E}, \tag{4.164d}$$

where n is the non-uniform refractive index of the medium. It is assumed that all field quantities vary in time like $\mathrm{e}^{-\mathrm{i}\,\omega t}$, where $\omega = kc$. Note that, in the following, k is the wavenumber in free space, rather than the wavenumber in the dielectric medium.

4.16 The W.K.B. approximation

Consider a radio wave which is vertically incident, from below, on the horizontally stratified ionosphere. Since the wave normal is initially aligned along the z-axis, and since $n = n(z)$, we expect all field components to be functions of z only, so that

$$\frac{\partial}{\partial x} \equiv \frac{\partial}{\partial y} \equiv 0. \tag{4.165}$$

In this situation, Eqs. (4.164) reduce to $E_z = c B_z = 0$, with

$$-\frac{\partial E_y}{\partial z} = \mathrm{i}\,k\,cB_x, \tag{4.166a}$$

$$\frac{\partial\, cB_x}{\partial z} = -\mathrm{i}\,k\,n^2\,E_y, \tag{4.166b}$$

and

$$\frac{\partial E_x}{\partial z} = \mathrm{i}\,k\,cB_y, \tag{4.167a}$$

$$-\frac{\partial\, cB_y}{\partial z} = -\mathrm{i}\,k\,n^2\,E_x. \tag{4.167b}$$

Note that Eqs. (4.166) and (4.167) are isomorphic and completely independent of one another. It follows that, without loss of generality, we can assume that the wave is linearly polarized with its electric vector parallel to the y-axis. This means that we are only going to consider the solution of Eqs. (4.166). The solution of Eqs. (4.167) is of exactly the same form, except that it describes a linear polarized wave with its electric vector parallel to the x-axis.

Equations (4.166) can be combined to give

$$\frac{d^2 E_y}{dz^2} + k^2 n^2\, E_y = 0. \tag{4.168}$$

Since E_y is a function of z only, we now use the total derivative sign d/dz instead of the partial derivative sign $\partial/\partial z$. The solution of the above equation for the case of a uniform medium, where n is constant, is straightforward:

$$E_y = A\,\mathrm{e}^{\,\mathrm{i}\,\phi(z)}, \tag{4.169}$$

where A is a constant, and

$$\phi = \pm k\,n\,z. \tag{4.170}$$

Note that the $e^{-i\omega t}$ time dependence of all wave quantities is taken as read during this investigation. The solution (4.169) represents a *wave* of constant amplitude A and phase $\phi(z)$. According to Eq. (4.170), there are, in fact, two independent waves which can propagate through the medium in question. The upper sign corresponds to a wave which propagates vertically upwards, and the lower sign corresponds to a wave which propagates vertically downwards. Both waves propagate with the constant phase velocity c/n.

In general, if $n = n(z)$ the solution of Eq. (4.168) does not remotely resemble the wave-like solution (4.169). However, in the limit in which $n(z)$ is a "slowly varying" function of z (exactly how slowly varying is something which we shall establish later), we expect to recover wave-like solutions. Let us suppose that $n(z)$ is indeed a "slowly varying" function, and let us try substituting the wave solution (4.169) into Eq. (4.168). We obtain

$$\left(\frac{d\phi}{dz}\right)^2 = k^2 n^2 + i\,\frac{d^2\phi}{dz^2}. \tag{4.171}$$

This is a non-linear differential equation which, in general, is very difficult to solve. However, we note that if n is a constant then $d^2\phi/dz^2 = 0$. It is, therefore, reasonable to suppose that if $n(z)$ is a "slowly varying" function then the last term on the right-hand side of the above equation can be regarded as being small. Thus, to a first approximation Eq. (4.171) yields

$$\frac{d\phi}{dz} \simeq \pm k\,n, \tag{4.172}$$

and

$$\frac{d^2\phi}{dz^2} \simeq \pm k\,\frac{dn}{dz}. \tag{4.173}$$

It is clear from a comparison of Eqs. (4.171) and (4.173) that $n(z)$ can be regarded as a "slowly varying" function of z as long as its variation length-scale is far longer than the wavelength of the wave. In other words, provided that $(dn/dz)/(k\,n^2) \ll 1$.

The second approximation to the solution is obtained by substituting Eq. (4.173) into the right-hand side of Eq. (4.171):

$$\frac{d\phi}{dz} \simeq \pm\left(k^2 n^2 \pm \mathrm{i}\,k\,\frac{dn}{dz}\right)^{1/2}. \tag{4.174}$$

This gives

$$\frac{d\phi}{dz} \simeq \pm k\,n\left(1 \pm \frac{\mathrm{i}}{k\,n^2}\,\frac{dn}{dz}\right)^{1/2} \simeq \pm k\,n + \frac{\mathrm{i}}{2n}\,\frac{dn}{dz}, \tag{4.175}$$

where use has been made of the binomial expansion. The above expression can be integrated to give

$$\phi \sim \pm k \int^z n\,dz + \mathrm{i}\,\log(n^{1/2}). \tag{4.176}$$

Substitution of Eq. (4.176) into Eq. (4.169) yields the final result

$$E_y \simeq A\,n^{-1/2}\,\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right). \tag{4.177}$$

It follows from Eq. (4.166a) that

$$cB_x \simeq \mp A\,n^{1/2}\,\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right) - \frac{\mathrm{i}\,A}{2k\,n^{3/2}}\,\frac{dn}{dz}\,\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right). \tag{4.178}$$

Note that the second term is small compared to the first, and can usually be neglected.

Let us test to what extent the expression (4.177) is a good solution of Eq. (4.168) by substituting this expression into the left-hand side of the equation. The result is

$$\frac{A}{n^{1/2}}\left\{\frac{3}{4}\left(\frac{1}{n}\,\frac{dn}{dz}\right)^2 - \frac{1}{2n}\,\frac{d^2n}{dz^2}\right\}\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right). \tag{4.179}$$

This must be small compared with either term on the left-hand side of Eq. (4.168). Hence, the condition for Eq. (4.177) to be a good solution of Eq. (4.168) becomes

$$\frac{1}{k^2}\left|\frac{3}{4}\left(\frac{1}{n^2}\,\frac{dn}{dz}\right)^2 - \frac{1}{2n^3}\,\frac{d^2n}{dz^2}\right| \ll 1. \tag{4.180}$$

The solutions

$$E_y \simeq A\,n^{-1/2}\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right), \tag{4.181a}$$

$$cB_x \simeq \mp A\,n^{1/2}\exp\left(\pm\mathrm{i}\,k\int^z n\,dz\right), \tag{4.181b}$$

to the non-uniform wave equations (4.166) are most commonly called the *W.K.B. solutions*, in honor of G. Wentzel, H.A. Kramers, and L. Brillouin, who are credited with independently discovering these solutions (in a quantum mechanical context) in 1926. Actually, H. Jeffries wrote a paper on these solutions (in a wave propagation context) in 1923. Hence, some people call these the W.K.B.J. solutions (or even the J.W.K.B. solutions). In fact, these solutions were first discussed by Liouville and Green in 1837, and again by Rayleigh in 1912. We shall refer to Eqs. (4.181) as the W.K.B. solutions, since this is what they are most commonly called. However, it should be understand that, in doing so, we are not making any statement as to the credit due to various scientists in discovering these solutions. After all, this is not a history of science course!

Recall, that when a propagating wave is normally incident on an *interface*, where the refractive index suddenly changes (for instance, when a light wave propagating in the air is normally incident on a glass slab), there is generally significant reflection of the wave. However, according to the W.K.B. solutions (4.181), when a propagating wave is normally incident on a medium in which the refractive index changes *slowly* along the direction of propagation of the wave, then the wave is not reflected at all. This is true even if the refractive index varies very substantially along the path of propagation of the wave, as long as it varies *slowly*. The W.K.B. solutions imply that as the wave propagates through the medium its wavelength gradually changes. In fact, the wavelength at position z is approximately $\lambda(z) = 2\pi/k\,n(z)$. Equations (4.181) also imply that the amplitude of the wave gradually changes as it propagates. In fact, the amplitude of the electric field component is inversely proportional to $n^{1/2}$, whereas the amplitude of the magnetic field component is directly proportional to $n^{1/2}$. Note, however, that the energy flux in the z-direction, given by the the Poynting vector $-(E_y {B_x}^* + {E_y}^* B_x)/(4\mu_0)$, remains constant (assuming that n is predominately real).

Of course, the W.K.B. solutions (4.181) are only *approximations.* In reality, a wave propagating into a medium in which the refractive index is a slowly varying function of position is subject to a small amount of reflection. However, it is easily demonstrated that the ratio of the reflected amplitude to the incident amplitude is of order $(dn/dz)/(k\,n^2)$. Thus, as long as the refractive index varies on a much longer length-scale than the wavelength of the radiation, the reflected wave is negligibly small. This conclusion remains valid as long as the inequality (4.180) is satisfied. There are two main reasons why this inequality might fail to be satisfied. First of all, if there is a localized region in the dielectric medium in which the refractive index suddenly changes (*i.e.*, if there is an interface), then (4.180) is likely to break down in this region, allowing strong reflection of the incident wave. Secondly, the inequality obviously breaks down in the vicinity of a point where $n = 0$. We would, therefore, expect strong reflection of the incident wave from such a point.

4.17 The reflection coefficient

Consider an ionosphere in which the refractive index is a slowly varying function of height z above the surface of the Earth. Let n^2 be positive for $z < z_0$, and negative for $z > z_0$. Suppose that an upgoing radio wave of amplitude E_0 is generated at ground level ($z = 0$). The complex amplitude of the wave in the region $0 < z < z_0$ is given by the upgoing W.K.B. solution

$$E_y = E_0\, n^{-1/2} \exp\left(\mathrm{i}\, k \int_0^z n\, dz\right), \tag{4.182a}$$

$$cB_x = -E_0\, n^{1/2} \exp\left(\mathrm{i}\, k \int_0^z n\, dz\right). \tag{4.182b}$$

The upgoing energy flux is given by $-(E_y {B_x}^* + {E_y}^* B_x)/(4\mu_0) = (\epsilon_0/\mu_0)^{1/2}\, |E_0|^2/2$. In the region $z > z_0$ the W.K.B. solutions take the form

$$E_y = A\, \mathrm{e}^{\mathrm{i}\,\pi/4}\, |n|^{-1/2} \exp\left(\pm k \int^z |n|\, dz\right), \tag{4.183a}$$

$$cB_x = \pm A\, \mathrm{e}^{-\mathrm{i}\pi/4}\, |n|^{1/2} \exp\left(\pm k \int^z |n|\, dz\right), \tag{4.183b}$$

where A is a constant. These solutions correspond to exponentially growing and decaying waves. Note that the magnetic components of the waves are in *phase quadrature* with the electric components. This implies that the Poynting fluxes of the waves are zero; *i.e.*, the waves do not transmit energy. Thus, there is a non-zero incident energy flux in the region $z < z_0$, and zero energy flux in the region $z > z_0$. Clearly, the incident wave is either absorbed or reflected in the vicinity of the plane $z = z_0$ (where $n = 0$). In fact, as we shall prove later on, the wave is *reflected*. The complex amplitude of the reflected wave in the region $0 < z < z_0$ is given by the downgoing W.K.B. solution

$$E_y = E_0\, R\, n^{-1/2} \exp\left(-\mathrm{i}\, k \int_0^z n\, dz\right), \tag{4.184a}$$

$$cB_x = E_0\, R\, n^{1/2} \exp\left(-\mathrm{i}\, k \int_0^z n\, dz\right), \tag{4.184b}$$

where R is the coefficient of reflection. Suppose, for the sake of argument, that the plane $z = z_0$ acts like a perfect conductor, so that $E_y(z_0) = 0$. It follows that

$$R = -\exp\left(2\mathrm{i}\, k \int_0^{z_0} n\, dz\right). \tag{4.185}$$

In fact, as we shall prove later on, the correct answer is

$$R = -\mathrm{i}\, \exp\left(2\mathrm{i}\, k \int_0^{z_0} n\, dz\right). \tag{4.186}$$

Thus, there is only a $-\pi/2$ phase shift at the reflection point, instead of the $-\pi$ phase shift which would be obtained if the plane $z = z_0$ acted like a perfect conductor.

4.18 Extension to oblique incidence

We have discussed the W.K.B. solutions for radio waves propagating vertically through an ionosphere whose refractive index varies slowly. Let us now generalize these solutions to allow for radio waves which propagate at an angle to the vertical axis.

The refractive index of the ionosphere varies continuously with height z. However, let us, for the sake of clarity, imagine that the ionosphere is replaced by a number of thin discrete strata in which the medium is homogeneous. By making these strata sufficiently thin and numerous we can approximate as closely as is desired to the real ionosphere. Suppose that a plane wave is incident on the ionosphere, from below, and suppose that the wave normal lies in the x-z plane and makes an angle θ_I with the vertical axis. At the lower boundary of the first stratum the wave is partially reflected and partially transmitted. The transmitted wave is partially reflected and partially transmitted at the second boundary between the strata, and so on. However, in the limit of many strata, where the difference in refractive indices between neighbouring strata is very small, the amount of reflection at the boundaries becomes negligible. In the nth stratum, let n_n be the refractive index, and let θ_n be the angle between the wave normal and the vertical axis. According to Snell's law,

$$n_{n-1}\,\sin\theta_{n-1} = n_n\,\sin\theta_n. \tag{4.187}$$

Below the ionosphere $n = 1$, and so

$$n_n\,\sin\theta_n = \sin\theta_I. \tag{4.188}$$

For a wave in the nth stratum, any field quantity depends on z and x through factors

$$A\exp\left[\mathrm{i}\,k\,n_n(\pm z\cos\theta_n + x\sin\theta_n)\right], \tag{4.189}$$

where A is a constant. The $\pm$ signs denote upgoing and downgoing waves, respectively. When the operator $\partial/\partial x$ acts on the above expression, it is equivalent to multiplication by $\mathrm{i}\,k\,n_n\,\sin\theta_n = \mathrm{i}\,k\,\sin\theta_I$, which is independent of x and z. It is convenient to use the notation $S = \sin\theta_I$. Hence, we may write symbolically

$$\frac{\partial}{\partial x} \equiv \mathrm{i}\,k\,S, \tag{4.190a}$$

$$\frac{\partial}{\partial y} \equiv 0. \tag{4.190b}$$

This result is true no matter how thin the strata are, so it must also hold for the real ionosphere. Note that, according to Snell's law, if the wave normal starts off in the x-z plane then it will remain in this plane as it propagates through the ionosphere.

Equations (4.164) and (4.190) can be combined to give

$$-\frac{\partial E_y}{\partial z} = \mathrm{i}\,k\,cB_x, \tag{4.191a}$$

$$\mathrm{i}\,kS\,E_y = \mathrm{i}\,k\,cB_z, \tag{4.191b}$$

$$\frac{\partial\,cB_x}{\partial z} - \mathrm{i}\,kS\,cB_z = -\mathrm{i}\,k\,n^2\,E_y, \tag{4.191c}$$

and

$$\frac{\partial E_x}{\partial z} - \mathrm{i}\,kS\,E_z = \mathrm{i}\,k\,cB_y, \tag{4.192a}$$

$$-\frac{\partial\,cB_y}{\partial z} = -\mathrm{i}\,k\,n^2\,E_x, \tag{4.192b}$$

$$\mathrm{i}\,kS\,cB_y = -\mathrm{i}\,k\,n^2\,E_z. \tag{4.192c}$$

As before, Maxwell's equations can be split into two independent groups, corresponding to two independent polarizations of radio waves propagating through the ionosphere. For the first set of equations, the electric field is always parallel to the y-axis. The corresponding waves are, therefore, said to be *horizontally polarized*. For the second set of equations, the electric field always lies in the x-z plane. The corresponding waves are, therefore, said to be *vertically polarized* (*n.b.*, the term "vertically polarized" does not necessarily imply that the electric field is parallel to the vertical axis). Note that the equations governing horizontally polarized waves are *not* isomorphic to those governing vertically polarized waves, so both types of waves must be dealt with separately.

For the case of horizontally polarized waves, Eqs. (4.191b) and (4.191c) yield

$$\frac{\partial\,cB_x}{\partial z} = -\mathrm{i}\,k\,q^2\,E_y, \tag{4.193}$$

where

$$q^2 = n^2 - S^2. \tag{4.194}$$

The above equation can be combined with Eq. (4.191a) to give

$$\frac{\partial^2 E_y}{\partial z^2} + k^2 q^2 \, E_y = 0. \tag{4.195}$$

Equations (4.193) and (4.195) have exactly the same form as Eqs. (4.166b) and (4.168), except that n^2 is replaced by q^2, so the results of Section 4.16 can be immediately employed to find the W.K.B. solutions, which take the form

$$E_y = A\, q^{-1/2} \exp\left(\pm\mathrm{i}\, k \int^z q\, dz\right), \tag{4.196a}$$

$$cB_x = \mp A\, q^{1/2} \exp\left(\pm\mathrm{i}\, k \int^z q\, dz\right), \tag{4.196b}$$

where A is a constant. Of course, both expressions should also contain a multiplicative factor $\mathrm{e}^{\,\mathrm{i}\,(kS\,x-\omega t)}$, but this is usually omitted for the sake of clarity. By analogy with Eq. (4.180), the W.K.B. solutions are valid as long as

$$\frac{1}{k^2}\left|\frac{3}{4}\left(\frac{1}{q^2}\frac{dq}{dz}\right)^2 - \frac{1}{2q^3}\frac{d^2q}{dz^2}\right| \ll 1. \tag{4.197}$$

This inequality clearly fails in the vicinity of $q = 0$, no matter how slowly q varies with z. Hence, $q = 0$, or $n^2 = S^2$, specifies the height at which reflection takes place. By analogy with Eq. (4.186), the reflection coefficient at ground level ($z = 0$) is given by

$$R = -\mathrm{i}\,\exp\left(2\mathrm{i}\, k \int_0^{z_0} q\, dz\right), \tag{4.198}$$

where z_0 is the height at which $q = 0$.

For the case of vertical polarization, Eqs. (4.192a) and (4.192c) yield

$$\frac{\partial E_x}{\partial z} = \mathrm{i}\, k\, \frac{q^2}{n^2}\, cB_y. \tag{4.199}$$

This equation can be combined with Eq. (4.192b) to give

$$\frac{\partial^2 B_y}{\partial z^2} - \frac{1}{n^2}\frac{d(n^2)}{dz}\frac{\partial B_y}{\partial z} + k^2 q^2 \, B_y = 0. \tag{4.200}$$

Clearly, the differential equation which governs the propagation of vertically polarized waves is considerably more complicated than the corresponding equation for horizontally polarized waves.

The W.K.B. solution for vertically polarized waves is obtained by substituting the wave-like solution

$$cB_y = A\,\mathrm{e}^{\,\mathrm{i}\,\phi(z)}, \tag{4.201}$$

where A is a constant and $\phi(z)$ is the generalized phase, into Eq. (4.200). The differential equation thus obtained for the phase is

$$\mathrm{i}\,\frac{d^2\phi}{dz^2} - \left(\frac{d\phi}{dz}\right)^2 - \frac{\mathrm{i}}{n^2}\frac{d(n^2)}{dz}\frac{d\phi}{dz} + k^2 q^2\,\phi = 0. \tag{4.202}$$

Since the medium is slowly varying, the first and third term in the above equation are small, and so to a first approximation

$$\frac{d\phi}{dz} = \pm k\,q, \tag{4.203a}$$

$$\frac{d^2\phi}{dz^2} = \pm k\,\frac{dq}{dz}. \tag{4.203b}$$

These expressions can be inserted into the first and third terms of Eq. (4.202) to give the second approximation

$$\frac{d\phi}{dz} = \pm\left[k^2 q^2 \pm \mathrm{i}\,k\left(\frac{dq}{dz} - \frac{2q}{n}\frac{dn}{dz}\right)\right]^{1/2}. \tag{4.204}$$

The final two terms on the right-hand side of the above equation are small, so expanding the right-hand side using the binomial theorem yields

$$\frac{d\phi}{dz} = \pm kq + \frac{\mathrm{i}}{2q}\frac{dq}{dz} - \frac{\mathrm{i}}{n}\frac{dn}{dz}. \tag{4.205}$$

This expression can be integrated, and the result inserted into Eq. (4.201), to give the W.K.B. solution

$$cB_y = A\,n\,q^{-1/2}\,\exp\left(\pm\mathrm{i}\,k\int^z q\,dz\right). \tag{4.206}$$

The corresponding W.K.B. solution for E_x is obtained from Eq. (4.199):

$$E_x = \pm A\, n^{-1}\, q^{1/2}\, \exp\left(\pm \mathrm{i}\, k \int^z q\, dz\right). \tag{4.207}$$

Here, any terms involving derivatives of n and q have been neglected.

Substituting Eq. (4.206) into the differential equation (4.200), and demanding that the result be small compared to the original terms in the differential equation, yields the following condition for the validity of the above W.K.B. solutions:

$$\frac{1}{k^2}\left|\frac{3}{4}\left(\frac{1}{q^2}\frac{dq}{dz}\right)^2 - \frac{1}{2q^3}\frac{d^2q}{dz^2} + \frac{1}{q^2}\left[\frac{1}{n}\frac{d^2n}{dz^2} - 2\left(\frac{1}{n}\frac{dn}{dz}\right)^2\right]\right| \ll 1. \tag{4.208}$$

This criterion fails close to $q = 0$, no matter how slowly n and q vary with z. Hence, $q = 0$ gives the height at which reflection takes place. The condition also fails close to $n = 0$, which does not correspond to the reflection level. If, as is usually the case, the electron density in the ionosphere increases monotonically with height, then the level where $n = 0$ lies above the reflection level, where $q = 0$. If the two levels are well separated then the reflection process is unaffected by the failure of the above inequality at the level $n = 0$, and the reflection coefficient is given by Eq. (4.198), just as for the case of horizontal polarization. If, however, the level $n = 0$ lies close to the level $q = 0$ then the reflection coefficient may be affected, and a more accurate treatment of the differential equation (4.200) is required in order to obtain the true value of the reflection coefficient.

4.19 Pulse propagation in the ionosphere

Suppose that we possess a generator of radio waves which sends radio pulses vertically upwards into the ionosphere. For the sake of argument, we shall assume that these pulses are linearly polarized such that the electric field vector lies parallel to the y-axis. The pulse structure can be represented as

$$E_y(t) = \int_{-\infty}^{\infty} F(\omega)\, \mathrm{e}^{-\mathrm{i}\,\omega\, t}\, d\omega, \tag{4.209}$$

where $E_y(t)$ is the electric field produced by the generator (*i.e.*, the field at $z = 0$). Suppose that the pulse is a signal of roughly constant (angular) frequency ω_0, which lasts a time T, where T is long compared to $1/\omega_0$. It follows that $F(\omega)$ possesses narrow maxima around $\omega = \pm\omega_0$. In other words, only those frequencies which lie very close to the central frequency ω_0 play a significant role in the propagation of the pulse.

Each component frequency of the pulse yields a wave which travels independently up into the ionosphere, in a manner specified by the appropriate W.K.B. solution [see Eqs. (4.181)]. Thus, if Eq. (4.209) specifies the signal at ground level ($z = 0$), then the signal at height z is given by

$$E_y(z,t) = \int_{-\infty}^{\infty} \frac{F(\omega)}{n^{1/2}(\omega,z)}\, \mathrm{e}^{\,\mathrm{i}\,\phi(\omega,z,t)}\, d\omega, \tag{4.210}$$

where

$$\phi(\omega,z,t) = \frac{\omega}{c}\int_0^z n(\omega,z)\, dz - \omega\, t. \tag{4.211}$$

Here, we have used $k = \omega/c$.

Equation (4.210) can be regarded as a contour integral in ω-space. The quantity $F/n^{1/2}$ is a relatively slowly varying function of ω, whereas the phase ϕ is a large and rapidly varying function of ω. As described in Section 4.11, the rapid oscillations of $\exp(\,\mathrm{i}\,\phi)$ over most of the path of integration ensure that the integrand averages almost to zero. However, this cancellation argument does not apply to those points on the path of integration where the phase is *stationary*; *i.e.*, those points where $\partial\phi/\partial\omega = 0$. It follows that the left-hand side of Eq. (4.210) averages to a very small value, expect for those special values of z and t at which one of the points of stationary phase in ω-space coincides with one of the peaks of $F(\omega)$. The locus of these special values of z and t can obviously be regarded as the equation of motion of the pulse as it propagates through the ionosphere. Thus, the equation of motion is specified by

$$\left(\frac{\partial\phi}{\partial\omega}\right)_{\omega=\omega_0} = 0, \tag{4.212}$$

which yields

$$t = \frac{1}{c}\int_0^z \left[\frac{\partial(\omega\, n)}{\partial\omega}\right]_{\omega=\omega_0} dz. \tag{4.213}$$

Suppose that the z-velocity of a pulse of central frequency ω_0 at height z is given by $u_z(\omega_0, z)$. The differential equation of motion of the pulse is then $dt = dz/u_z$. This can be integrated, using the boundary condition $z = 0$ at $t = 0$, to give the full equation of motion:

$$t = \int_0^z \frac{dz}{u_z}. \tag{4.214}$$

A comparison of Eqs. (4.213) and (4.214) yields

$$u_z(\omega_0, z) = c \left/ \left\{\frac{\partial[\omega\, n(\omega, z)]}{\partial\omega}\right\}_{\omega=\omega_0}\right. . \tag{4.215}$$

The velocity u_z is usually called the *group velocity*. It is easily demonstrated that the above expression for the group velocity is entirely consistent with that given previously [see Eq. (4.135)].

The dispersion relation (4.164) yields

$$n(\omega, z) = \left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{1/2}, \tag{4.216}$$

in the limit where electron collisions are negligible. The phase velocity of radio waves of frequency ω propagating vertically through the ionosphere is given by

$$v_z(\omega, z) = \frac{c}{n(\omega, z)} = c\left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{-1/2}. \tag{4.217}$$

According to Eqs. (4.215) and (4.216), the corresponding group velocity is

$$u_z(\omega, z) = c\left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{1/2}. \tag{4.218}$$

It follows that

$$v_z\,u_z = c^2. \tag{4.219}$$

Note that as the reflection point $z = z_0$ [defined as the solution of $\omega = \omega_p(z_0)$] is approached from below, the phase velocity tends to infinity, whereas the group velocity tends to zero.

Let τ be the time taken for the pulse to travel from the ground to the reflection level, and back to the ground again. The product $c\,\tau/2$ is termed the *equivalent height of reflection*, and is denoted $h(\omega)$, since it is a function of the pulse frequency, ω. The equivalent height is the height to which the pulse would have to go if it always traveled with the velocity c. Since we know that a pulse of dominant frequency ω propagates at height z with the z-velocity $u_z(\omega, z)$ (this is true for both upgoing and downgoing pulses), and also that the pulse is reflected at the height $z_0(\omega)$, where $\omega = \omega_p(z_0)$, it follows that

$$\tau = 2\int_0^{z_0(\omega)} \frac{dz}{u_z(\omega, z)}. \tag{4.220}$$

Hence,

$$h(\omega) = \int_0^{z_0(\omega)} \frac{c}{u_z(\omega, z)}\, dz. \tag{4.221}$$

Note that the equivalent height of reflection, $h(\omega)$, is always *greater* than the actual height of reflection, $z_0(\omega)$, since the group velocity u_z is always less than the velocity of light. The above equation can be combined with Eq. (4.218) to give

$$h(\omega) = \int_0^{z_0(\omega)} \left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{-1/2} dz. \tag{4.222}$$

Note that the integrand diverges as the reflection point is approached, but the integral remains finite.

4.20 Determining the ionospheric electron density profile

We can measure the equivalent height of the ionosphere in a fairly straightforward manner, by timing how long it takes a radio pulse fired vertically upwards to return to ground level again. We can, therefore, determine the function $h(\omega)$ experimentally by performing this procedure many times over, using pulses of different central frequencies. But, is it possible to use this information to determine the number density of free electrons in the ionosphere as a function of height? In mathematical terms, the problem is as follows. Does a knowledge of the function

$$h(\omega) = \int_0^{z_0(\omega)} \frac{\omega}{[\omega^2 - {\omega_p}^2(z)]^{1/2}}\,dz, \tag{4.223}$$

where ${\omega_p}^2(z_0) = \omega^2$, necessarily imply a knowledge of the function ${\omega_p}^2(z)$? Note, of course, that ${\omega_p}^2(z) \propto N(z)$.

Let $\omega^2 = v$ and ${\omega_p}^2(z) = u(z)$. Equation (4.223) then becomes

$$v^{-1/2}\,h(v^{1/2}) = \int_0^{z_0(v^{1/2})} \frac{dz}{[v - u(z)]^{1/2}}, \tag{4.224}$$

where $u(z_0) = v$, and $u(z) < v$ for $0 < z < z_0$. Let us multiply both sides of the above equation by $(w - v)^{-1/2}/\pi$ and integrate from $v = 0$ to w. We obtain

$$\frac{1}{\pi}\int_0^w v^{-1/2}\,(w - v)^{-1/2}\,h(v^{1/2})\,dv = \frac{1}{\pi}\int_0^w \left[\int_0^{z_0(v^{1/2})} \frac{dz}{(w - v)^{1/2}\,(v - u)^{1/2}}\right] dv. \tag{4.225}$$

Consider the double integral on the right-hand side. The region of v-z space over which this integral is performed is sketched in Fig. 15. It can be seen that, as long as $z_0(v^{1/2})$ is a *monotonically increasing* function of z, we can swap the order of integration to give

$$\frac{1}{\pi}\int_0^{z_0(w^{1/2})} \left[\int_{u(z)}^w \frac{dv}{(w - v)^{1/2}\,(v - u)^{1/2}}\right] dz. \tag{4.226}$$

Here, we have used the fact that the curve $z = z_0(v^{1/2})$ is identical with the curve $v = u(z)$. Note that if $z_0(v^{1/2})$ is *not* a monotonically increasing function of v then we can still swap the order of integration, but the limits of integration are, in general, far more complicated than those indicated above. The integral over v in the above expression can be evaluated using standard methods (by making the substitution $v = w\,\cos^2\theta + u\,\sin^2\theta$): the result is simply π. Thus, the expression (4.226) reduces to $z_0(w^{1/2})$. It follows from Eq. (4.225) that

$$z_0(w^{1/2}) = \frac{1}{\pi}\int_0^w v^{-1/2}\,(w-v)^{-1/2}\,h(v^{1/2})\,dv. \tag{4.227}$$

Making the substitutions $v = w\,\sin^2\alpha$ and $w^{1/2} = \omega$, we obtain

$$z_0(\omega) = \frac{2}{\pi}\int_0^{\pi/2} h(\omega\,\sin\alpha)\,d\alpha. \tag{4.228}$$

By definition, $\omega = \omega_p$ at the reflection level $z = z_0$. Hence, the above equation reduces to

$$z(\omega_p) = \frac{2}{\pi}\int_0^{\pi/2} h(\omega_p\,\sin\alpha)\,d\alpha. \tag{4.229}$$

Thus, we can obtain z as a function of ω_p (and, hence, ω_p as a function of z) simply by taking the appropriate integral of the experimentally determined function $h(\omega)$. Since $\omega_p(z) \propto \sqrt{N(z)}$, this means that we can determine the electron number density profile in the ionosphere provided we know the variation of the equivalent height of the ionosphere with pulse frequency. The constraint that $z_0(\omega)$ must be a monotonically increasing function of ω translates to the constraint that $N(z)$ must be a monotonically increasing function of z. Note that we can still determine $N(z)$ from $h(\omega)$ for the case where the former function is non-monotonic, it is just a far more complicated procedure than that outlined above. Incidentally, the technique by which we have inverted Eq. (4.222), which specifies $h(\omega)$ as some integral over $\omega_p(z)$, to give $\omega_p(z)$ as some integral over $h(\omega)$ is known as *Abel inversion.*

4.21 Ray tracing in the ionosphere

Suppose that we possess a radio antenna which is capable of launching radio waves of constant frequency ω into the ionosphere at an angle to the vertical. Let us consider the paths traced out by these waves in the x-z plane. For the sake of simplicity, we shall assume that the waves are horizontally polarized, so that the

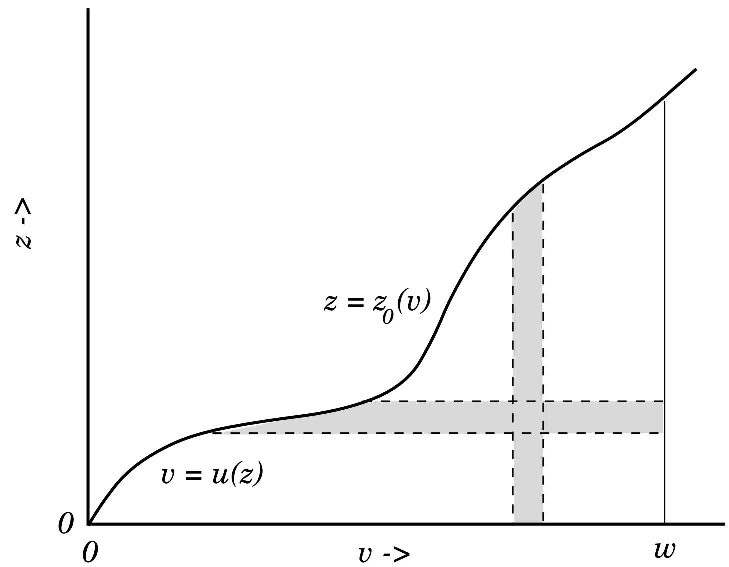


Figure 15: A sketch of the region of v-z space over which the integral on the right-hand side of Eq. (4.223) is evaluated

electric field vector always lies parallel to the y-axis. The signal emitted by the antenna (located at $z = 0$) can be represented as

$$E_y(x) = \int_0^1 F(S)\, \mathrm{e}^{\,\mathrm{i}\, k\, S\, x}\, dS, \tag{4.230}$$

where $k = \omega/c$. Here, the $\mathrm{e}^{-\mathrm{i}\,\omega\, t}$ time dependence of the signal is neglected for the sake of clarity. Suppose that the signal emitted by the antenna is mostly concentrated in a direction making an angle θ_I with the vertical. It follows that $F(S)$ possesses a narrow maximum around $S = S_0$, where $S_0 = \sin\theta_I$.

If Eq. (4.230) represents the signal at ground level, then the signal at height z in the ionosphere is easily obtained by making use of the W.K.B. solutions for horizontally polarized waves described in Section 4.18. We obtain

$$E_y(x, z) = \int_0^1 \frac{F(S)}{q^{1/2}(z, S)}\, \mathrm{e}^{\,\mathrm{i}\,\phi(x,z,S)}\, dS, \tag{4.231}$$

where

$$\phi(x, z, S) = k \int_0^z q(z, S)\, dz + k\, S\, x. \tag{4.232}$$

Equation (4.231) is basically a line integral in S-space. The quantity $F/q^{1/2}$ is a relatively slowly varying function of S, whereas the phase ϕ is a large and rapidly varying function of S. As described in Section 4.11, the rapid oscillations of $\exp(\mathrm{i}\,\phi)$ over most of the path of integration ensure that the integrand averages almost to zero. In fact, only those points on the path of integration where the phase is stationary (*i.e.*, where $\partial\phi/\partial S = 0$) make a significant contribution to the integral. It follows that the left-hand side of Eq. (4.231) averages to a very small value, except for those special values of x and z at which one of the points of stationary phase in S-space coincides with the peak of $F(S)$. The locus of these special values of x and z can clearly be regarded as the trajectory of the radio signal emitted by the antenna as it passes through the ionosphere. Thus, the signal trajectory is specified by

$$\left(\frac{\partial\phi}{\partial S}\right)_{S=S_0} = 0, \tag{4.233}$$

which yields

$$x = -\int_0^z \left(\frac{\partial q}{\partial S}\right)_{S=S_0} dz. \tag{4.234}$$

We can think of this equation as tracing the path of a *ray* of radio frequency radiation, emitted by the antenna at an angle θ_I to the vertical (where $S_0 = \sin\theta_I$), as it propagates through the ionosphere.

Now

$$q^2 = n^2 - S^2, \tag{4.235}$$

so the ray tracing equation becomes

$$x = S\int_0^z \frac{dz}{\sqrt{n^2(z) - S^2}}, \tag{4.236}$$

where S is the sine of the initial (*i.e.*, at the antenna) angle of incidence of the ray with respect to the vertical axis. Of course, Eq. (4.236) only holds for *upgoing* rays. For *downgoing* rays, a simple variant of the previous analysis using the downgoing W.K.B. solutions yields

$$x = S\int_0^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}} + S\int_z^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}}, \tag{4.237}$$

where $n(z_0) = S$. Thus, the ray ascends into the ionosphere after being launched from the antenna, reaches a maximum height z_0 above the surface of the Earth, and then starts to descend. The ray eventually intersects the Earth's surface again a horizontal distance

$$x_0 = 2S \int_0^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}} \tag{4.238}$$

away from the antenna.

The angle ξ which the ray makes with the vertical is given by $\tan\xi = dx/dz$. It follows from Eqs. (4.236) and (4.237) that

$$\tan\xi = \pm \frac{S}{\sqrt{n^2(z) - S^2}} \tag{4.239}$$

where the upper and lower signs correspond to the upgoing and downgoing parts of the ray trajectory, respectively. Note that $\xi = \pi/2$ at the reflection point, where $n = S$. Thus, the ray is horizontal at the reflection point.

Let us investigate the reflection process in more detail. In particular, we wish to prove that radio waves are reflected at the $q = 0$ surface, rather than being absorbed. We would also like to understand the origin of the $-\pi/2$ phase shift of radio waves at reflection which is evident in Eq. (4.198). In order to achieve these goals, we shall need to review the mathematics of asymptotic series.

4.22 Asymptotic series: A mathematical aside

It is often convenient to expand a function of the complex variable $f(z)$ in inverse powers of z:

$$f(z) = \phi(z) \left[A_0 + \frac{A_1}{z} + \frac{A_2}{z^2} + \cdots \right], \tag{4.240}$$

where $\phi(z)$ is a function whose behaviour for large values of z is known. If $f(z)/\phi(z)$ is singular as $|z| \to \infty$ then the above series diverges. Nevertheless, under certain circumstances, the series may still be useful.

The circumstance needed to make this possible is that the difference between $f(z)/\phi(z)$ and the first $n+1$ terms of the series be of order $1/z^{n+1}$, so that for sufficiently large z this difference becomes vanishingly small. More precisely, the series is said to represent $f(z)/\phi(z)$ *asymptotically*, that is

$$f(z) \simeq \phi(z) \sum_{p=0}^{\infty} \frac{A_p}{z^p}, \tag{4.241}$$

provided that

$$\lim_{|z|\to\infty} \left\{ z^n \left[\frac{f(z)}{\phi(z)} - \sum_{p=0}^{n} \frac{A_p}{z^p} \right] \right\} \to 0. \tag{4.242}$$

In other words, for a given value of n, the first $n+1$ terms of the series may be made as close as desired to the ratio $f(z)/\phi(z)$ by making z sufficiently large. For each value of z and n there is an error of order $1/z^{n+1}$. Since the series actually diverges, there is an optimum number of terms in the series used to represent $f(z)/\phi(z)$ for a given value of z. Associated with this is an unavoidable error. As z increases, the optimal number of terms increases and the error decreases.

Consider a simple example. The exponential integral is defined

$$\mathrm{E}_1(x) = \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t}\, dt. \tag{4.243}$$

The asymptotic series for this function can be generated via a series of partial integrations. For example,

$$\mathrm{E}_1(x) = \frac{\mathrm{e}^{-x}}{x} - \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t^2}\, dt. \tag{4.244}$$

Continuing this procedure yields

$$\begin{aligned} \mathrm{E}_1(x) \quad &= \quad \frac{\mathrm{e}^{-x}}{x} \left[1 - \frac{1}{x} + \frac{2!}{x^2} - \frac{3!}{x^3} + \cdots + \frac{(-1)^n n!}{x^n} \right] \\ &\quad + (-1)^{n+1}(n+1)! \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t^{n+2}}\, dt. \end{aligned} \tag{4.245}$$

The infinite series obtained by taking the limit $n \rightarrow \infty$ diverges, since the Cauchy convergence test yields

$$\lim_{n\rightarrow\infty}\left|\frac{u_{n+1}}{u_n}\right| = \lim_{n\rightarrow\infty}\left[\frac{n}{x}\right] \rightarrow \infty. \tag{4.246}$$

Note that two successive terms in the series become equal in magnitude for $n = x$, indicating that the optimum number of terms for a given x is roughly the integer nearest x. To prove that the series is asymptotic, we need to show that

$$\lim_{x\rightarrow 0} x^{n+1}\,\mathrm{e}^{x}(-1)^{n+1}(n+1)!\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{n+2}}\,dt = 0. \tag{4.247}$$

This immediately follows, since

$$\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{n+2}}\,dt < \frac{1}{x^{n+2}}\int_x^\infty \mathrm{e}^{-t}\,dt = \frac{\mathrm{e}^{-x}}{x^{n+2}}. \tag{4.248}$$

Thus, the error involved in using the first n terms is less than $(n+1)!\,\mathrm{e}^{-x}/x^{n+2}$, which is exactly the next term in the series. We can see that as n increases, this estimate of the error first decreases and then increases without limit. In order to visualize this phenomenon more exactly, let $f(x) = x\exp(x)\,\mathrm{E}(x)$, and let

$$f_n(x) = \sum_{p=0}^{n}\frac{(-1)^p p!}{x^p} \tag{4.249}$$

be the asymptotic series representation of this function which contains $n+1$ terms. Figure 16 shows the relative error in the asymptotic series $|f_n(x) - f(x)|/f(x)$ plotted as a function of the approximate number of terms in the series n for $x = 10$. It can be seen that as n increases the error initially falls, reaches a minimum value at about $n = 10$, and then increases rapidly. Clearly, the optimum number of terms in the asymptotic series used to represent $f(10)$ is about 10.

Asymptotic series are fundamentally different to conventional power law expansions, such as

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots. \tag{4.250}$$

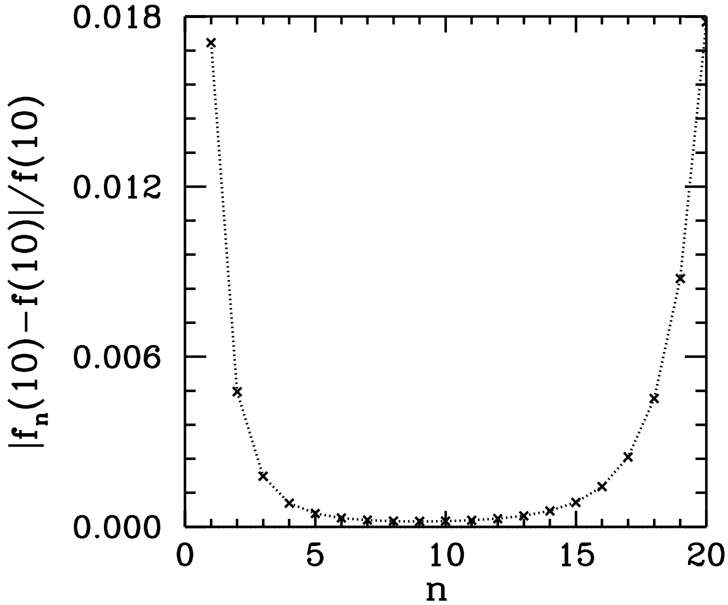


Figure 16: The relative error in a typical asymptotic series plotted as a function of the number of terms in the series

This series representation of $\sin z$ *converges* absolutely for all finite values of z. Thus, at any z the error associated with the series can be made as small as is desired by including a sufficiently large number of terms. In other words, unlike an asymptotic series, there is no intrinsic, or unavoidable, error associated with a convergent series. It follows that a convergent power law series representation of a function is *unique* inside the domain of convergence of the series. On the other hand, an asymptotic series representation of a function is *not unique*. It is perfectly possible to have two different asymptotic series representations of the same function, as long as the difference between the two series is less than the intrinsic error associated with each series. Furthermore, it is often the case that *different* asymptotic series are used to represent the *same* single-valued analytic function in different regions of the complex plane.

For example, consider the asymptotic expansion of the confluent hypergeometric function $F(a, c, z)$. This function is the solution of the differential equation

$$zF'' + (c - z)F' - aF = 0 \tag{4.251}$$

which is analytic at $z = 0$ [in fact, $F(a, c, 0) = 1$]. Here, $'$ denotes d/dz. The asymptotic expansion of $F(a,c,z)$ takes the form:

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{-\mathrm{i}\,\pi a}\,[1+O(1/z)] \qquad (4.252\mathrm{a})$$

for $-\pi < \arg(z) < 0$, and

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{\mathrm{i}\,\pi a}\,[1+O(1/z)] \qquad (4.252\mathrm{b})$$

for $0 < \arg(z) < \pi$, and

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{-\mathrm{i}\,2\pi(a-c)}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{\mathrm{i}\,\pi a}\,[1+O(1/z)] \qquad (4.252\mathrm{c})$$

for $\pi < \arg(z) < 2\pi$, *etc.* It can be seen that the expansion consists of a linear combination of two asymptotic series (only the first term in each series is shown). For $|z| \gg 1$, the first series is exponentially larger than the second whenever $\mathrm{Re}(z) > 0$. We say that the first series is *dominant* in this region, whereas the second series is *subdominant*. Likewise, the first series is exponentially smaller than the second whenever $\mathrm{Re}(z) < 0$. We say that the first series is subdominant and the second series is dominant in this region.

Consider a region in which one or other of the series is dominant. Strictly speaking, it is not mathematically consistent to include the subdominant series in the asymptotic expansion because its contribution is actually less than the intrinsic error associated with the dominant series [this error is $O(1/z)$ times the dominant series, since we are only including the first term in this series]. Thus, at a general point in the complex plane the asymptotic expansion simply consists of the dominant series. However, this is not the case in the immediate vicinity of the lines $\mathrm{Re}(z) = 0$: these are called the *anti-Stokes lines*. When an anti-Stokes line is crossed, a dominant series becomes subdominant and *vice versa*. In the immediate vicinity of an anti-Stokes line neither series is dominant, so it is mathematically consistent to include both series in the asymptotic expansion.

The hypergeometric function $F(a, c, z)$ is a perfectly well behaved, single-valued, analytic function in the complex plane. However, our two asymptotic series are, in general, multi-valued functions in the complex plane [see Eq. (4.252a)]. Can a single-valued function be represented asymptotically by a multi-valued function? The short answer is no. We have to employ different combinations of the two series in different regions of the complex plane in order to ensure that $F(a, c, z)$ remains single-valued. Equations (4.252) show how this is achieved. Basically, the coefficient in front of the subdominant series changes *discontinuously* at certain values of $\arg(z)$. This is perfectly consistent with $F(a, c, z)$ being an analytic function because the subdominant series is "invisible"; *i.e.*, the contribution of the subdominant series to the asymptotic solution falls below the intrinsic error associated with the dominant series, so it does not really matter if the coefficient in front of the former series changes discontinuously. Imagine tracing a large circle, centred on the origin, in the complex plane. Close to an anti-Stokes line, neither series is dominant, so we must include both series in the asymptotic expansion. As we move away from the anti-Stokes line, one series becomes dominant, which means that the other series becomes subdominant and, therefore, drops out of our asymptotic expansion. Eventually, we approach a second anti-Stokes line, and the subdominant series reappears in our asymptotic expansion. However, the coefficient in front of the subdominant series when it reappears is different to that which it had when it disappeared. This new coefficient is carried across the second anti-Stokes line into the region where the subdominant series becomes dominant. In this new region, the dominant series becomes subdominant and disappears from our asymptotic expansion. Eventually, a third anti-Stokes line is approached and the series reappears, but, again, with a different coefficient in front. The jumps in the coefficients of the subdominant series are chosen in such a manner that if we perform a complete circuit in the complex plane then the value of the asymptotic expansion is the same at the beginning and the end points. In other words, the asymptotic expansion is single-valued, despite the fact that it is built up out of two asymptotic series which are not single-valued. The jumps in the coefficient of the subdominant series, which are needed to keep the asymptotic expansion single-valued, are called *Stokes phenomena*, after the celebrated nineteenth century British mathematician Sir George Gabriel Stokes, who first drew attention to this effect.

Where exactly does the jump in the coefficient of the subdominant series occur? All we can really say is "somewhere in the region between two anti-Stokes lines where the series in question is subdominant." The problem is that we only retain the first term in each asymptotic series. Consequently, the intrinsic error in the dominant series is relatively large and we lose track of the subdominant series very quickly after moving away from an anti-Stokes line. However, we could include more terms in each asymptotic series. This would enable us to reduce the intrinsic error in the dominant series and, thereby, expand the region of the complex plane in the vicinity of the anti-Stokes lines where we can see both the dominant and subdominant series. If we were to keep adding terms to our asymptotic series, so as to minimize the error in the dominant solution, we would eventually be forced to conclude that a jump in the coefficient of the subdominant series can only take place on those lines in the complex plane on which $\mathrm{Im}(z) = 0$: these are called *Stokes lines*. This result was first proved by Stokes in 1857.[15] On a Stokes line the magnitude of the dominant series achieves its *maximum* value with respect to that of the subdominant series. Once we know that a jump in the coefficient of the subdominant series can only take place at a Stokes line, we can retain the subdominant series in our asymptotic expansion in all regions of the complex plane. What we are basically saying is that, although, in practice, we cannot actually see the subdominant series very far away from an anti-Stokes line because we are only retaining the first term in each asymptotic series, we could, in principle, see the subdominant series at all values of $\arg(z)$ provided that we retained a sufficient number of terms in our asymptotic series.

Figure 17 shows the location in the complex plane of the Stokes and anti-Stokes lines for the asymptotic expansion of the hypergeometric function. Also shown is a branch cut, which is needed to make z single-valued. The branch cut is chosen such that $\arg(z) = 0$ on the positive real axis. Every time we cross an anti-Stokes line the dominant series becomes subdominant and *vice versa*. Every time we cross a Stokes line the coefficient in front of the dominant series stays the same, but that in front of the subdominant series jumps discontinuously [see Eqs. (4.252)]. Finally, the jumps in the coefficient of the subdominant series are such as to ensure that the asymptotic expansion is single-valued.

[15]G.G. Stokes, *Trans. Camb. Phil. Soc.* **10**, 106–128 (1857)

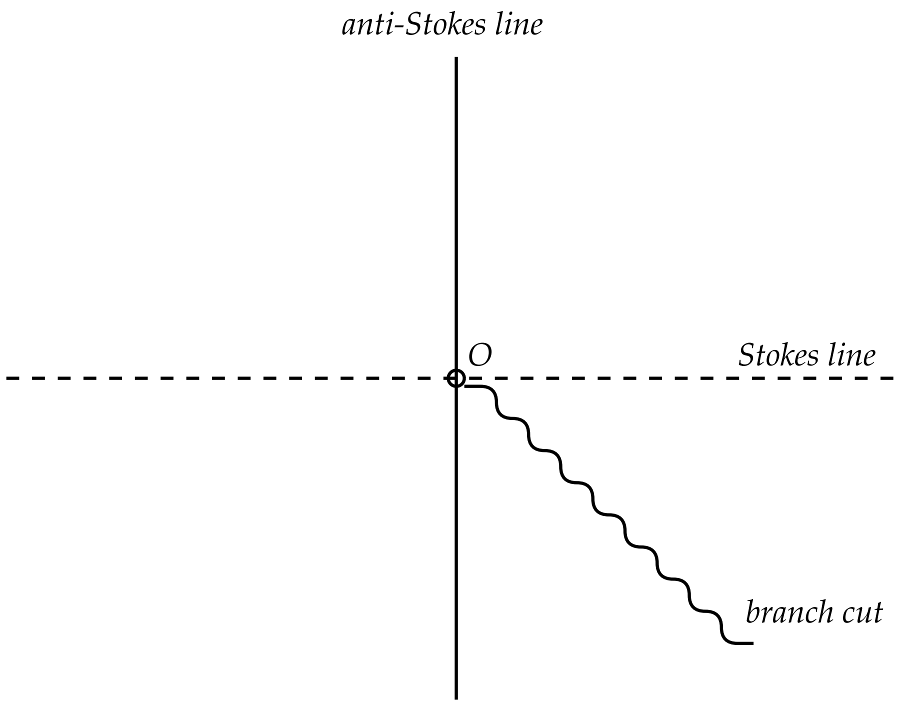


Figure 17: The location of the Stokes lines (dashed), the anti-Stokes lines (solid), and the branch cut (wavy) in the complex plane for the asymptotic expansion of the hypergeometric function

4.23 The W.K.B. solutions as asymptotic series

We have seen that the W.K.B. solution

$$E_y = n^{-1/2} \exp\left(\pm\mathrm{i}\,k \int^z n\,dz\right) \tag{4.253}$$

is an approximate solution of the differential equation

$$\frac{d^2 E_y}{dz^2} + k^2 n^2(z)\,E_y = 0 \tag{4.254}$$

in the limit where the typical wavelength, $2\pi/nk$, is much smaller than the typical variation length-scale of the refractive index. But, what sort of approximation is involved in writing this solution?

It is convenient to define the scaled variable

$$\hat{z} = \frac{z}{L}, \tag{4.255}$$

where L is the typical variation length-scale of the refractive index, $n(z)$. Equation (4.254) can then be written

$$w'' + h^2\,q\,w = 0, \tag{4.256}$$

where $w(\hat{z}, h) \equiv E_y(L\,\hat{z})$, $q(\hat{z}) \equiv n^2(L\,\hat{z})$, $' \equiv d/d\hat{z}$, and $h = kL$. Note that, in general, $q(\hat{z})$, $q'(\hat{z})$, $q''(\hat{z})$, *etc.* are $O(1)$ quantities. The non-dimensional constant h is of order the ratio of the variation length-scale of the refractive index to the wavelength. Let us seek the solutions to Eq. (4.256) in the limit $h \gg 1$.

We can write

$$w(\hat{z}, h) = \exp\left[\,\mathrm{i}\,h\,\phi(\hat{z}, h)\right]. \tag{4.257}$$

Equation (4.256) transforms to

$$\frac{\mathrm{i}}{h}\,\phi'' - (\phi')^2 + q = 0. \tag{4.258}$$

Expanding in powers of $1/h$, we obtain

$$\phi' = \pm q^{1/2} + \frac{\mathrm{i}}{4h}\frac{q'}{q} + O\left(\frac{1}{h^2}\right), \tag{4.259}$$

which yields

$$w(\hat{z}, h) = q^{-1/4} \exp\left(\pm\mathrm{i}\, h \int^{\hat{z}} q\, d\hat{z}\right) \left[1 + O\left(\frac{1}{h}\right)\right]. \tag{4.260}$$

Of course, we immediately recognize this expression as a W.K.B. solution.

Suppose that we keep expanding in powers of $1/h$ in Eq. (4.259). The appropriate generalization of Eq. (4.260) is a series solution of the form

$$w(\hat{z}, h) = q^{-1/4} \exp\left(\pm\mathrm{i}\, h \int^{\hat{z}} q\, d\hat{z}\right) \left[1 + \sum_{p=1}^{\infty} \frac{A_p(\hat{z})}{h^p}\right]. \tag{4.261}$$

This is, in fact, an *asymptotic series* in h. We can now appreciate that a W.K.B. solution is just a highly truncated asymptotic series in h, in which only the first term in the series is retained.

But, why is it so important that we recognize that W.K.B. solutions are highly truncated asymptotic series? The point is that the W.K.B. method was initially rather controversial after it was popularized in the 1920s. A lot of people thought that the method was completely wrong. Let us try to understand what the problem was. Suppose that we have never heard of an asymptotic series. Looking at Eq. (4.261), we would imagine that the expression in square brackets is a power law expansion in $1/h$. The W.K.B. approximation involves neglecting all terms in this expansion except the first. This sounds fine, as long as h is much greater than unity. But, surely, to be mathematically rigorous, we have to check that the sum of all of the terms in the expansion which we are neglecting is *small* compared to the first term? However, if we attempt this we discover, much to our consternation, that the expansion is *divergent*. In other words, the sum of all of the neglected terms is infinite! Thus, if we interpret Eq. (4.261) as a conventional power law expansion in $1/h$, the W.K.B. method is clearly nonsense: the W.K.B. solution is the first approximation to infinity. However, once we appreciate that Eq. (4.261) is actually an asymptotic series in h, the fact that the series diverges becomes irrelevant. If we retain the first n terms in the series, the series approximates the exact solution of Eq. (4.261) with an intrinsic (fractional) error which is of order $1/h^n$ (*i.e.*, the first neglected term in the series). The error is minimized at a particular value of h. As the number of terms in the series is increased, the intrinsic error decreases, and the value of h at which the error is minimized increases. In particular, we can see that there is an intrinsic error associated with a W.K.B. solution which is of order $1/h$ times the solution.

It is amusing to note that if Eq. (4.261) were not a divergent series then it would be impossible to obtain total reflection of the W.K.B. solutions at the point $q = 0$. As we shall discover, the reflection is directly associated with the fact that the expansion (4.261) exhibits a Stokes phenomenon. It is, of course, impossible for a convergent power series expansion to exhibit a Stokes phenomenon.

4.24 Stokes constants

We have seen that the differential equation

$$w'' + h^2\, q(\hat{z})\, w = 0, \tag{4.262}$$

where $' \equiv d/d\hat{z}$, possesses approximate W.K.B. solutions of the form

$$(a, \hat{z}) \;=\; q^{-1/4} \exp\left(\mathrm{i}\, h \int_a^{\hat{z}} q^{1/2}\, d\hat{z}\right) \left[1 + O\left(\frac{1}{h}\right)\right], \tag{4.263a}$$

$$(\hat{z}, a) \;=\; q^{-1/4} \exp\left(-\mathrm{i}\, h \int_a^{\hat{z}} q^{1/2}\, d\hat{z}\right) \left[1 + O\left(\frac{1}{h}\right)\right]. \tag{4.263b}$$

Here, we have adopted an arbitrary phase reference level $\hat{z} = a$. The convenient notation $(a, \hat{z})$ is fairly self explanatory: a and $\hat{z}$ refer to the lower and upper bounds of integration, respectively, inside the exponential. It follows that the other W.K.B. solution can be written $(\hat{z}, a)$ (we can reverse the limits of integration inside the exponential to obtain *minus* the integral in $\hat{z}$ from $\hat{z} = a$ to $\hat{z} = \hat{z}$).

Up to now we have thought of $\hat{z}$ as a *real* variable representing scaled height in the ionosphere. Let us now generalize our analysis somewhat and think of $\hat{z}$ as a *complex* variable. There is nothing in our derivation of the W.K.B. solutions which depends crucially on $\hat{z}$ being a real variable, so we expect these solutions to remain valid when $\hat{z}$ is reinterpreted as a complex variable. Incidentally, we must now interpret $q(\hat{z})$ as some well behaved function of the complex variable. An approximate general solution of the differential equation (4.262) in the complex $\hat{z}$ plane can be written as as a linear superposition of the two W.K.B. solutions (4.263).

The parameter h is assumed to be much larger than unity. It is clear from Eqs. (4.263) that in some regions of the complex plane one of the W.K.B. solutions is going to be exponentially larger than the other. In such regions, it is not mathematically consistent to retain the smaller W.K.B. solution in the expression for the general solution, since the contribution of the smaller W.K.B. solution is less than the intrinsic error associated with the larger solution. Adopting the terminology introduced in Section 4.22, the larger W.K.B. solution is said to be *dominant*, and the smaller solution is said to be *subdominant*. Let us denote the W.K.B. solution (4.263a) as $(a,\hat{z})_d$ in regions of the complex plane where it is dominant, and as $(a,\hat{z})_s$ in regions where it is subdominant. An analogous notation is adopted for the second W.K.B. solution (4.263b).

Suppose that $q(\hat{z})$ possesses a simple zero at the point $\hat{z} = \hat{z}_0$ (chosen to be the origin for the sake of convenience). It follows that in the immediate neighbourhood of the origin we can write

$$q = a_1\,\hat{z} + a_2\,\hat{z}^2 + \cdots, \tag{4.264}$$

where $a_1 \neq 0$. It is convenient to adopt the origin as the phase reference point (*i.e.*, $a = 0$), so the two W.K.B. solutions become $(0,\hat{z})$ and $(\hat{z},0)$. We can define *anti-Stokes lines* in the complex $\hat{z}$ plane (see Section 4.22). These are lines which satisfy

$$\mathrm{Re}\left[\mathrm{i}\int_0^{\hat{z}} q^{1/2}\,d\hat{z}\right] = 0. \tag{4.265}$$

As we cross an anti-Stokes line, a dominant W.K.B. solution becomes subdominant, and *vice versa*. Thus, $(0,\hat{z})_d \leftrightarrow (0,\hat{z})_s$ and $(\hat{z},0)_d \leftrightarrow (\hat{z},0)_s$. In the immediate vicinity of an anti-Stokes line the two W.K.B. solutions have about the same magnitude, so it is mathematically consistent to include the contributions from both solutions in the expression for the general solution. In such a region, we can drop the subscripts d and s, since the W.K.B. solutions are neither dominant nor subdominant, and write the W.K.B. solutions simply as $(0, \hat{z})$ and $(\hat{z}, 0)$.

It is clear from Eqs. (4.263) that the W.K.B. solutions are not single-valued functions of $\hat{z}$, since they depend on $q^{1/2}(\hat{z})$, which is a double-valued function. Thus, if we wish to write an approximate *analytic* solution to the differential equation (4.262) we cannot express this solution as the *same* linear combination of W.K.B. solutions in all regions of the complex $\hat{z}$-plane. This implies that there must exist certain lines in the complex $\hat{z}$-plane across which the mix of W.K.B. solutions in our expression for the general solution changes discontinuously. These lines are called *Stokes lines* (see Section 4.22), and satisfy

$$\mathrm{Im}\left[\mathrm{i}\int_0^{\hat{z}} q^{1/2}\, d\hat{z}\right] = 0. \tag{4.266}$$

As we cross a Stokes line, the coefficient of the dominant W.K.B. solution in our expression for the general solution must remain unchanged, but the coefficient of the subdominant solution is allowed to change discontinuously. Incidentally, this is perfectly consistent with the fact that the general solution is analytic: the jump in our expression for the general solution due to the jump in the coefficient of the subdominant W.K.B. solution is *less* than the intrinsic error in this expression due to the intrinsic error in the dominant W.K.B. solution. Once we appreciate that the coefficient of the subdominant solution can only change at a Stokes line, we can retain both W.K.B. solutions in our expression for the general solution throughout the complex $\hat{z}$ plane. In practice, we can only see a subdominant solution in the immediate vicinity of an anti-Stokes line, but if we were to evaluate the W.K.B. solutions to higher accuracy [*i.e.* retain more terms in the asymptotic series in Eqs. (4.263)]we could, in principle, follow a subdominant solution all the way to a neighbouring Stokes line.

In the immediate vicinity of the origin

$$\int_0^{\hat{z}} q^{1/2}\, d\hat{z} \simeq \frac{2\sqrt{a_1}}{3}\,\hat{z}^{3/2}. \tag{4.267}$$

It follows from Eqs. (4.265) and (4.266) that *three* Stokes lines and *three* anti-Stokes lines radiate from a zero of $q(\hat{z})$. The general arrangement of Stokes and

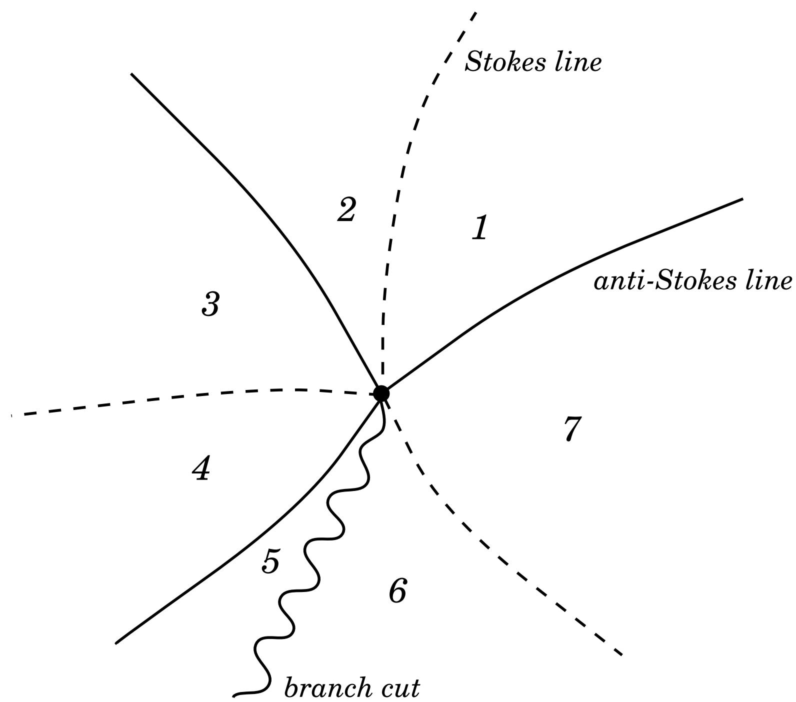


Figure 18: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) around a simple zero of $q(\hat{z})$. Also shown is the branch cut (wavy line). All of the lines radiate from the point $q = 0$.

anti-Stokes lines in the vicinity of a $q = 0$ point is sketched in Fig. 18. Note that a branch cut must also radiate from the $q = 0$ point in order to uniquely specify the function $q^{1/2}(\hat{z})$. Thus, in general, *seven* lines radiate from a zero of $q(\hat{z})$, dividing the complex $\hat{z}$ plane into seven domains (numbered 1 through 7).

Let us write our general solution as

$$w(\hat{z}, h) = A\,(0, \hat{z}) + B\,(\hat{z}, 0) \tag{4.268}$$

on the anti-Stokes line between domains 1 and 7, where A and B are arbitrary constants. Suppose that the W.K.B. solution $(0, \hat{z})$ is dominant in domain 7. Thus, in domain 7 the general solution takes the form

$$w(7) = A\,(0, \hat{z})_d + B\,(\hat{z}, 0)_s. \tag{4.269}$$

Let us move into domain 1. In doing so, we cross an anti-Stokes line, so the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 1 the general solution takes the form

$$w(1) = A\,(0,\hat{z})_s + B\,(\hat{z},0)_d. \tag{4.270}$$

Let us now move into domain 2. In doing so, we cross a Stokes line, so the coefficient of the dominant solution, B, must remain constant, but the coefficient of the subdominant solution, A, is allowed to change. Suppose that the coefficient of the subdominant solution jumps by t times the coefficient of the dominant solution, where t is an undetermined constant. It follows that in domain 2 the general solution takes the form

$$w(2) = (A + t\,B)\,(0,\hat{z})_s + B\,(\hat{z},0)_d. \tag{4.271}$$

Let us now move into domain 3. In doing so, we cross an anti-Stokes line, so the the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 3 the general solution takes the form

$$w(3) = (A + t\,B)\,(0,\hat{z})_d + B\,(\hat{z},0)_s. \tag{4.272}$$

Let us now move into domain 4. In doing so, we cross a Stokes line, so the coefficient of the dominant solution must remain constant, but the coefficient of the subdominant solution is allowed to change. Suppose that the coefficient of the subdominant solution jumps by u times the coefficient of the dominant solution, where u is an undetermined constant. It follows that in domain 4 the general solution takes the form

$$w(4) = (A + t\,B)\,(0,\hat{z})_d + (B + u\,[A + t\,B])\,(\hat{z},0)_s. \tag{4.273}$$

Let us now move into domain 5. In doing so, we cross an anti-Stokes line, so the the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 5 the general solution takes the form

$$w(5) = (A + t\,B)\,(0,\hat{z})_s + (B + u\,[A + t\,B])\,(\hat{z},0)_d. \tag{4.274}$$

Let us now move into domain 6. In doing so, we cross the branch cut in an anti-clockwise direction. Thus, the argument of $\hat{z}$ decreases by 2π. It follows from Eq. (4.264) that $q^{1/2} \to -q^{1/2}$ and $q^{1/4} \to -\mathrm{i}\,q^{1/4}$. The following rules for tracing the W.K.B. solutions across the branch cut (in an anti-clockwise direction) ensure that the general solution is continuous across the cut [see Eqs. (4.261)]:

$$(0,\hat{z}) \quad\rightarrow\quad -\mathrm{i}\,(\hat{z},0), \tag{4.275a}$$

$$(\hat{z},0) \quad\rightarrow\quad -\mathrm{i}\,(0,\hat{z}). \tag{4.275b}$$

Note that the properties of dominancy and subdominancy are preserved when the branch cut is crossed. It follows that in domain 6 the general solution takes the form

$$w(6) = -\mathrm{i}\,(A + t\,B)\,(\hat{z},0)_s - \mathrm{i}\,(B + u\,[A + t\,B])\,(0,\hat{z})_d. \tag{4.276}$$

Let us, finally, move into domain 7. In doing so, we cross a Stokes line, so the coefficient of the dominant solution must remain constant, but the coefficient of the subdominant solution is allowed to change. Suppose that the coefficient of the subdominant solution jumps by v times the coefficient of the dominant solution, where v is an undetermined constant. It follows that in domain 7 the general solution takes the form

$$w(7) = -\mathrm{i}\,(A{+}t\,B{+}v\,\{B{+}u\,[A{+}t\,B]\})\,(\hat{z},0)_s{-}\mathrm{i}\,(B{+}u\,[A{+}t\,B])\,(0,\hat{z})_d. \tag{4.277}$$

Now, we expect our general solution to be an *analytic* function, so it follows that the solutions (4.269) and (4.277) must be identical. Thus, we can compare the coefficients of the two W.K.B. solutions, $(\hat{z},0)_s$ and $(0,\hat{z})_d$. Since A and B are arbitrary, we can also compare the coefficients of A and B. Thus, comparing the coefficients of $A\,(0,\hat{z})_d$, we find

$$1 = -\mathrm{i}\,u. \tag{4.278}$$

Comparing the coefficients of $B\,(0,\hat{z})_d$ yields

$$0 = 1 + ut. \tag{4.279}$$

Comparing the coefficients of $A\,(\hat{z},0)_s$ gives

$$0 = 1 + vu. \tag{4.280}$$

Finally, comparing the coefficients of $B\,(\hat{z},0)_s$ yields

$$1 = -\mathrm{i}\,(t + v + v u t). \tag{4.281}$$

Equations (4.278)–(4.281) imply that

$$t = u = v = \mathrm{i}. \tag{4.282}$$

In other words, if we adopt the simple rule that every time we cross a Stokes line in an anti-clockwise direction the coefficient of the subdominant solution jumps by i times the coefficient of the dominant solution, then this *ensures* that our expression for the general solution (4.268) behaves as an analytic function. Here, the constant i is usually called a *Stokes constant.* Note that if we cross a Stokes line in a clockwise direction then the coefficient of the subdominant solution has to jump by $-\mathrm{i}$ times the coefficient of the dominant solution in order to ensure that our general solution behaves as an analytic function.

4.25 The reflection coefficient

Let us write $\hat{z} = x + \mathrm{i}\,y$, where x and y are real variables. Consider the solution of the differential equation

$$w'' + h^2\, q(x)\, w = 0, \tag{4.283}$$

where $q(x)$ is a real function, h is a large number, $q > 0$ for $x < 0$, and $q < 0$ for $x > 0$. It is clear that $\hat{z} = 0$ represents a simple zero of $q(\hat{z})$. Here, we assume, as seems eminently reasonable, that we can find a well behaved function of the complex variable $q(\hat{z})$ such that $q(\hat{z}) = q(x)$ along the real axis. The arrangement of Stokes and anti-Stokes lines in the immediate vicinity of the point $\hat{z} = 0$ is sketched in Fig. 19. The argument of $q(\hat{z})$ on the positive x-axis is chosen to be $-\pi$. Thus, the argument of $q(\hat{z})$ on the negative x-axis is 0.

On OB, the two W.K.B. solutions (4.261) can be written

$$(0,x) \;=\; q^{-1/4}(x)\,\exp\left(\mathrm{i}\,h\int_0^x q^{1/2}(x)\,dx\right), \tag{4.284a}$$

$$(x,0) \;=\; q^{-1/4}(x)\,\exp\left(-\mathrm{i}\,h\int_0^x q^{1/2}(x)\,dx\right). \tag{4.284b}$$

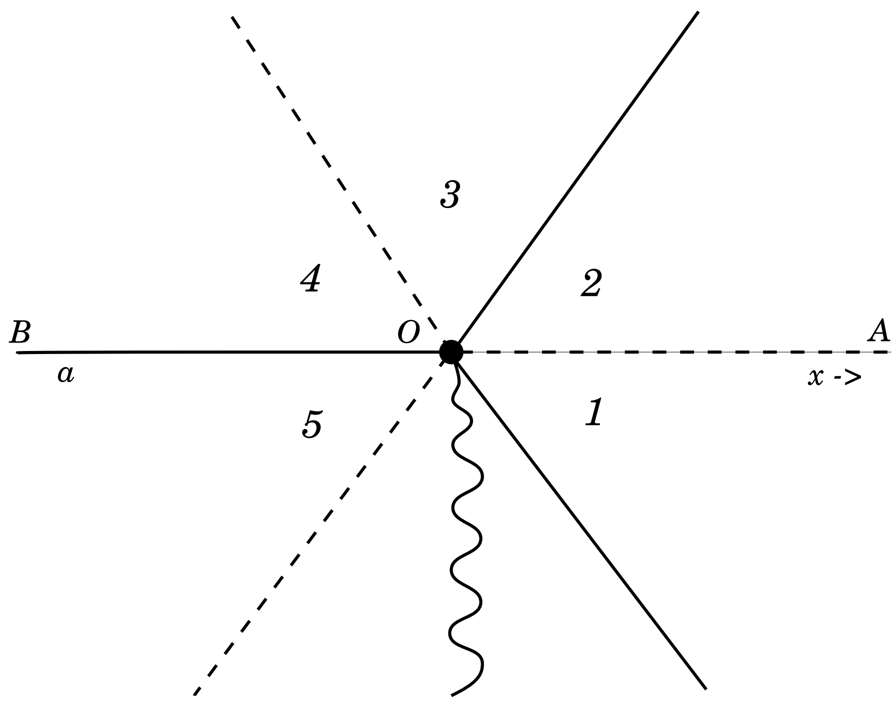


Figure 19: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) in the complex $\hat{z}$ plane. Also shown is the branch cut (wavy line).

Here, we can interpret $(0, x)$ as a wave propagating to the right along the x-axis, and $(x, 0)$ as a wave propagating to the left. On OA, the W.K.B. solutions take the form

$$(0, x)_d \quad = \quad \mathrm{e}^{\,\mathrm{i}\,\pi/4}\,|q(x)|^{-1/4}\,\exp\left(+h\int_0^x |q(x)|^{1/2}\,dx\right), \tag{4.285a}$$

$$(x, 0)_s \quad = \quad \mathrm{e}^{\,\mathrm{i}\,\pi/4}\,|q(x)|^{-1/4}\,\exp\left(-h\int_0^x |q(x)|^{1/2}\,dx\right). \tag{4.285b}$$

Clearly, $(x, 0)_s$ represents an evanescent wave which decays to the right along the x-axis, whereas $(0, x)_d$ represents an evanescent wave which decays to the left. If we adopt the boundary condition that there is no incident wave from the region $x \to +\infty$, the most general asymptotic solution to Eq. (4.283) on OA is written

$$w(x, h) = A\,(x, 0)_s, \tag{4.286}$$

where A is an arbitrary constant.

Let us assume that we can find an analytic solution $w(\hat{z}, h)$ to the differential equation

$$w'' + h^2\,q(\hat{z})\,w = 0, \tag{4.287}$$

which satisfies $w(\hat{z}, h) = w(x, h)$ along the real axis, where $w(x, h)$ is the physical solution. From a mathematical point of view, this seems eminently reasonable. In the domains 1 and 2 the solution (4.286) becomes

$$w(1) = A\,(\hat{z}, 0)_s, \tag{4.288}$$

and

$$w(2) = A\,(\hat{z}, 0)_s. \tag{4.289}$$

Note that the solution is continuous across the Stokes line OA, since the coefficient of the dominant solution $(0, \hat{z})$ is zero: thus, the jump in the coefficient of the subdominant solution is zero times the Stokes constant, i; *i.e.*, it is zero. Let us move into domain 3. In doing so, we cross an anti-Stokes line, so the solution becomes

$$w(3) = A\,(\hat{z}, 0)_d. \tag{4.290}$$

Let us now move into domain 4. In doing so, we cross a Stokes line. Applying the general rule derived in the preceding section, the solution becomes

$$w(4) = A\,(\hat{z},0)_d + \mathrm{i}\,A\,(0,\hat{z})_s. \tag{4.291}$$

Finally, on OB the solution becomes

$$w(x,h) = A\,(x,0) + \mathrm{i}\,A\,(0,x). \tag{4.292}$$

Suppose that there is a point a on the negative x-axis where $q(x) = 1$. It follows from Eqs. (4.286) and (4.292) that we can write the asymptotic solution to Eq. (4.283) as

$$\begin{aligned} w(x,h) &= q^{-1/4}\exp\left(\mathrm{i}\,h\int_a^x q^{1/2}\,dx\right) \qquad (4.293)\\ &\quad -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right)\,q^{-1/4}\exp\left(-\mathrm{i}\,h\int_a^x q^{1/2}\,dx\right), \end{aligned}$$

in the region $x < 0$, and

$$w(x,h) = \exp\left(\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right)\,\mathrm{e}^{-\mathrm{i}\,\pi/4}\,|q|^{-1/4}\exp\left(-h\int_0^x |q|^{1/2}\,dx\right) \tag{4.294}$$

in the region $x > 0$. Here, we have chosen

$$A = -\mathrm{i}\,\exp\left(\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right). \tag{4.295}$$

If we interpret x as a normalized altitude in the ionosphere, $q(x)$ as the square of the refractive index in the ionosphere, the point a as ground level, and w as the electric field strength of a radio wave propagating vertically upwards into the ionosphere, then Eq. (4.293) tells us that a unit amplitude wave fired vertically upwards from ground level into the ionosphere is *reflected* at the level where the refractive index is zero. The first term in Eq. (4.293) is the incident wave and the second term is the reflected wave. The reflection coefficient (*i.e.*, the ratio of the reflected to the incident wave at ground level) is given by

$$R = -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right). \tag{4.296}$$

Note that $|R| = 1$, so the amplitude of the reflected wave equals that of the incident wave. In other words, there is no absorption of the wave at the level of reflection. The phase shift of the reflected wave at ground level, with respect to that of the incident wave, is that associated with the wave propagating from ground level to the reflection level and back to ground level again, *plus* a $-\pi/2$ phase shift at reflection. According to Eq. (4.294), the wave attenuates fairly rapidly (in the space of a few wavelengths) above the reflection level. Of course, Eq. (4.296) is completely equivalent to Eq. (4.186).

Note that the reflection of the incident wave at the point where the refractive index is zero is directly associated with the Stokes phenomenon. Without the jump in the coefficient of the subdominant solution, as we go from domain 3 to domain 4, there is no reflected wave on the OB axis. Note, also, that the W.K.B. solutions (4.293) and (4.294) break down in the immediate vicinity of $q = 0$ (*i.e.*, the reflection point). Thus, it is possible to demonstrate that the incident wave is totally reflected at the point $q = 0$, with a $-\pi/2$ phase shift, without having to solve for the wave structure in the immediate vicinity of the reflection point. This demonstrates that the reflection of the incident wave at $q = 0$ is an intrinsic property of the W.K.B. solutions, and does not depend on the detailed behaviour of the wave in the region where the W.K.B. solutions break down.

4.26 The Jeffries connection formula

In the preceding section there is a tacit assumption that the square of the refractive index, $q(x) \equiv n^2(x)$, is a *real* function. As is apparent from Eq. (4.162), this is only the case in the ionosphere as long as electron collisions are negligible. Let us generalize our analysis to take electron collisions into account. In fact, the main effect of electron collisions is to move the zero of $q(\hat{z})$ a short distance off the real axis (the distance is relatively short provided that we adopt the physical ordering $\nu \ll \omega$). The arrangement of Stokes and anti-Stokes lines around the new zero point, located at $\hat{z} = \hat{z}_0$, is sketched in Fig. 20. Note that electron collisions only significantly modify the form of $q(\hat{z})$ in the immediate vicinity of the zero point. Thus, sufficiently far away from $\hat{z} = \hat{z}_0$ in the complex $\hat{z}$-plane, the W.K.B. solutions and the locations of the Stokes and anti-Stokes lines are exactly the same as in the preceding section.

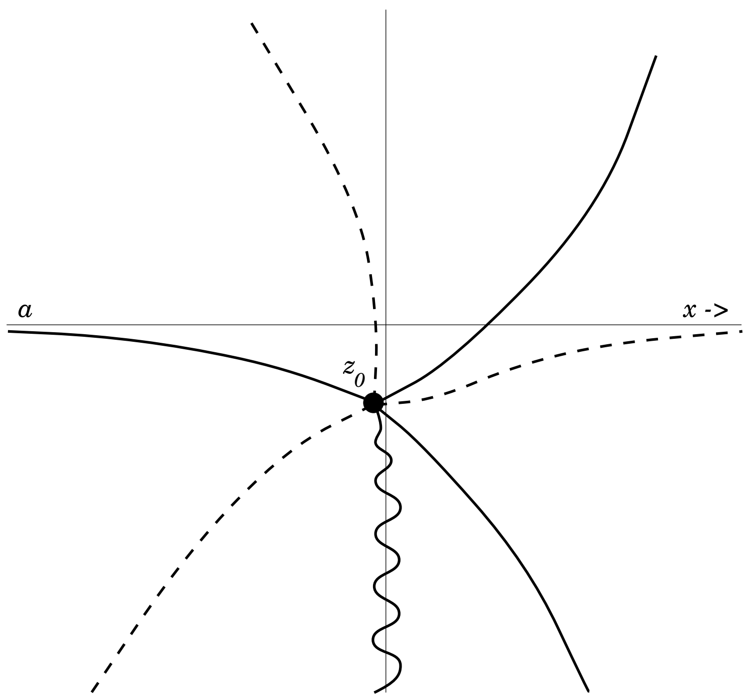


Figure 20: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) in the complex $\hat{z}$ plane. Also shown is the branch cut (wavy line).

The W.K.B. solutions (4.284) and (4.285) are valid all the way along the real axis, except for a small region close to the origin where electron collisions significantly modify the form of $q(\hat{z})$. Thus, we can still adopt the physically reasonable decaying solution (4.286) on the positive real axis. Let us trace this solution in the complex $\hat{z}$-plane until we reach the negative real axis. We can achieve this by moving in a semi-circle in the upper half-plane. Since we never move out of the region in which the W.K.B. solutions (4.284) and (4.285) are valid, we conclude, by analogy with the preceding section, that the solution on the negative real axis is given by Eq. (4.292). Of course, in all of the W.K.B. solutions the point $\hat{z} = 0$ must be replaced by the new zero point $\hat{z} = \hat{z}_0$. The new formula for the reflection coefficient, which is just a straightforward generalization of Eq. (4.296), is

$$R = -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^{\hat{z}_0} q^{1/2}\,d\hat{z}\right). \tag{4.297}$$

This is called the *Jeffries connection formula*, after H. Jeffries, who discovered it in 1923. The general expression for the reflection coefficient is incredibly simple. We just integrate the W.K.B. solution in the complex $\hat{z}$-plane from the phase reference level $\hat{z} = a$ to the zero point, square the result, and multiply by $-\mathrm{i}$. Note that the path of integration between $\hat{z} = a$ and $\hat{z} = \hat{z}_0$ does not matter, because of Cauchy's theorem. Note, also, that since $q^{1/2}$ is, in general, *complex* along the path of integration, we no longer have $|R| = 1$. In fact, it is easily demonstrated that $|R| \leq 1$. Thus, when electron collisions are included in the analysis we no longer obtain perfect reflection of radio waves from the ionosphere. Instead, some (small) fraction of the radio energy is *absorbed* at each reflection event. This energy is ultimately transfered to the particles in the ionosphere with which the electrons collide.

5 Radiation and scattering

5.1 Basic antenna theory

It possible to solve exactly for the radiation pattern emitted by a linear antenna fed with a sinusoidal current pattern. Assuming that all fields and currents vary in time like $\mathrm{e}^{-\mathrm{i}\,\omega t}$, and adopting the Lorentz gauge, it is easily demonstrated that the vector potential obeys the inhomogeneous Helmholtz equation

$$(\nabla^2 + k^2)\boldsymbol{A} = -\mu_0\,\boldsymbol{j}, \tag{5.1}$$

where $k = \omega/c$. The Green's function for this equation, subject to the Sommerfeld radiation condition (which ensures that sources radiate waves instead of absorbing them), is given by Eq. (2.123). Thus, we can invert Eq. (5.1) to obtain

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')\,\mathrm{e}^{\,\mathrm{i}\,k|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{5.2}$$

The electric field in the source free region follows from the Ampère-Maxwell equation and $\boldsymbol{B} = \nabla\wedge\boldsymbol{A}$,

$$\boldsymbol{E} = \frac{\mathrm{i}}{k}\nabla\wedge c\boldsymbol{B}. \tag{5.3}$$

Now

$$|\boldsymbol{r}-\boldsymbol{r}'| = r\sqrt{1 - 2\,\boldsymbol{n}\cdot\boldsymbol{r}'/r + r'^2/r^2}, \tag{5.4}$$

where $\boldsymbol{n} = \boldsymbol{r}/r$. Assuming that $r' \ll r$, this expression can be expanded binomially to give

$$|\boldsymbol{r}-\boldsymbol{r}'| = r\left[1 - \frac{\boldsymbol{n}\cdot\boldsymbol{r}'}{r} + \frac{r'^2}{2r^2} - \frac{1}{8}\left(\frac{2\boldsymbol{n}\cdot\boldsymbol{r}'}{r}\right)^2 + \cdots\right], \tag{5.5}$$

where we have retained all terms up to order $(r'/r)^2$. This expansion occurs in the complex exponential of Eq. (5.2); *i.e.*, it determines the oscillation phase of each element of the antenna. The quadratic terms in the expansion can be neglected provided they can be shown to contribute a phase shift which is significantly less than 2π. Since the maximum possible value of r' is $d/2$, for a linear antenna which extends along the z-axis from $z = -d/2$ to $z = d/2$, the phase shift associated with the quadratic terms is insignificant as long as

$$r \gg \frac{k d^2}{16\pi} = \frac{d^2}{8\lambda}, \tag{5.6}$$

where $\lambda = 2\pi/k$ is the wavelength of the radiation. This constraint is known as the *Fraunhofer limit*.

In the Fraunhofer limit we can approximate the phase variation of the complex exponential in Eq. (5.2) by a linear function of r':

$$|\boldsymbol{r} - \boldsymbol{r}'| \rightarrow r - \boldsymbol{n}\cdot\boldsymbol{r}'. \tag{5.7}$$

The denominator $|\boldsymbol{r} - \boldsymbol{r}'|$ in the integrand of Eq. (5.2) can be approximated as r provided that the distance from the antenna is much greater than its length; *i.e.*, provided that

$$r \gg d. \tag{5.8}$$

Thus, Eq. (5.2) reduces to

$$\boldsymbol{A}(\boldsymbol{r}) \simeq \frac{\mu_0}{4\pi}\frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{r}\int \boldsymbol{j}(\boldsymbol{r}')\,\mathrm{e}^{-\mathrm{i}\,k\boldsymbol{n}\cdot\boldsymbol{r}'}\,d^3\boldsymbol{r}' \tag{5.9}$$

when the constraints (5.6) and (5.8) are satisfied. If the additional constraint

$$kr \gg 1 \tag{5.10}$$

is also satisfied, then the electromagnetic fields associated with Eq. (5.9) take the form

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{r}) &\simeq \mathrm{i}\,k\,\boldsymbol{n}\wedge\boldsymbol{A} = \mathrm{i}\,k\,\frac{\mu_0}{4\pi}\frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{r}\int \boldsymbol{n}\wedge\boldsymbol{j}(\boldsymbol{r}')\,\mathrm{e}^{-\mathrm{i}\,k\boldsymbol{n}\cdot\boldsymbol{r}'}\,d^3\boldsymbol{r}', && \text{(5.11a)}\\ \boldsymbol{E}(\boldsymbol{r}) &\simeq c\boldsymbol{B}\wedge\boldsymbol{n} = \mathrm{i}\,ck\,(\boldsymbol{n}\wedge\boldsymbol{A})\wedge\boldsymbol{n}. && \text{(5.11b)} \end{aligned}$$

These are clearly radiation fields, since they are mutually orthogonal, transverse to the radius vector, and satisfy $E = cB \propto r^{-1}$. The three constraints (5.6), (5.8), and (5.10), can be summed up in a single inequality:

$$d \ll \sqrt{\lambda r} \ll r. \tag{5.12}$$

The current density associated with a linear, sinusoidal, centre-fed antenna is

$$\boldsymbol{j}(\boldsymbol{r}) = I\sin(kd/2 - k|z|)\,\delta(x)\,\delta(y)\,\hat{\boldsymbol{z}} \tag{5.13}$$

for $|z| < d/2$, with $\boldsymbol{j}(\boldsymbol{r}) = 0$ for $|z| \geq d/2$. In this case, Eq. (5.9) yields

$$\boldsymbol{A}(\boldsymbol{r}) = \hat{\boldsymbol{z}}\,\frac{\mu_0\,I}{4\pi}\,\frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{r}\int_{-d/2}^{d/2}\sin(kd/2 - k|z|)\,\mathrm{e}^{-\mathrm{i}\,kz\cos\theta}\,dz, \tag{5.14}$$

where $\cos\theta = \boldsymbol{n}\cdot\hat{\boldsymbol{z}}$. The result of this straightforward integration is

$$\boldsymbol{A}(\boldsymbol{r}) = \hat{\boldsymbol{z}}\,\frac{\mu_0\,I}{4\pi}\,\frac{2\,\mathrm{e}^{\,\mathrm{i}\,kr}}{kr}\left[\frac{\cos(kd\,\cos\theta/2) - \cos(kd/2)}{\sin^2\theta}\right]. \tag{5.15}$$

Note from Eqs. (5.11) that the electric field lies in the plane containing the antenna and the radius vector to the observation point. The time-averaged power radiated by the antenna per unit solid angle is

$$\frac{dP}{d\Omega} = \frac{\mathrm{Re}\,(\boldsymbol{n}\cdot\boldsymbol{E}\wedge\boldsymbol{B}^*)\,r^2}{2\mu_0} = \frac{ck^2\,\sin^2\theta\,|A|^2\,r^2}{2\mu_0}. \tag{5.16}$$

Thus,

$$\frac{dP}{d\Omega} = \frac{\mu_0 c\,I^2}{8\pi^2}\left|\frac{\cos(kd\,\cos\theta/2) - \cos(kd/2)}{\sin\theta}\right|^2. \tag{5.17}$$

The angular distribution of power depends on the value of kd. In the long wavelength limit $kd \ll 1$ the distribution reduces to

$$\frac{dP}{d\Omega} = \frac{\mu_0 c\,{I_0}^2}{128\pi^2}\,(kd)^2\sin^2\theta, \tag{5.18}$$

where $I_0 = I\,kd/2$ is the peak current in the antenna. It is easily shown from Eq. (5.13) that the current distribution in the antenna is linear:

$$I(z) = I_0(1 - 2|z|/d) \tag{5.19}$$

for $|z| < d/2$. This type of antenna corresponds to a short (compared to the wavelength) oscillating electric dipole, and is generally known as a *Hertzian oscillating dipole*. The total power radiated is

$$P = \frac{\mu_0 c\,{I_0}^2\,(kd)^2}{48\pi}. \tag{5.20}$$

In order to maintain the radiation, power must be supplied continuously to the oscillating dipole from some generator. By analogy with the heating power produced in a resistor,

$$\langle P\rangle_{\rm heat} = \langle I^2\rangle R = \frac{I_0^{\,2}\,R}{2}, \tag{5.21}$$

we can define the factor which multiplies $I_0^{\,2}/2$ in Eq. (5.20) as the *radiation resistance* of the dipole antenna:

$$R_{\rm rad} = \sqrt{\frac{\mu_0}{\epsilon_0}}\,\frac{(kd)^2}{24\pi} = 197\left(\frac{d}{\lambda}\right)^2 \text{ ohms.} \tag{5.22}$$

Since we have assumed that $\lambda \gg d$, this radiation resistance is necessarily very small. Typically, in devices of this sort the radiated power is swamped by the ohmic losses appearing as heat. Thus, a "short" dipole is a very inefficient radiator. Practical antennas have dimensions which are comparable with the wavelength of the emitted radiation.

Probably the most common practical antennas are the half-wave antenna ($kd = \pi$) and the full-wave antenna ($kd = 2\pi$). In the former case, Eq. (5.17) reduces to

$$\frac{dP}{d\Omega} = \frac{\mu_0 c\, I^2}{8\pi^2}\,\frac{\cos^2(\pi\cos\theta/2)}{\sin^2\theta}. \tag{5.23}$$

In the latter case, Eq. (5.17) yields

$$\frac{dP}{d\Omega} = \frac{\mu_0 c\, I^2}{2\pi^2}\,\frac{\cos^4(\pi\cos\theta/2)}{\sin^2\theta}. \tag{5.24}$$

The half-wave antenna radiation pattern is very similar to the characteristic $\sin^2\theta$ pattern of a Hertzian dipole. However, the full-wave antenna radiation pattern is considerably sharper (*i.e.*, it is more concentrated in the transverse directions $\theta = \pm\pi/2$).

The total power radiated by a half-wave antenna is

$$P = \frac{\mu_0 c\, I^2}{4\pi}\int_0^\pi \frac{\cos^2(\pi\cos\theta/2)}{\sin\theta}\,d\theta. \tag{5.25}$$

The integral can be evaluated numerically to give 1.2188. Thus,

$$P = 1.2188\, \frac{\mu_0 c\, I^2}{4\pi}. \tag{5.26}$$

Note from Eq. (5.13) that I is equivalent to the peak current flowing in the antenna. Thus, the radiation resistance of a half-wave antenna is given by $P/(I^2/2)$, or

$$R_{\rm rad} = \frac{0.6094}{\pi}\sqrt{\frac{\mu_0}{\epsilon_0}} = 73 \text{ ohms}. \tag{5.27}$$

This resistance is substantially larger than that for a Hertzian dipole (see Eq. (5.22)). In other words, a half-wave antenna is a far more efficient radiator of electromagnetic radiation than a Hertzian dipole. According to standard transmission line theory, if a transmission line is terminated by a resistor whose resistance matches the characteristic impedance of the line, then all of the power transmitted down the line is dissipated in the resistor. On the other hand, if the resistance does not match the impedance of the line then some of the power is reflected and returned to the generator. We can think of a half-wave antenna, centre-fed by a transmission line, as a 73 ohm resistor terminating the line. The only difference is that the power absorbed from the line is radiated rather than dissipated as heat. Thus, in order to avoid problems with reflected power the impedance of a transmission line feeding a half-wave antenna must be 73 ohms. Not surprisingly, 73 ohm impedance is one of the standard ratings for the co-axial cables used in amateur radio.

5.2 Antenna directivity and effective area

We have seen that standard antennas emit more radiation in some directions than in others. Indeed, it is topologically impossible for an antenna to emit *transverse* waves uniformly in all directions (for the same reason that it is impossible to comb the hair on a sphere in such a manner that there is no parting). One of the aims of antenna engineering is to design antennas which transmit most of their radiation in a particular direction. By a reciprocity argument, such an antenna, when used as a receiver, is preferentially sensitive to radiation incident from the same direction.

The *directivity* or *gain* of an antenna is defined as the ratio of the *maximum* value of the power radiated per unit solid angle, to the average power radiated per unit solid angle:

$$G = \frac{(dP/d\Omega)_{\rm max}}{P/4\pi}. \tag{5.28}$$

Thus, the directivity measures how much more intensely the antenna radiates in its preferred direction than a mythical "isotropic radiator" would when fed with the same total power. For a Hertzian dipole the gain is 3/2. For a half-wave antenna the gain is 1.64. To achieve a directivity which is significantly greater than unity, the antenna size needs to be much larger than the wavelength. This is usually achieved using a phased array of half-wave or full-wave antennas.

Antennas can be used to receive, as well as emit, electromagnetic radiation. The incoming wave induces a voltage in the antenna which can be detected in an electrical circuit connected to the antenna. In fact, this process is equivalent to the emission of electromagnetic waves by the antenna viewed in reverse. In the theory of electrical circuits, a receiving antenna is represented as an e.m.f connected in series with a resistor. The e.m.f., $V_0 \cos\omega t$, represents the voltage induced in the antenna by the incoming wave. The resistor, $R_{\rm rad}$, represents the power re-radiated by the antenna (here, the real resistance of the antenna is neglected). Let us represent the detector circuit as a single load resistor $R_{\rm load}$ connected in series with the antenna. The question is: how can we choose $R_{\rm load}$ so that the maximum power is extracted from the wave and transmitted to the load resistor? According to Ohm's law:

$$V_0 \cos\omega t = I_0 \cos\omega t\,(R_{\rm rad} + R_{\rm load}), \tag{5.29}$$

where $I = I_0 \cos\omega t$ is the current induced in the circuit.

The power input to the circuit is

$$P_{\rm in} = \langle V I\rangle = \frac{{V_0}^2}{2(R_{\rm rad} + R_{\rm load})}. \tag{5.30}$$

The power transferred to the load is

$$P_{\rm load} = \langle I^2 R_{\rm load}\rangle = \frac{R_{\rm load}\,{V_0}^2}{2(R_{\rm rad} + R_{\rm load})^2}. \tag{5.31}$$

The power re-radiated by the antenna is

$$P_{\rm rad} = \langle I^2 R_{\rm rad}\rangle = \frac{R_{\rm rad}\, {V_0}^{\,2}}{2(R_{\rm rad}+R_{\rm load})^2}. \tag{5.32}$$

Note that $P_{\rm in} = P_{\rm load} + P_{\rm rad}$. The maximum power transfer to the load occurs when

$$\frac{\partial P_{\rm load}}{\partial R_{\rm load}} = \frac{{V_0}^{\,2}}{2}\left[\frac{R_{\rm load}-R_{\rm rad}}{(R_{\rm rad}+R_{\rm load})^3}\right] = 0. \tag{5.33}$$

Thus, the maximum transfer rate corresponds to

$$R_{\rm load} = R_{\rm res}. \tag{5.34}$$

In other words, the resistance of the load circuit must match the radiation resistance of the antenna. For this optimum case,

$$P_{\rm load} = P_{\rm rad} = \frac{{V_0}^{\,2}}{8R_{\rm rad}} = \frac{P_{\rm in}}{2}. \tag{5.35}$$

So, even in the optimum case one *half* of the power absorbed by the antenna is immediately re-radiated. If $R_{\rm load} \neq R_{\rm res}$ then more than one half of the absorbed power is re-radiated. Clearly, an antenna which is receiving electromagnetic radiation is also emitting it. This is how the BBC catch people who do not pay their television license fee in England. They have vans which can detect the radiation emitted by a TV aerial whilst it is in use (they can even tell which channel you are watching!).

For a Hertzian dipole antenna interacting with an incoming wave whose electric field has an amplitude E_0 we expect

$$V_0 = E_0\, d/2. \tag{5.36}$$

Here, we have used the fact that the wavelength of the radiation is much longer than the length of the antenna, and that the relevant e.m.f. develops between the two ends and the centre of the antenna. We have also assumed that the antenna is properly aligned (*i.e.*, the radiation is incident perpendicular to the axis of the antenna). The Poynting flux of the incoming wave is

$$\langle u_{\rm in}\rangle = \frac{\epsilon_0 c\, {E_0}^{\,2}}{2}, \tag{5.37}$$

whereas the power transferred to a properly matched detector circuit is

$$P_{\rm load} = \frac{{E_0}^2 d^2}{32 R_{\rm rad}}. \tag{5.38}$$

Consider an idealized antenna in which all incoming radiation incident on some area $A_{\rm eff}$ is absorbed and then magically transferred to the detector circuit with no re-radiation. Suppose that the power absorbed from the idealized antenna matches that absorbed from the real antenna. This implies that

$$P_{\rm load} = \langle u_{\rm in}\rangle A_{\rm eff}. \tag{5.39}$$

The quantity $A_{\rm eff}$ is called the *effective area* of the antenna; it is the area of the idealized antenna which absorbs as much net power from the incoming wave as the actual antenna. Alternatively, $A_{\rm eff}$ is the area of the incoming wavefront which is captured by the receiving antenna and fed to its load circuit. Thus,

$$P_{\rm load} = \frac{{E_0}^2 d^2}{32 R_{\rm rad}} = \frac{\epsilon_0 c\, {E_0}^2}{2}\, A_{\rm eff}, \tag{5.40}$$

giving

$$A_{\rm eff} = \frac{d^2}{16\,\epsilon_0 c\, R_{\rm rad}} = \frac{3}{8\pi}\,\lambda^2. \tag{5.41}$$

It is clear that the effective area of a Hertzian dipole antenna is of order the wavelength squared of the incoming radiation.

We can generalize from this analysis of a special case. The directivity of a Hertzian dipole is 3/2. Thus, the effective area of the *isotropic radiator* (the mythical reference antenna against which directivities are measured) is

$$A_0 = \frac{2}{3} A_{\rm Hertzian\ dipole} = \frac{\lambda^2}{4\pi}, \tag{5.42}$$

or

$$A_0 = \pi \lambda\!\!\!^{-}{}^2, \tag{5.43}$$

where $\lambda\!\!\!^{-} = \lambda/2\pi$. Here, we have used the formal definition of the effective area of an antenna: $A_{\rm eff}$ is that area which, when multiplied by the time-averaged Poynting flux of the incoming wave, equals the maximum power received by the antenna (when its orientation is optimal). Clearly, the effective area of an isotropic radiator is the same as the area of a circle whose radius is the reduced wavelength $\lambda\!\!\!^{-}$.

We can take yet one more step and conclude that the effective area of any antenna of directivity G is

$$A_{\rm eff} = G\,\pi\lambda\!\!\!^{-2}. \tag{5.44}$$

Of course, to realize this full capture area the antenna must be orientated properly.

Let us calculated the coupling or *insertion loss* of an antenna-to-antenna communications link. Suppose that a generator delivers the power $P_{\rm in}$ to a transmitting antenna, which is aimed at a receiving antenna a distance r away. The receiving antenna (properly aimed) then captures and delivers the power $P_{\rm out}$ to its load circuit. From the definition of directivity, the transmitting antenna produces the time-averaged Poynting flux

$$\langle u\rangle = G_t\,\frac{P_{\rm in}}{4\pi r^2} \tag{5.45}$$

at the receiving antenna. The received power is

$$P_{\rm out} = \langle u\rangle\,G_r\,A_0. \tag{5.46}$$

Here, G_t is the gain of the transmitting antenna, and G_r is the gain of the receiving antenna. Thus,

$$\frac{P_{\rm out}}{P_{\rm in}} = G_t G_r\left(\frac{\lambda}{4\pi r}\right)^2 = \frac{A_t A_r}{\lambda^2 r^2}, \tag{5.47}$$

where A_t and A_r are the effective areas of the transmitting and receiving antennas, respectively. This result is known as the *Friis transmission formula*. Note that it depends on the product of the gains of the two antennas. Thus, a properly aligned communications link has the same insertion loss operating in either direction.

A thin wire linear antenna might appear to be essentially one dimensional. However, the concept of an effective area shows that it possesses a second dimension determined by the wavelength. For instance, for a half-wave antenna, the gain of which is 1.64, the effective area is

$$A_{\rm eff} = 1.64\,A_0 = \frac{\lambda}{2}\,(0.26\,\lambda). \tag{5.48}$$

Thus, we can visualize the capture area as a rectangle which is the physical length of the antenna in one direction, and approximately one quarter of the wavelength in the other.

5.3 Antenna arrays

Consider a linear array of N half-wave antennas arranged along the x-axis with a uniform spacing Δ. Suppose that each antenna is aligned along the z-axis, and also that all antennas are driven *in phase*. Let one end of the array coincide with the origin. The field produced in the radiation zone by the end-most antenna is given by (see Eq. (5.15))

$$\boldsymbol{A}(\boldsymbol{r}) = \hat{\boldsymbol{z}}\,\frac{\mu_0\,I}{4\pi}\,\frac{2}{kr}\,\frac{\cos(\pi\cos\theta/2)}{\sin^2\theta}\,{\rm e}^{\,{\rm i}\,(kr-\omega t)}, \tag{5.49}$$

where I is the peak current flowing in each antenna. The fields produced at a given point in the radiation zone by successive elements of the array differ in phase by an amount $\alpha = k\Delta\,\sin\theta\,\cos\varphi$. Here, r, θ, φ are conventional spherical polar coordinates. Thus, the total field is given by

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{r}) &= \hat{\boldsymbol{z}}\,\frac{\mu_0\,I}{4\pi}\,\frac{2}{kr}\,\frac{\cos(\pi\cos\theta/2)}{\sin^2\theta}\\ &\quad\times\left[1+{\rm e}^{\,{\rm i}\,\alpha}+{\rm e}^{\,2{\rm i}\,\alpha}+\cdots+{\rm e}^{\,(N-1){\rm i}\,\alpha}\right]\,{\rm e}^{\,{\rm i}\,(kr-\omega t)}.\end{aligned} \tag{5.50}$$

The series in square brackets is a geometric progression in $\beta = \exp(\,{\rm i}\,\alpha)$, the sum of which is

$$1+\beta+\beta^2+\cdots\beta^{N-1} = \frac{\beta^N-1}{\beta-1}. \tag{5.51}$$

Thus, the term in square brackets becomes

$$\frac{{\rm e}^{\,{\rm i}\,N\alpha}-1}{{\rm e}^{\,{\rm i}\,\alpha}-1} = {\rm e}^{\,{\rm i}\,(N-1)\alpha/2}\,\frac{\sin(N\alpha/2)}{\sin(\alpha/2)}. \tag{5.52}$$

It follows from Eq. (5.16) that the radiation pattern due to the array takes the form

$$\frac{dP}{d\Omega} = \left(\frac{\mu_0 c\, I^2}{8\pi^2}\frac{\cos^2(\pi\cos\theta/2)}{\sin^2\theta}\right)\left(\frac{\sin^2(N\alpha/2)}{\sin^2(\alpha/2)}\right). \tag{5.53}$$

We can think of this formula as the product of the two factors in large parentheses. The first is just the standard radiation pattern of a half-wave antenna. The second arises from the linear array of N elements. If we retained the same array, but replaced the elements by something other than half-wave antennas, then the first factor would change, but not the second. If we changed the array, but not the elements, then the second factor would change but the first would remain the same. Thus, we can think of the radiation pattern as the product of two independent factors, the *element function* and the *array function*. This independence follows from the Fraunhofer approximation (5.6), which justifies the linear phase shifts of Eq. (5.7).

The array function in this case is

$$f(\alpha) = \frac{\sin^2(N\alpha/2)}{\sin^2(\alpha/2)}, \tag{5.54}$$

where

$$\alpha = k\Delta\,\sin\theta\,\cos\varphi. \tag{5.55}$$

The function $f(\alpha)$ has nulls whenever the numerator vanishes; that is, whenever

$$\pm\alpha = \frac{2\pi}{N}, \frac{4\pi}{N}, \cdots \frac{(N-1)\,2\pi}{N}; \frac{(N+1)\,2\pi}{N}\cdots. \tag{5.56}$$

However, when $\pm\alpha = 0, 2\pi, \cdots$, the denominator also vanishes, and the l'Hôpital limit is easily seen to be $f(0, 2\pi, \cdots) \to N^2$. These limits are known as the *principle maxima* of the function. Secondary maxima occur approximately at the maxima of the numerator; that is, at

$$\pm\alpha = \frac{3\pi}{N}, \frac{5\pi}{N}, \cdots \frac{(2N-3)\,2\pi}{N}; \frac{(2N+3)\,2\pi}{N}\cdots. \tag{5.57}$$

There are $(N-2)$ secondary maxima between successive principal maxima.

Now, the maximum possible value of α is $k\Delta = 2\pi\Delta/\lambda$. Thus, when the element spacing Δ is less than the wavelength there is only one principle maximum (at $\alpha = 0$), directed perpendicular to the array (*i.e.*, at $\varphi = \pm\pi/2$). Such a system is called a *broadside array.* The secondary maxima of the radiation pattern are called *side lobes.* In the direction perpendicular to the array, all elements contribute in phase, and the intensity is proportional to the square of the sum of the individual amplitudes. Thus, the peak intensity for an N element array is N^2 times the intensity of a single antenna. The angular half-width of the principle maximum (in φ) is approximately $\Delta\varphi \simeq \lambda/N\Delta$. Although the principle lobe clearly gets narrower in the azimuthal angle φ as N increases, the lobe width in the polar angle θ is mainly controlled by the element function, and is thus little affected by the number of elements. A radiation pattern which is narrow in one angular dimension, but broad in the other, is called a *fan beam.*

Arranging a set of antennas in a regular array has the effect of taking the azimuthally symmetric radiation pattern of an individual antenna and concentrating it into some narrow region of azimuthal angle of extent $\Delta\varphi \simeq \lambda/N\Delta$. The net result is that the gain of the array is larger than that of an individual antenna by a factor of order

$$\frac{2\pi\, N\Delta}{\lambda}. \tag{5.58}$$

It is clear that the boost factor is of order the linear extent of the array divided by the wavelength of the emitted radiation. Thus, it is possible to construct a very high gain antenna by arranging a large number of low gain antennas in a regular pattern and driving them in phase. The optimum spacing between successive elements of the array is of order the wavelength of the radiation.

A linear array of antenna elements which are spaced $\Delta = \lambda/2$ apart and driven with alternating phases has its principle radiation maximum along $\varphi = 0$ and π, since the field amplitudes now add in phase in the plane of the array. Such a system is called an *end-fire array.* The direction of the principle maximum can be changed at will by introducing the appropriate phase shift between successive elements of the array. In fact, it is possible to produce a radar beam which sweeps around the horizon, without any mechanical motion of the array, by varying the phase difference between successive elements of the array electronically.

5.4 Thomson scattering

When an electromagnetic wave is incident on a charged particle, the electric and magnetic components of the wave exert a Lorentz force on the particle, setting it into motion. Since the wave is periodic in time, so is the motion of the particle. Thus, the particle is accelerated and, consequently, emits radiation. More exactly, energy is absorbed from the incident wave by the particle and re-emitted as electromagnetic radiation. Such a process is clearly equivalent to the scattering of the electromagnetic wave by the particle.

Consider a linearly polarized, monochromatic, plane wave incident on a particle carrying a charge q. The electric component of the wave can be written

$$\boldsymbol{E} = \boldsymbol{e}\, E_0\, \mathrm{e}^{\,\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \tag{5.59}$$

where E_0 is the peak amplitude of the electric field, $\boldsymbol{e}$ is the polarization vector, and $\boldsymbol{k}$ is the wave vector (of course, $\boldsymbol{e}\cdot\boldsymbol{k} = 0$). The particle is assumed to undergo small amplitude oscillations about an equilibrium position which coincides with the origin of the coordinate system. Furthermore, the particle's velocity is assumed to remain sub-relativistic, which enables us to neglect the magnetic component of the Lorentz force. The equation of motion of the charged particle is approximately

$$\boldsymbol{f} = q\boldsymbol{E} = m\ddot{\boldsymbol{s}}, \tag{5.60}$$

where m is the mass of the particle, $\boldsymbol{s}$ is its displacement from the origin, and $\dot{}$ denotes $\partial/\partial t$. According to Eq. (2.321), the time-averaged power radiated per unit solid angle by an accelerating, non-relativistic, charged particle is given by

$$\frac{dP}{d\Omega} = \frac{q^2\langle\ddot{s}^2\rangle}{16\pi^2\epsilon_0 c^3}\,\sin^2\theta, \tag{5.61}$$

where $\langle\cdots\rangle$ denotes a time average. However,

$$\langle\ddot{s}^2\rangle = \frac{q^2}{m^2}\langle E^2\rangle = \frac{q^2 {E_0}^2}{2m^2}. \tag{5.62}$$

Hence, the scattered power per unit solid angle is given by

$$\frac{dP}{d\Omega} = \left(\frac{q^2}{4\pi\epsilon_0\, mc^2}\right)^2 \frac{\epsilon_0 c\, {E_0}^2}{2}\sin^2\theta. \tag{5.63}$$

The time-averaged Poynting flux of the incident wave is

$$\langle u\rangle = \frac{\epsilon_0 c\, E_0{}^2}{2}. \tag{5.64}$$

It is convenient to define the *scattering cross section* as the equivalent area of the incident wavefront which delivers the same power as that re-radiated by the particle:

$$\sigma = \frac{\text{total re-radiated power}}{\langle u\rangle}. \tag{5.65}$$

By analogy, the *differential scattering cross section* is defined

$$\frac{d\sigma}{d\Omega} = \frac{dP/d\Omega}{\langle u\rangle}. \tag{5.66}$$

It follows from Eqs. (5.63), (5.64), and (5.66) that

$$\frac{d\sigma}{d\Omega} = \left(\frac{q^2}{4\pi\epsilon_0\, mc^2}\right)^2 \sin^2\theta. \tag{5.67}$$

The total scattering cross section is then

$$\sigma = \int_0^\pi \frac{d\sigma}{d\Omega}\, 2\pi\, \sin\theta\, d\theta = \frac{8\pi}{3}\left(\frac{q^2}{4\pi\epsilon_0\, mc^2}\right)^2. \tag{5.68}$$

The quantity θ appearing in Eq. (5.67) is the angle subtended between the direction of acceleration of the particle and the direction of the outgoing radiation (which is parallel to the unit vector $\boldsymbol{n}$). In the present case, the acceleration is due to the electric field, so it is parallel to the polarization vector $\boldsymbol{e}$. Thus, $\cos\theta = \boldsymbol{e}\cdot\boldsymbol{n}$.

Up to now, we have only considered the scattering of linearly polarized radiation by a charged particle. Let us now calculate the angular distribution of scattered radiation for the commonly occurring case of randomly polarized incident radiation. It is helpful to set up a right-handed coordinate system based on the three mutually orthogonal unit vectors $\boldsymbol{e}$, $\boldsymbol{e}\wedge\hat{\boldsymbol{k}}$, and $\hat{\boldsymbol{k}}$. In terms of these unit vectors, we can write

$$\boldsymbol{n} = \sin\varphi\cos\psi\, \boldsymbol{e} + \sin\varphi\sin\psi\, \boldsymbol{e}\wedge\hat{\boldsymbol{k}} + \cos\varphi\, \hat{\boldsymbol{k}}, \tag{5.69}$$

where φ is the angle subtended between the direction of the incident radiation and that of the scattered radiation, and ψ is an angle which specifies the orientation of the polarization vector in the plane perpendicular to $\boldsymbol{k}$ (assuming that $\boldsymbol{n}$ is known). It is easily seen that

$$\cos\theta = \boldsymbol{e}\cdot\boldsymbol{n} = \cos\psi\,\sin\varphi, \tag{5.70}$$

so

$$\sin^2\theta = 1 - \cos^2\psi\,\sin^2\varphi. \tag{5.71}$$

Averaging this result over all possible polarizations of the incident wave (*i.e.*, over all possible values of the polarization angle ψ), we obtain

$$\overline{\sin^2\theta} = 1 - \overline{\cos^2\psi}\,\sin^2\varphi = 1 - (\sin^2\varphi)/2 = \frac{1+\cos^2\varphi}{2}. \tag{5.72}$$

Thus, the differential scattering cross section for unpolarized incident radiation (obtained by substituting $\overline{\sin^2\theta}$ for $\sin^2\theta$ in Eq. (5.67)) is given by

$$\left(\frac{d\sigma}{d\Omega}\right)_{\rm unpolarized} = \left(\frac{q^2}{4\pi\epsilon_0\,mc^2}\right)^2\frac{1+\cos^2\varphi}{2}. \tag{5.73}$$

It is clear that the differential scattering cross section is independent of the frequency of the incident wave, and is also symmetric with respect to forward and backward scattering. The frequency of the scattered radiation is the same as that of the incident radiation. The total scattering cross section is obtained by integrating over the entire solid angle of the polar angle φ and the azimuthal angle ψ. Not surprisingly, the result is exactly the same as Eq. (5.68).

The classical scattering cross section (5.73) is modified by quantum effects when the energy of the incident photons, $\hbar\omega$, becomes comparable with the rest mass of the scattering particle, mc^2. The scattering of a photon by a charged particle is called *Compton scattering*, and the quantum mechanical version of the Compton scattering cross section is known as the *Klein-Nishina formula.* As the photon energy increases, and eventually becomes comparable with the rest mass energy of the particle, the Klein-Nishina formula predicts that forward scattering of photons becomes increasingly favored with respect to backward scattering. The Klein-Nishina cross section *does*, in general, depend on the frequency of the incident photons. Furthermore, energy and momentum conservation demand a shift in the frequency of scattered photons with respect to that of the incident photons.

If the charged particle in question is an electron then Eq. (5.68) reduces to the well known *Thomson scattering cross section*

$$\sigma_{\rm Thomson} = \frac{8\pi}{3}\left(\frac{e^2}{4\pi\epsilon_0\, m_e c^2}\right)^2 = 6.65\times 10^{-29}\,{\rm m}^2. \tag{5.74}$$

The quantity $e^2/(4\pi\epsilon_0\, m_e c^2) = 2.8\times 10^{-15}$ m is called the *classical electron radius* (it is the radius of spherical shell of total charge e whose electrostatic energy equals the rest mass energy of the electron). Thus, as a scatterer the electron acts rather like a solid sphere whose radius is of order the classical electron radius. Since this radius is extremely small, it is clear that scattering of radiation by a single electron (or any other charged particle) is a very weak process.

5.5 Rayleigh scattering

Let us now consider the scattering of electromagnetic radiation by a harmonically bound electron; *e.g.*, an electron orbiting an atomic nucleus. We have seen in Section 4.2 that such an electron satisfies an equation of motion of the form

$$\ddot{\boldsymbol{s}} + \gamma_0\,\dot{\boldsymbol{s}} + {\omega_0}^2\boldsymbol{s} = -\frac{e}{m_e}\,\boldsymbol{E}, \tag{5.75}$$

where ω_0 is the characteristic oscillation frequency of the electron, and $\gamma_0 \ll \omega_0$ is the damping rate of such oscillations. Assuming an ${\rm e}^{-{\rm i}\,\omega t}$ time dependence of both $\boldsymbol{s}$ and $\boldsymbol{E}$, we find that

$$\ddot{\boldsymbol{s}} = \frac{\omega^2}{{\omega_0}^2 - \omega^2 - {\rm i}\,\gamma_0\,\omega}\,\frac{e}{m_e}\,\boldsymbol{E}. \tag{5.76}$$

It follows, by analogy with the analysis in the previous section, that the total scattering cross section is given by

$$\sigma = \sigma_{\rm Thomson}\,\frac{\omega^4}{({\omega_0}^2-\omega^2)^2 + (\gamma_0\,\omega)^2}. \tag{5.77}$$

The angular distribution of the radiation is the same as that in the case of a free electron.

The maximum value of the cross section (5.77) is obtained when $\omega \simeq \omega_0$; *i.e.*, for resonant scattering. In this case, the scattering cross section can become very large. In fact,

$$\sigma \simeq \sigma_{\rm Thomson} \left(\frac{\omega_0}{\gamma_0}\right)^2, \tag{5.78}$$

which is generally far greater than the Thomson scattering cross section.

For strong binding, $\omega \ll \omega_0$, Eq. (5.77) reduces to

$$\sigma \simeq \sigma_{\rm Thomson} \left(\frac{\omega}{\omega_0}\right)^4, \tag{5.79}$$

giving a scattering cross section which depends on the inverse fourth power of the wavelength of the incident radiation. Equation (5.79) is known as the *Rayleigh scattering cross section*, and is appropriate to the scattering of visible radiation by gas molecules. This is Rayleigh's famous explanation of the blue sky: the air molecules of the atmosphere preferentially scatter the shorter wavelength components out of "white" sunlight which grazes the atmosphere. Conversely, sunlight viewed directly through the long atmospheric path at sunset appears reddened. The Rayleigh scattering cross section is much less than the Thompson scattering cross section (for $\omega \ll \omega_0$). However, this effect is offset to some extent by the fact that the density of neutral molecules in a gas (*e.g.*, the atmosphere) is much larger than the density of free electrons typically encountered in a plasma.

6 Resonant cavities and wave guides

6.1 Introduction

Let us investigate the solution of the homogeneous wave equation in regions containing various geometric boundaries, particularly in regions bounded by conductors. The boundary value problem is of great theoretical significance and also has many practical electromagnetic applications, particularly in the microwave region of the spectrum.

6.2 Boundary conditions

Let us review the general boundary conditions on the field vectors at a surface between medium 1 and medium 2:

$$\boldsymbol{n}\cdot(\boldsymbol{D}_1 - \boldsymbol{D}_2) = \tau, \tag{6.1a}$$

$$\boldsymbol{n}\wedge(\boldsymbol{E}_1 - \boldsymbol{E}_2) = 0, \tag{6.1b}$$

$$\boldsymbol{n}\cdot(\boldsymbol{B}_1 - \boldsymbol{B}_2) = 0, \tag{6.1c}$$

$$\boldsymbol{n}\wedge(\boldsymbol{H}_1 - \boldsymbol{H}_2) = \boldsymbol{K}, \tag{6.1d}$$

where τ is used for the surface change density (to avoid confusion with the conductivity), and $\boldsymbol{K}$ is the surface current density. Here, $\boldsymbol{n}$ is a unit vector normal to the surface, directed from medium 2 to medium 1. We have seen in Section 4.4 that for normal incidence an electromagnetic wave falls off very rapidly inside the surface of a good conductor. Equation (4.35) implies that in the limit of perfect conductivity ($\sigma \to \infty$) the tangential component of $\boldsymbol{E}$ vanishes, whereas that of $\boldsymbol{H}$ may remain finite. Let us examine the behaviour of the normal components.

Let medium 1 be a good conductor for which $\sigma/\epsilon\epsilon_0\omega \gg 1$, whilst medium 2 is a perfect insulator. The surface change density is related to the currents flowing inside the conductor. In fact, the conservation of charge requires that

$$\boldsymbol{n}\cdot\boldsymbol{j} = \frac{\partial\tau}{\partial t} = -\mathrm{i}\,\omega\,\tau. \tag{6.2}$$

However, $\boldsymbol{n}\cdot\boldsymbol{j} = \boldsymbol{n}\cdot\sigma\boldsymbol{E}_1$, so it follows from Eq. (6.1)(a) that

$$\left(1 + \frac{\mathrm{i}\,\omega\epsilon_0\epsilon_1}{\sigma}\right)\boldsymbol{n}\cdot\boldsymbol{E}_1 = \frac{\mathrm{i}\,\omega\epsilon_0\epsilon_2}{\sigma}\,\boldsymbol{n}\cdot\boldsymbol{E}_2. \tag{6.3}$$

It is clear that the normal component of $\boldsymbol{E}$ within the conductor also becomes vanishingly small as the conductivity approaches infinity.

If $\boldsymbol{E}$ vanishes inside a perfect conductor then the curl of $\boldsymbol{E}$ also vanishes, and the time rate of change of $\boldsymbol{B}$ is correspondingly zero. This implies that there are no oscillatory fields whatever inside such a conductor, and that the boundary values of the fields outside are given by

$$\boldsymbol{n}\cdot\boldsymbol{D} = -\tau, \tag{6.4a}$$

$$\boldsymbol{n}\wedge\boldsymbol{E} = 0, \tag{6.4b}$$

$$\boldsymbol{n}\cdot\boldsymbol{B} = 0, \tag{6.4c}$$

$$\boldsymbol{n}\wedge\boldsymbol{H} = -\boldsymbol{K}. \tag{6.4d}$$

Here, $\boldsymbol{n}$ is a unit normal at the surface of the conductor pointing *into* the conductor. Thus, the electric field is normal and the magnetic field tangential at the surface of a perfect conductor. For good conductors these boundary conditions yield excellent representations of the geometrical configurations of external fields, but they lead to the neglect of some important features of real fields, such as losses in cavities and signal attenuation in wave guides.

In order to estimate such losses it is useful to see how the tangential and normal fields compare when σ is large but finite. Equations (4.5) and (4.34) yield

$$\boldsymbol{H} = \frac{1+\mathrm{i}}{\sqrt{2}}\sqrt{\frac{\sigma}{\mu_0\omega}}\,\boldsymbol{n}\wedge\boldsymbol{E} \tag{6.5}$$

at the surface of a conductor (provided that the wave propagates into the conductor). Let us assume, without obtaining a complete solution, that a wave with $\boldsymbol{H}$ very nearly tangential and $\boldsymbol{E}$ very nearly normal is propagated along the surface of the metal. According to the Faraday-Maxwell equation

$$|H_{\parallel}| \simeq \frac{k}{\mu_0\omega}\,|E_{\perp}| \tag{6.6}$$

just outside the surface, where k is the component of the propagation vector along the surface. However, Eq. (6.5) implies that a tangential component of $\boldsymbol{H}$ is accompanied by a small tangential component of $\boldsymbol{E}$. By comparing these two expressions, we obtain

$$\frac{|E_{\|}|}{|E_{\perp}|} \simeq k\sqrt{\frac{2}{\mu_0\omega\sigma}} = \frac{d}{\lambda\!\!\!^{-}}, \tag{6.7}$$

where d is the skin depth (see Eq. (4.36)) and $\lambda\!\!\!^{-} \equiv 1/k$. It is clear that the ratio of the tangential component of $\boldsymbol{E}$ to its normal component is of order the skin depth divided by the wavelength. It is readily demonstrated that the ratio of the normal component of $\boldsymbol{H}$ to its tangential component is of this same magnitude. Thus, we can see that in the limit of high conductivity, which means vanishing skin depth, no fields penetrate the conductor, and the boundary conditions are those given by Eqs. (6.4). Let us investigate the solution of the homogeneous wave equation subject to such boundary conditions.

6.3 Cavities with rectangular boundaries

Consider a vacuum region totally enclosed by rectangular conducting walls. In this case, all of the field components satisfy the wave equation

$$\nabla^2\psi - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = 0, \tag{6.8}$$

where ψ represents any component of $\boldsymbol{E}$ or $\boldsymbol{H}$. The boundary conditions (6.4) require that the electric field is normal to the walls at the boundary whereas the magnetic field is tangential. If a, b, and c are the dimensions of the cavity, then it is readily verified that the electric field components are

$$E_x = E_1\cos(k_1 x)\sin(k_2 y)\sin(k_3 z)\,\mathrm{e}^{-\mathrm{i}\,\omega t}, \tag{6.9a}$$

$$E_y = E_2\sin(k_1 x)\cos(k_2 y)\sin(k_3 z)\,\mathrm{e}^{-\mathrm{i}\,\omega t}, \tag{6.9b}$$

$$E_z = E_3\sin(k_1 x)\sin(k_2 y)\cos(k_3 z)\,\mathrm{e}^{-\mathrm{i}\,\omega t}, \tag{6.9c}$$

where

$$k_1 = \frac{l\,\pi}{a}, \tag{6.10a}$$

$$k_2 = \frac{m\,\pi}{b}, \tag{6.10b}$$

$$k_3 = \frac{n\,\pi}{c}, \tag{6.10c}$$

with l, m, n integers. The allowed frequencies are given by

$$\frac{\omega^2}{c^2} = \pi^2\left(\frac{l^2}{a^2} + \frac{m^2}{b^2} + \frac{n^2}{c^2}\right). \tag{6.11}$$

It is clear from Eq. (6.9) that at least two of the integers l, m, n must be different from zero in order to have non-vanishing fields. The magnetic fields obtained by the use of $\nabla\wedge\boldsymbol{E} = \mathrm{i}\,\omega\boldsymbol{B}$ automatically satisfy the appropriate boundary conditions, and are in phase quadrature with the electric fields. Thus, the sum of the total electric and magnetic energies within the cavity is constant, although the two terms oscillate separately.

The amplitudes of the electric field components are not independent, but are related by the divergence condition $\nabla\cdot\boldsymbol{E} = 0$, which yields

$$k_1\,E_1 + k_2\,E_2 + k_3\,E_3 = 0. \tag{6.12}$$

There are, in general, two linearly independent vectors $\boldsymbol{E}$ that satisfy this condition, corresponding to two polarizations. (The exception is the case that one of the integers l, m, n is zero, in which case $\boldsymbol{E}$ is fixed in direction.) Each vector is accompanied by a magnetic field at right angles. The fields corresponding to a given set of integers l, m, and n constitute a particular mode of vibration of the cavity. It is evident from standard Fourier theory that the different modes are *orthogonal* (*i.e.*, they are normal modes) and that they form a *complete set*. In other words, any general electric and magnetic fields which satisfy the boundary conditions (6.4) can be unambiguously decomposed into some linear combination of all of the various possible normal modes of the cavity. Since each normal mode oscillates at a specific frequency it is clear that if we are given the electric and magnetic fields inside the cavity at time $t = 0$ then the subsequent behaviour of the fields is *uniquely* determined for all time.

The conducting walls gradually absorb energy from the cavity, due to their finite resistivity, at a rate which can easily be calculated. For finite σ the small tangential component of $\boldsymbol{E}$ at the walls can be estimated using Eq. (6.5):

$$\boldsymbol{E}_\parallel = \frac{1-\mathrm{i}}{\sqrt{2}}\sqrt{\frac{\mu_0\omega}{\sigma}}\,\boldsymbol{H}_\parallel \wedge \boldsymbol{n}. \tag{6.13}$$

Now, the tangential component of $\boldsymbol{H}$ at the walls is slightly different from that given by the ideal solution. However, this is a small effect and can be neglected to leading order in σ^{-1}. The time averaged energy flux into the walls is given by

$$\overline{\boldsymbol{N}} = \frac{1}{2}\,\mathrm{Re}\,(\boldsymbol{E}_\parallel \wedge \boldsymbol{H}_\parallel) = \frac{1}{2}\sqrt{\frac{\mu_0\omega}{2\sigma}}\,H_{\parallel 0}^{\;2}\,\boldsymbol{n} = \frac{H_{\parallel 0}^{\;2}}{2\sigma d}\,\boldsymbol{n}, \tag{6.14}$$

where $H_{\parallel 0}$ is the peak value of the tangential magnetic field at the walls predicted by the ideal solution. According to the boundary condition (6.4)(d), $H_{\parallel 0}$ is equal to the peak value of the surface current density K_0. It is helpful to define a surface resistance,

$$\overline{\boldsymbol{N}} = \overline{K^2} R_s\,\boldsymbol{n}, \tag{6.15}$$

where

$$R_s = \frac{1}{\sigma d}. \tag{6.16}$$

This approach makes it clear that the dissipation of energy is due to ohmic heating in a thin layer, whose thickness is of order the skin depth, on the surface of the conducting walls.

6.4 The quality factor of a resonant cavity

The quality factor Q of a resonant cavity is defined

$$Q = 2\pi\,\frac{\text{energy stored in cavity}}{\text{energy lost per cycle to walls}}. \tag{6.17}$$

For a specific normal mode of the cavity this quantity is independent of the mode amplitude. By conservation of energy the power dissipated in ohmic losses is minus the rate of change of the stored energy U. We can write a differential equation for the behaviour of U as a function of time:

$$\frac{dU}{dt} = -\frac{\omega_0}{Q}\,U, \tag{6.18}$$

where ω_0 is the oscillation frequency of the normal mode in question. The solution to the above equation is

$$U(t) = U(0)\,\mathrm{e}^{-\omega_0 t/Q}. \tag{6.19}$$

This time dependence of the stored energy suggests that the oscillations of the fields in the cavity are damped as follows:

$$E(t) = E_0\,\mathrm{e}^{-\omega_0 t/2Q}\,\mathrm{e}^{-\mathrm{i}\,(\omega_0+\Delta\omega)t}, \tag{6.20}$$

where we have allowed for a shift $\Delta\omega$ of the resonant frequency as well as the damping. A damped oscillation such as this does not consist of a pure frequency. Instead, it is made up of a superposition of frequencies around $\omega = \omega_0 + \Delta\omega$. Standard Fourier analysis yields

$$E(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} E(\omega)\,\mathrm{e}^{-\mathrm{i}\,\omega t}\,d\omega, \tag{6.21}$$

where

$$E(\omega) = \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} E_0\,\mathrm{e}^{-\omega_0 t/2Q}\,\mathrm{e}^{\mathrm{i}\,(\omega-\omega_0-\Delta\omega)t}\,dt. \tag{6.22}$$

It follows that

$$|E(\omega)|^2 \propto \frac{1}{(\omega-\omega_0-\Delta\omega)^2 + (\omega_0/2Q)^2}. \tag{6.23}$$

The resonance shape has a full width Γ at half-maximum equal to ω_0/Q. For a constant input voltage, the energy of oscillation in the cavity as a function of frequency follows the resonance curve in the neighbourhood of a particular resonant frequency. It can be seen that the ohmic losses, which determine Q for a particular mode, also determine the maximum amplitude of the oscillation when the resonance condition is exactly satisfied, as well as the width of the resonance (*i.e.*, how far off the resonant frequency the system can be driven and still yield a significant oscillation amplitude).

6.5 Axially symmetric cavities

The rectangular cavity which we have just considered has many features in common with axially symmetric cavities of arbitrary cross section. In every cavity the allowed values of the wave vector $\boldsymbol{k}$, and thus the allowed frequencies, are determined by the geometry of the cavity. We have seen that for each set of k_1, k_2, k_3 in a rectangular cavity there are, in general, two linearly independent modes; *i.e.*, the polarization remains arbitrary. We can take advantage of this fact to classify modes into two kinds according to the orientation of the field vectors. Let us choose one type of mode such that the electric field vector lies in the cross-sectional plane, and the other so that the magnetic field vector lies in this plane. This classification into transverse electric (TE) and transverse magnetic (TM) modes turns out to be possible for all axially symmetric cavities, although the rectangular cavity is unique in having one mode of each kind corresponding to each allowed frequency.

Suppose that the direction of symmetry is along the z-axis, and that the length of the cavity in this direction is L. The boundary conditions at $z = 0$ and $z = L$ demand that the z dependence of wave quantities be either $\sin k_3 z$ or $\cos k_3 z$, where $k_3 = n\pi/L$. In other words, every field component satisfies

$$\left(\frac{\partial^2}{\partial z^2} + {k_3}^2\right)\psi = 0, \tag{6.24}$$

as well as

$$(\nabla^2 + k^2)\psi = 0, \tag{6.25}$$

where ψ stands for any component of $\boldsymbol{E}$ or $\boldsymbol{H}$. The field equations

$$\begin{aligned} \nabla \wedge \boldsymbol{E} &= \mathrm{i}\,\omega\mu_0\,\boldsymbol{H}, && \text{(6.26a)}\\ \nabla \wedge \boldsymbol{H} &= -\mathrm{i}\,\omega\epsilon_0\,\boldsymbol{E} && \text{(6.26b)} \end{aligned}$$

must also be satisfied.

Let us write each vector and each operator in the above equations as the sum of a transverse part, designated by the subscript s, and a component along z. We find that for the transverse fields

$$\begin{aligned} \mathrm{i}\,\omega\mu_0\,\boldsymbol{H}_s &= \nabla_s \wedge \boldsymbol{E}_z + \nabla_z \wedge \boldsymbol{E}_s, && \text{(6.27a)}\\ -\mathrm{i}\,\omega\epsilon_0\,\boldsymbol{E}_s &= \nabla_s \wedge \boldsymbol{H}_z + \nabla_z \wedge \boldsymbol{H}_s. && \text{(6.27b)} \end{aligned}$$

When one side of Eqs. (6.27) is substituted for the transverse field on the right-hand side of the other, and use is made of Eq. (6.24), we obtain

$$\boldsymbol{E}_s = \frac{\nabla_s(\partial E_z/\partial z)}{k^2 - {k_3}^2} + \frac{\mathrm{i}\,\omega\mu_0}{k^2 - {k_3}^2}\,\nabla_s \wedge \boldsymbol{H}_z, \tag{6.28a}$$

$$\boldsymbol{H}_s = \frac{\nabla_s(\partial H_z/\partial z)}{k^2 - {k_3}^2} - \frac{\mathrm{i}\,\omega\epsilon_0}{k^2 - {k_3}^2}\,\nabla_s \wedge \boldsymbol{E}_z. \tag{6.28b}$$

Thus, all transverse fields can be expressed in terms of the z components of the fields, each of which satisfies the differential equation

$$\left[\nabla_s^2 + (k^2 - {k_3}^2)\right] A_z = 0, \tag{6.29}$$

where A_z stands for either E_z or H_z, and ∇_s^2 is the two-dimensional Laplacian operator.

The conditions on E_z and H_z at the boundary (in the transverse plane) are quite different: E_z must vanish on the boundary, whereas the normal derivative of H_z must vanish so that $\boldsymbol{H}_s$ in Eq. (6.28)(b) satisfies the appropriate boundary condition. When the cross section is a rectangle, these two conditions lead to the same eigenvalues of $(k^2 - {k_3}^2) = {k_s}^2 = {k_1}^2 + {k_2}^2$, as we have seen. Otherwise, they correspond to two *different* frequencies, one for which E_z is permitted but $H_z = 0$, and the other where the opposite is true. In every case, it is possible to classify the modes as transverse magnetic or transverse electric. Thus, the field components E_z and H_z play the role of independent potentials, from which the other field components of the TE and TM modes, respectively, can be derived using Eqs. (6.28).

The mode frequencies are determined by the eigenvalues of Eqs. (6.24) and (6.29). If we denote the functional dependence of E_z or H_z on the plane cross section coordinates by $f(x, y)$, then we can write Eq. (6.29) as

$$\nabla_s^2 f = -{k_s}^2 f. \tag{6.30}$$

Let us first show that ${k_s}^2 > 0$, and hence that $k > k_3$. Now,

$$f\,\nabla_s^2 f = \nabla_s\cdot(f\,\nabla_s f) - (\nabla_s f)^2. \tag{6.31}$$

It follows that

$$-{k_s}^2 \int f^2\, dV + \int (\nabla_s f)^2\, dV = \int f \nabla f \cdot d\boldsymbol{S}, \tag{6.32}$$

where the integration is over the transverse cross section. If either f or its normal derivative is to vanish on S, the conducting surface, then

$$ {k_s}^2 = \frac{\int (\nabla_s f)^2\, dV}{\int f^2\, dV} > 0. \tag{6.33}$$

We have already seen that $k_3 = n\pi/L$. The allowed values of k_s depend both on the geometry of the cross section and the nature of the mode.

For TM modes $H_z = 0$, and the z dependence of E_z is given by $\cos(n\pi\, z/L)$. Equation (6.30) must be solved subject to the condition that f vanish on the boundaries of the plane cross section, thus completing the determination of E_z and k. The transverse fields are special cases of Eqs. (6.28):

$$\boldsymbol{E}_s = \frac{1}{{k_s}^2} \nabla_s \frac{\partial E_z}{\partial z}, \tag{6.34a}$$

$$\boldsymbol{H} = \frac{\mathrm{i}\,\omega\epsilon_0}{{k_s}^2}\, \hat{\boldsymbol{z}} \wedge \nabla_s E_z. \tag{6.34b}$$

For TE modes, in which $E_z = 0$, the condition that H_z vanish at the ends of the cylinder demands the use of $\sin(n\pi\, z/L)$, and k_s must be such that the normal derivative of H_z is zero at the walls. Equations (6.28), giving the transverse fields, then become

$$\boldsymbol{H}_s = \frac{1}{{k_s}^2} \nabla_s \frac{\partial H_z}{\partial z}, \tag{6.35a}$$

$$\boldsymbol{E} = -\frac{\mathrm{i}\,\omega\mu_0}{{k_s}^2}\, \hat{\boldsymbol{z}} \wedge \nabla_s H_z, \tag{6.35b}$$

and the mode determination is completed.

6.6 Cylindrical cavities

Let us apply the methods of the previous section to the TM modes of a right circular cylinder of radius a. We can write

$$E_z = A f(r,\varphi)\cos(k_3 z)\,\mathrm{e}^{-\mathrm{i}\,\omega t}, \tag{6.36}$$

where $f(r,\varphi)$ satisfies the equation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\,\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial\varphi^2} + {k_s}^2 f = 0, \tag{6.37}$$

and (r,φ,z) are cylindrical polar coordinates. Let

$$f(r,\varphi) = g(r)\,\mathrm{e}^{\mathrm{i}\,m\varphi}. \tag{6.38}$$

It follows that

$$\frac{1}{r}\frac{d}{dr}\left(r\,\frac{dg}{dr}\right) + \left({k_s}^2 - \frac{m^2}{r^2}\right) g = 0, \tag{6.39}$$

or

$$z^2\,\frac{d^2 g}{dz^2} + z\,\frac{dg}{dz} + (z^2 - m^2)\,g = 0, \tag{6.40}$$

where $z = k_s r$. The above equation is known as *Bessel's equation*. The two linearly independent solutions of Bessel's equation are denoted $J_m(z)$ and $Y_m(z)$. In the limit $|z| \ll 1$ these solutions behave as z^m and z^{-m}, respectively, to lowest order . More exactly[16]

$$J_m(z) = \left(\frac{z}{2}\right)^m \sum_{k=0}^{\infty} \frac{(-z^2/4)^k}{k!(m+k)!}, \tag{6.41a}$$

$$\begin{aligned} Y_m(z) = &-\frac{(z/2)^{-m}}{\pi}\sum_{k=0}^{m-1}\frac{(m-k-1)!\,(z^2/4)^k}{k!} + \frac{2}{\pi}\,\ln(z/2)\,J_m(z) \\ &-\frac{(z/2)^m}{\pi}\sum_{k=0}^{\infty}\left[\psi(k+1)+\psi(m+k+1)\right]\frac{(-z^2/4)^k}{k!(m+k)!} \end{aligned} \tag{6.41b}$$

[16] M. Abramowitz, and I.A. Stegun, *Handbook of mathematical functions*, (Dover, New York, 1965), Cha. 9

for $|z| \ll 1$, where

$$\psi(1) = -\gamma, \tag{6.42a}$$

$$\psi(n) = -\gamma + \sum_{k=1}^{n-1} k^{-1}, \tag{6.42b}$$

and $\gamma = \sum_{k=1}^{\infty} k^{-1} = 0.57722$ is Euler's constant. Clearly, the J_m are well behaved in the limit $|z| \to 0$, whereas the Y_m are badly behaved.

The asymptotic behaviour of both solutions at large $|z|$ is

$$J_m(z) = \sqrt{\frac{2}{\pi z}} \cos(z - m\,\pi/2 - \pi/4) + O(1/z), \tag{6.43a}$$

$$Y_m(z) = \sqrt{\frac{2}{\pi z}} \sin(z - m\,\pi/2 - \pi/4) + O(1/z). \tag{6.43b}$$

Thus, for $|z| \gg 1$ the solutions take the form of gradually decaying oscillations which are in phase quadrature. The behaviour of $J_0(z)$ and $Y_0(z)$ is shown in Fig. 21.

Since the axis $r = 0$ is included in the cavity the radial eigenfunction must be regular at the origin. This immediately rules out the $Y_m(k_s r)$ solutions. Thus, the most general solution for a TM mode is

$$E_z = A\, J_m(k_l r)\, \mathrm{e}^{\mathrm{i}\, m\varphi} \cos(k_3 z)\, \mathrm{e}^{-\mathrm{i}\,\omega t}. \tag{6.44}$$

The k_l are the eigenvalues of k_s, and are determined by the solutions of

$$J_m(k_l a) = 0. \tag{6.45}$$

The above constraint ensures that the tangential electric field is zero on the conducting walls surrounding the cavity ($r = a$).

The most general solution for a TE mode is

$$H_z = A\, J_m(k_l r)\, \mathrm{e}^{\mathrm{i}\, m\varphi} \sin(k_3 z)\, \mathrm{e}^{-\mathrm{i}\,\omega t}. \tag{6.46}$$

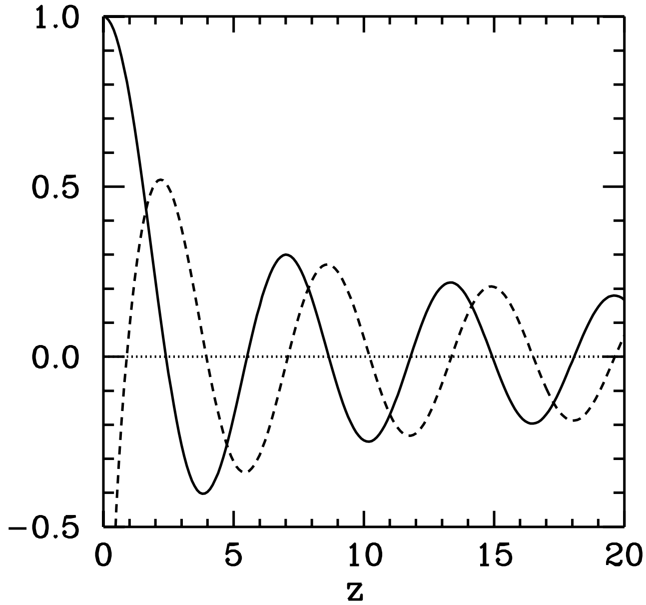


Figure 21: The Bessel functions $J_0(z)$ (solid line) and $Y_0(z)$ (dotted line)

In this case, the k_l are determined by the solution of

$$J_m'(k_l a) = 0, \tag{6.47}$$

where $'$ denotes differentiation with respect to the argument. The above constraint ensures that the normal magnetic field is zero on the conducting walls surrounding the cavity. The oscillation frequency of both the TM and TE modes is given by

$$\frac{\omega^2}{c^2} = k^2 = {k_l}^2 + \frac{n^2\pi^2}{L^2}. \tag{6.48}$$

If l is the ordinal number of a zero of a particular Bessel function of order m (l increases with increasing values of the argument), then each mode is characterized by three integers, l, m, n, as in the rectangular case. The lth zero of J_m is conventionally denoted $j_{m,l}$ [so, $J_m(j_{m,l}) = 0$]. Likewise, the lth zero of J_m' is denoted $j_{m,l}'$. Table 2 shows the first few zeros of J_0, J_0', J_1, and J_1'. It is clear that for fixed n and m the lowest frequency mode (i.e., the mode with the lowest value of k_l) is a TE mode. The mode with the next highest frequency is also a TE mode. The next highest frequency mode is a TM mode, and so on.

l	$j_{0,l}$	$j_{1,l}$	$j'_{0,l}$	$j'_{1,l}$
1	2.4048	3.8317	0.0000	1.8412
2	5.5201	7.0156	3.8317	5.3314
3	8.6537	10.173	7.0156	8.5363
4	11.792	13.324	10.173	11.706

Table 2: The first few values of $j_{0,l}$, $j_{1,l}$, $j'_{0,l}$, and $j'_{1,l}$

6.7 Wave guides

Let us consider the transmission of electromagnetic waves along the axis of a wave guide, which is simply a long, axially symmetric, hollow conductor with open ends. In order to represent a wave propagating along the z-direction, we can write the dependence of the fields on the coordinate variables and the time as

$$f(x,y)\,\mathrm{e}^{\,\mathrm{i}\,(k_g z-\omega t)}. \tag{6.49}$$

The *guide propagation constant*, k_g, is just the k_3 of previous sections, except that it is no longer restricted by the boundary conditions to take discrete values. The general considerations of Section 6.5 still apply, so that we can treat TM and TE modes separately. The solutions for f are identical to those for axially symmetric cavities already discussed. Although k_g is not restricted in magnitude, we note that for every eigenvalue of the two-dimensional equation, k_s, there is a lowest value of k, namely $k = k_s$ (often designated k_c for wave guides), for which k_g is real. This corresponds to the *cutoff frequency* below which waves are not transmitted by that mode, and the fields fall off exponentially with increasing z. In fact, the wave guide dispersion relation for a particular mode can easily be shown to take the form

$$k_g = \frac{\sqrt{\omega^2-{\omega_c}^2}}{c}, \tag{6.50}$$

where

$$\omega_c = k_c\,c \equiv k_s\,c \tag{6.51}$$

is the cutoff frequency. There is an absolute cutoff frequency associated with the mode of lowest frequency; *i.e.*, the mode with the lowest value of k_c.

For real k_g (*i.e.*, $\omega > \omega_c$) it is clear from Eq. (6.50) that the wave is propagated along the guide with a phase velocity

$$u_p = \frac{\omega}{k_g} = \frac{c}{\sqrt{1 - {\omega_c}^2/\omega^2}}. \tag{6.52}$$

It is evident that the phase velocity is greater than that of electromagnetic waves in free space. This velocity is not constant, however, but depends on the frequency. The wave guide thus behaves as a dispersive medium. The group velocity of a wave pulse propagated along the guide is given by

$$u_g = \frac{d\omega}{dk_g} = c\,\sqrt{1 - {\omega_c}^2/\omega^2}. \tag{6.53}$$

It can be seen that u_g is always smaller than c, and also that

$$u_p\, u_g = c^2. \tag{6.54}$$

For a TM mode ($H_z = 0$) Eqs. (6.34) yield

$$\boldsymbol{E}_s = \frac{\mathrm{i}\,k_g}{{k_s}^2}\,\nabla_s E_z, \tag{6.55a}$$

$$\boldsymbol{H}_s = \frac{\omega\epsilon_0}{k_g}\,\hat{\boldsymbol{z}}\wedge\boldsymbol{E}_s, \tag{6.55b}$$

where use has been made of $\partial/\partial z = \mathrm{i}\,k_g$. For TE modes ($E_z = 0$) Eqs. (6.35) give

$$\boldsymbol{H}_s = \frac{\mathrm{i}\,k_g}{{k_s}^2}\,\nabla_s H_z, \tag{6.56a}$$

$$\boldsymbol{E}_s = -\frac{\omega\mu_0}{k_g}\,\hat{\boldsymbol{z}}\wedge\boldsymbol{H}_s. \tag{6.56b}$$

The time-average z component of the Poynting vector $\boldsymbol{N}$ is given by

$$\overline{N_z} = \frac{|\boldsymbol{E}_s\wedge{\boldsymbol{H}_s}^*|}{2}. \tag{6.57}$$

It follows that

$$\overline{N_z} = \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{1}{\sqrt{1 - {\omega_c}^2/\omega^2}} \frac{H_{s0}^{\,2}}{2} \tag{6.58}$$

for TE modes, and

$$\overline{N_z} = \sqrt{\frac{\mu_0}{\epsilon_0}} \sqrt{1 - {\omega_c}^2/\omega^2} \frac{H_{s0}^{\,2}}{2} \tag{6.59}$$

for TM modes. The subscript 0 denotes the peak value of a wave quantity.

Wave guide losses can be estimated by integrating Eq. (6.14) over the wall of the guide for any given mode. The energy flow of a propagating wave attenuates as e^{-Kz}, where

$$K = \frac{\text{power loss per unit length of guide}}{\text{power transmitted through guide}}. \tag{6.60}$$

Thus,

$$K = \frac{1}{2\sigma d} \int \left({H_s}^2 + {H_z}^2\right)\, dS \bigg/ \int \overline{N_z}\, dS\;, \tag{6.61}$$

where the numerator is integrated over unit length of the wall and the denominator is integrated over the transverse cross section of the guide. It is customary to define a *guide impedance* Z_g by writing

$$\int \overline{N_z}\, dS = \frac{Z_g}{2} \int H_{s0}^{\,2}\, dS. \tag{6.62}$$

It follows from Eqs. (6.58) and (6.59) that

$$Z_g = \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{1}{\sqrt{1 - {\omega_c}^2/\omega^2}} \tag{6.63}$$

for TE modes, and

$$Z_g = \sqrt{\frac{\mu_0}{\epsilon_0}} \sqrt{1 - {\omega_c}^2/\omega^2} \tag{6.64}$$

for TM modes. For both types of mode $\boldsymbol{H}_s = (1/Z_g)\, \hat{\boldsymbol{z}} \wedge \boldsymbol{E}_s$.

6.8 Dielectric wave guides

We have seen that it is possible to propagate electromagnetic waves down a hollow conductor. However, other types of guiding structures are also possible. The general requirement for a guide of electromagnetic waves is that there be a flow of energy along the axis of the guiding structure but not perpendicular to it. This implies that the electromagnetic fields are appreciable only in the immediate neighbourhood of the guiding structure.

Consider an axisymmetric tube of arbitrary cross section made of some dielectric material and surrounded by a vacuum. This structure can serve as a wave guide provided that the dielectric constant of the material is sufficiently large. Note, however, that the boundary conditions satisfied by the electromagnetic fields are significantly different to those of a conventional wave guide. The transverse fields are governed by two equations; one for the region inside the dielectric, and the other for the vacuum region. Inside the dielectric we have

$$\left[\nabla_s^2 + \left(\epsilon_1\,\frac{\omega^2}{c^2} - {k_g}^2\right)\right]\psi = 0. \tag{6.65}$$

In the vacuum region we have

$$\left[\nabla_s^2 + \left(\frac{\omega^2}{c^2} - {k_g}^2\right)\right]\psi = 0. \tag{6.66}$$

Here, $\psi(x, y)\,{\rm e}^{\,{\rm i}\,k_g z}$ stands for either E_z or H_z, ϵ_1 is the relative permittivity of the dielectric material, and k_g is the guide propagation constant. The guide propagation constant must be the same both inside and outside the dielectric in order to satisfy the electromagnetic boundary conditions at all points on the surface of the tube.

Inside the dielectric the transverse Laplacian must be negative, so that the constant

$${k_s}^2 = \epsilon_1\,\frac{\omega^2}{c^2} - {k_g}^2 \tag{6.67}$$

is positive. Outside the cylinder the requirement of no transverse flow of energy can only be satisfied if the fields fall off exponentially (instead of oscillating).

Thus,

$$k_t{}^2 = k_g{}^2 - \frac{\omega^2}{c^2} \tag{6.68}$$

must be positive.

The oscillatory solutions (inside) must be matched to the exponentiating solutions (outside). The boundary conditions are the continuity of normal $\boldsymbol{B}$ and $\boldsymbol{D}$ and tangential $\boldsymbol{E}$ and $\boldsymbol{H}$ on the surface of the tube. These boundary conditions are far more complicated than those in a conventional wave guide. For this reason, the normal modes cannot usually be classified as either pure TE or pure TM modes. In general, the normal modes possess both electric and magnetic field components in the transverse plane. However, for the special case of a cylindrical tube of dielectric material the normal modes can have either pure TE or pure TM characteristics. Let us examine this case in detail.

Consider a dielectric cylinder of radius a and dielectric constant ϵ_1. For the sake of simplicity, let us only search for normal modes whose electromagnetic fields have no azimuthal variation. Equations (6.65) and (6.67) yield

$$\left(r^2\,\frac{d^2}{dr^2} + r\,\frac{d}{dr} + r^2 k_s{}^2\right)\psi = 0 \tag{6.69}$$

for $r < a$. The general solution to this equation is some linear combination of the Bessel functions $J_0(k_s r)$ and $Y_0(k_s r)$. However, since $Y_0(k_s r)$ is badly behaved at the origin ($r = 0$) the physical solution is $\psi \propto J_0(k_s r)$.

Equations (6.66) and (6.68) yield

$$\left(r^2\,\frac{d^2}{dr^2} + r\,\frac{d}{dr} - r^2 k_t{}^2\right)\psi = 0. \tag{6.70}$$

This can be rewritten

$$\left(z^2\,\frac{d^2}{dz^2} + z\,\frac{d}{dz} - z^2\right)\psi = 0, \tag{6.71}$$

where $z = k_t r$. This is type of *modified Bessel's equation*, whose most general form is

$$\left[z^2\,\frac{d^2}{dz^2} + z\,\frac{d}{dz} - (z^2 + m^2)\right]\psi = 0. \tag{6.72}$$

The two linearly independent solutions of the above equation are denoted $I_m(z)$ and $K_m(z)$. The asymptotic behaviour of these solutions at small $|z|$ is as follows:

$$I_m(z) = \left(\frac{z}{2}\right)^m \sum_{k=0}^{\infty} \frac{(z^2/4)^k}{k!(k+m)!}, \tag{6.73a}$$

$$\begin{aligned} K_m(z) &= \frac{1}{2}\left(\frac{z}{2}\right)^{-m} \sum_{k=0}^{m-1} \frac{(m-k-1)!}{k!}(-z^2/4)^k + (-1)^{m+1}\ln(z/2) I_m(z) \\ &\quad + (-1)^m \frac{1}{2}\left(\frac{z}{2}\right)^m \sum_{k=0}^{\infty} [\psi(k+1)+\psi(m+k+1)]\,\frac{(z^2/4)^m}{k!(m+k)!}. \end{aligned} \tag{6.73b}$$

Hence, I_m is well behaved in the limit $|z| \to 0$, whereas K_m is badly behaved. The asymptotic behaviour at large $|z|$ is

$$I_m(z) \simeq \frac{\mathrm{e}^z}{\sqrt{2\pi z}}\left[1 + O\left(\frac{1}{z}\right)\right], \tag{6.74a}$$

$$K_m(z) \simeq \sqrt{\frac{\pi}{2z}}\,\mathrm{e}^{-z}\left[1 + O\left(\frac{1}{z}\right)\right]. \tag{6.74b}$$

Hence, I_m is badly behaved in the limit $|z| \to \infty$, whereas K_m is well behaved. The behaviour of $I_0(z)$ and $K_0(z)$ is shown in Fig. 22. It is clear that the physical solution to Eq. (6.70) (*i.e.*, the one which decays as $|r| \to \infty$) is $\psi \propto K_0(k_t r)$.

The physical solution is

$$\psi = J_0(k_s r) \tag{6.75}$$

for $r \leq a$, and

$$\psi = A\,K_0(k_t r) \tag{6.76}$$

for $r > a$. Here, A is an arbitrary constant, and $\psi(r)\,\mathrm{e}^{\,\mathrm{i}\,k_g z}$ stands for either E_z or H_z. It follows from Eqs. (6.28) (using $\partial/\partial\theta = 0$) that

$$H_r = \mathrm{i}\,\frac{k_g}{{k_s}^2}\frac{\partial H_z}{\partial r}, \tag{6.77a}$$

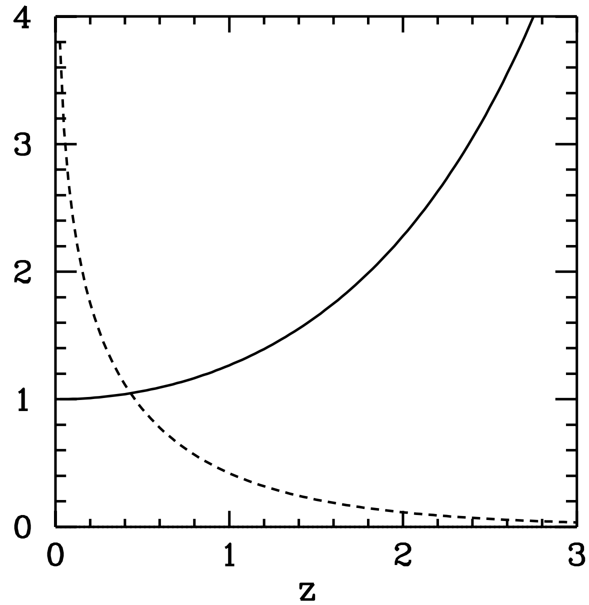


Figure 22: The Bessel functions $I_0(z)$ (solid line) and $K_0(z)$ (dotted line)

$$E_\theta = -\frac{\omega\mu_0}{k_g}\,H_r, \tag{6.77b}$$

$$H_\theta = \mathrm{i}\,\frac{\omega\epsilon_0\epsilon_1}{{k_s}^2}\frac{\partial E_z}{\partial r}, \tag{6.77c}$$

$$E_r = \frac{k_g}{\omega\epsilon_0\epsilon_1}\,H_\theta \tag{6.77d}$$

for $r \leq a$. There are an analogous set of relationships for $r > a$. The fact that the field components form two groups; (H_r, E_θ), which depend on H_z, and (H_θ, E_r), which depend on E_z; means that the normal modes take the form of either pure TE modes or pure TM modes.

For a TE mode ($E_z = 0$) we find that

$$H_z = J_0(k_s r), \tag{6.78a}$$

$$H_r = -\mathrm{i}\,\frac{k_g}{k_s}\,J_1(k_s r), \tag{6.78b}$$

$$E_\theta = \mathrm{i}\,\frac{\omega\mu_0}{k_s}\,J_1(k_s r) \tag{6.78c}$$

for $r \leq a$, and

$$H_z = A\,K_0(k_t r), \tag{6.79a}$$

$$H_r = \mathrm{i}\,A\,\frac{k_g}{k_t}\,K_1(k_t r), \tag{6.79b}$$

$$E_\theta = -\mathrm{i}\,A\,\frac{\omega\mu_0}{k_t}\,K_1(k_t r) \tag{6.79c}$$

for $r > a$. Here we have used

$$J_0'(z) = -J_1(z), \tag{6.80a}$$

$$K_0'(z) = -K_1(z), \tag{6.80b}$$

where $'$ denotes differentiation with respect to z. The boundary conditions require H_z, H_r, and E_θ to be continuous across $r = a$. Thus, it follows that

$$A\,K_0(k_t a) = J_0(k_s a), \tag{6.81a}$$

$$-A\,\frac{K_1(k_t r)}{k_t} = \frac{J_1(k_s a)}{k_s}. \tag{6.81b}$$

Eliminating the arbitrary constant A between the above two equations yields the dispersion relation

$$\frac{J_1(k_s a)}{k_s\,J_0(k_s a)} + \frac{K_1(k_t a)}{k_t\,K_0(k_t a)} = 0, \tag{6.82}$$

where

$$k_t^{\,2} + k_s^{\,2} = (\epsilon_1 - 1)\,\frac{\omega^2}{c^2}. \tag{6.82}$$

Figure 23 shows a graphical solution of the above dispersion relation. The roots correspond to the crossing points of the two curves; $-J_1(k_s a)/k_s J_0(k_s a)$ and $K_1(k_t a)/k_t K_0(k_t a)$. The vertical asymptotes of the first curve are given by the roots of $J_0(k_s a) = 0$. The vertical asymptote of the second curve occurs when $k_t = 0$; *i.e.*, when $k_s^{\,2} a^2 = (\epsilon_1 - 1)\,\omega^2 a^2/c^2$. Note from Eq. (6.82) that k_t decreases as k_s increases. In Fig. 23 there are two crossing points, corresponding to two distinct propagating modes of the system. It is evident that if the point $k_t = 0$

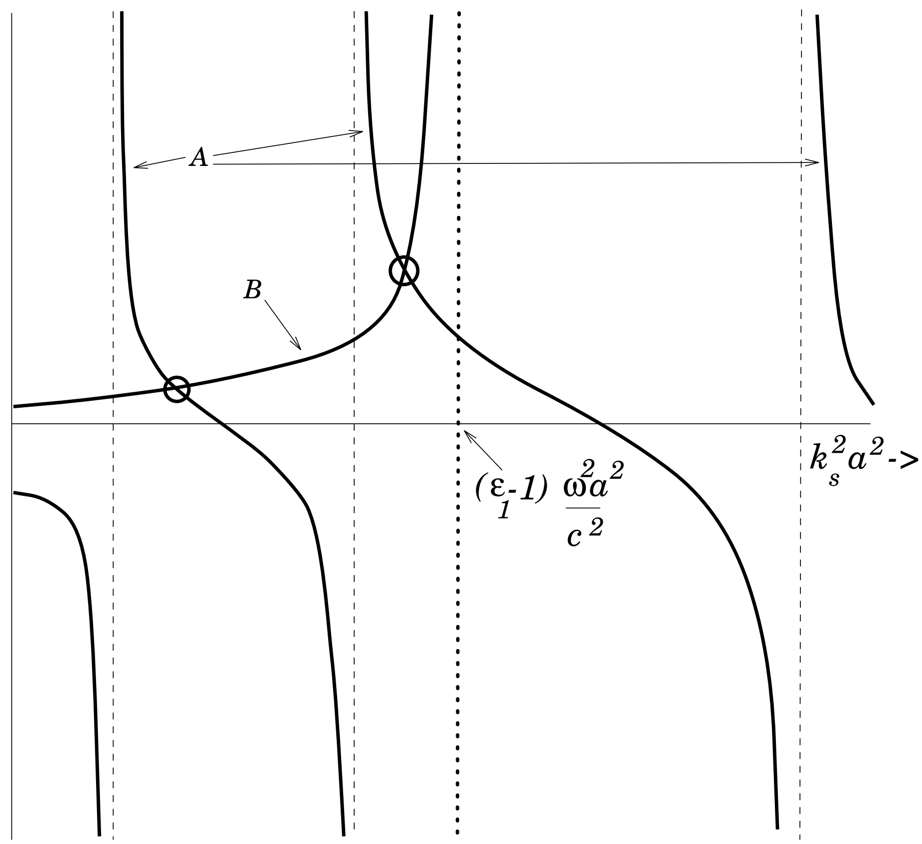


Figure 23: Graphical solution of the dispersion relation (6.82). The curve A represents $-J_1(k_s/a)/k_s J_0(k_s a)$. The curve B represents $K_1(k_t a)/k_t K_0(k_t a)$.

corresponds to a value of $k_s a$ which is less than the first root of $J_0(k_s a) = 0$, then there is no crossing of the two curves, and, hence, there are no propagating modes. Since the first root of $J_0(z) = 0$ occurs at $z = 2.4048$ (see Table 2) the condition for the existence of propagating modes can be written

$$\omega > \omega_{01} = \frac{2.4048\, c}{\sqrt{\epsilon_1 - 1}\, a}. \tag{6.83}$$

In other words, the mode frequency must lie above the cutoff frequency ω_{01} for the TE_{01} mode (here, the 0 corresponds to the number of nodes in the azimuthal direction, and the 1 refers to the 1st root of $J_0(z) = 0$). It is also evident that as the mode frequency is gradually increased the point $k_t = 0$ eventually crosses the second vertical asymptote of $-J_1(k_s/a)/k_s J_0(k_s a)$, at which point the TE_{02} mode can propagate. As ω is further increased more and more TE modes can propagate. The cutoff frequency for the TE_{0l} mode is given by

$$\omega_{0l} = \frac{j_{0l}\, c}{\sqrt{\epsilon_1 - 1}\, a}, \tag{6.84}$$

where j_{0l} is lth root of $J_0(z) = 0$ (in order of increasing z).

At the cutoff frequency for a particular mode $k_t = 0$, which implies from Eq. (6.68) that $k_g = \omega/c$. In other words, the mode propagates along the guide at the velocity of light in vacuum. Immediately below this cutoff frequency the system no longer acts as a guide but as an antenna, with energy being radiated radially. For frequencies well above the cutoff, k_t and k_g are of the same order of magnitude, and are large compared to k_s. This implies that the fields do not extend appreciably outside the dielectric cylinder.

For a TM mode ($H_z = 0$) we find that

$$E_z = J_0(k_s r), \tag{6.85a}$$

$$H_\theta = -\mathrm{i}\,\frac{\omega \epsilon_0 \epsilon_1}{k_s}\, J_1(k_s r), \tag{6.85b}$$

$$E_r = -\mathrm{i}\,\frac{k_g}{k_s}\, J_1(k_s r) \tag{6.85c}$$

for $r \leq a$, and

$$E_z = A\, K_0(k_t r), \tag{6.86a}$$

$$H_\theta = \mathrm{i}\,A\,\frac{\omega\epsilon_0}{k_t}\,K_1(k_t r), \tag{6.86b}$$

$$E_r = \mathrm{i}\,A\,\frac{k_g}{k_t}\,K_1(k_t r) \tag{6.86c}$$

for $r > a$. The boundary conditions require E_z, H_θ, and D_r to be continuous across $r = a$. Thus, it follows that

$$A\,K_0(k_t a) = J_0(k_s a), \tag{6.87a}$$

$$-A\,\frac{K_1(k_t r)}{k_t} = \epsilon_1\frac{J_1(k_s a)}{k_s}. \tag{6.87b}$$

Eliminating the arbitrary constant A between the above two equations yields the dispersion relation

$$\epsilon_1\frac{J_1(k_s a)}{k_s\,J_0(k_s a)} + \frac{K_1(k_t a)}{k_t\,K_0(k_t a)} = 0. \tag{6.88}$$

It is clear from this dispersion relation that the cutoff frequency for the TM_{0l} mode is exactly the same as that for the TE_{0l} mode. It is also clear that in the limit $\epsilon_1 \gg 1$ the propagation constants are determined by the roots of $J_1(k_s a) \simeq 0$. However, this is exactly the same as the determining equation for TE modes in a metallic wave guide of circular cross section (filled with dielectric of relative permittivity ϵ_1).

Modes with azimuthal dependence (*i.e.*, $m > 0$) have longitudinal components of both $\boldsymbol{E}$ and $\boldsymbol{H}$. This makes the mathematics somewhat more complicated. However, the basic results are the same as for $m = 0$ modes: for frequencies well above the cutoff frequency the modes are localized in the immediate vicinity of the cylinder.

7 The multipole expansion

7.1 Multipole expansion of the scalar wave equation

Consider the emission and scattering of electromagnetic radiation. This type of problem involves solving the vector wave equation. The solutions of this equation in free space are conveniently written as an expansion in orthogonal spherical waves. This expansion is known as the *multipole expansion*. Let us examine this expansion in more detail.

Before considering the vector wave equation, let us consider the somewhat simpler scalar wave equation. A scalar field $\psi(\boldsymbol{r}, t)$ satisfying the homogeneous wave equation

$$\nabla^2\psi - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = 0 \tag{7.1}$$

can be Fourier analyzed in time

$$\psi(\boldsymbol{r}, t) = \int_{-\infty}^{\infty} \psi(\boldsymbol{r}, \omega)\, \mathrm{e}^{-\mathrm{i}\,\omega t}\, d\omega \tag{7.2}$$

with each Fourier harmonic satisfying the Helmholtz wave equation

$$(\nabla^2 + k^2)\,\psi(\boldsymbol{r}, \omega) = 0, \tag{7.3}$$

where $k^2 = \omega^2/c^2$. We can write the Helmholtz equation in terms of spherical polar coordinates (r, θ, φ):

$$\left[\frac{1}{r^2}\frac{\partial}{\partial r}\, r^2\frac{\partial}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\,\sin\theta\,\frac{\partial}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} + k^2\right]\psi = 0. \tag{7.4}$$

As is well known, it is possible to solve this equation via the separation of variables:

$$\psi(\boldsymbol{r}, \omega) = \sum_{l,m} f_{lm}(r)\, Y_{lm}(\theta, \varphi). \tag{7.5}$$

Here, we restrict our attention to physical solutions which are well behaved in the angular variables θ and φ. The spherical harmonics $Y_{lm}(\theta, \varphi)$ satisfy the following equations:

$$
\begin{aligned}
-\frac{\partial^2 Y_{lm}}{\partial\varphi^2} &= m^2\, Y_{lm}, && (7.6a)\\
-\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\,\sin\theta\,\frac{\partial}{\partial\theta}+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right]Y_{lm} &= l(l+1)\,Y_{lm}, && (7.6b)
\end{aligned}
$$

where l is a non-negative integer, and m is an integer which satisfies the inequality $|m|\leq l$. The radial functions $f_{lm}(r)$ satisfy

$$\left[\frac{d^2}{dr^2}+\frac{2}{r}\frac{d}{dr}+k^2-\frac{l(l+1)}{r^2}\right]f_l(r)=0, \tag{7.7}$$

where there is no dependence on m. With the substitution

$$f_l(r)=\frac{u_l(r)}{r^{1/2}}, \tag{7.8}$$

Eq. (7.7) is transformed into

$$\left[\frac{d^2}{dr^2}+\frac{1}{r}\frac{d}{dr}+k^2-\frac{(l+1/2)^2}{r^2}\right]u_l(r)=0. \tag{7.9}$$

It can be seen, by comparison with Eq. (5.39), that this is a type of Bessel's equation of half-integer order $l+1/2$. Thus, we can write the solution for $f_{lm}(r)$ as

$$f_{lm}(r)=\frac{A_{lm}}{r^{1/2}}\,J_{l+1/2}(kr)+\frac{B_{lm}}{r^{1/2}}\,Y_{l+1/2}(kr), \tag{7.10}$$

where A_{lm} and B_{lm} are arbitrary constants. The half-integer order Bessel functions $J_{l+1/2}(z)$ and $Y_{l+1/2}(z)$ have analogous properties to the integer order Bessel functions $J_m(z)$ and $Y_m(z)$. In particular, the $J_{l+1/2}(z)$ are well behaved in the limit $|z|\to 0$, whereas the $Y_{l+1/2}(z)$ are badly behaved. The asymptotic expansions (5.43) remain valid when $m\to l+1/2$.

It is convenient to define the *spherical Bessel functions* $j_l(r)$ and $y_l(r)$, where

$$
\begin{aligned}
j_l(z) &= \left(\frac{\pi}{2z}\right)^{1/2} J_{l+1/2}(z), && (7.11a)\\
y_l(z) &= \left(\frac{\pi}{2z}\right)^{1/2} Y_{l+1/2}(z). && (7.11b)
\end{aligned}
$$

It is also convenient to define the spherical Hankel functions

$$h_l^{(1,2)}(z) = j_l(z) \pm \mathrm{i}\, y_l(z). \tag{7.12}$$

For real z, $h_l^{(2)}(z)$ is the complex conjugate of $h_l^{(1)}(z)$. It turns out that the spherical Bessel functions can be expressed in the closed form

$$j_l(z) \;=\; (-z)^l \left(\frac{1}{z}\frac{d}{dz}\right)^l \left(\frac{\sin z}{z}\right), \tag{7.13a}$$

$$y_l(z) \;=\; -(-z)^l \left(\frac{1}{z}\frac{d}{dz}\right)^l \left(\frac{\cos z}{z}\right). \tag{7.13b}$$

In the limit of small argument

$$j_l(z) \;\to\; \frac{z^l}{(2l+1)!!}\left[1 + O(z^2)\right], \tag{7.14a}$$

$$y_l(z) \;\to\; -\frac{(2l-1)!!}{z^{l+1}}\left[1 + O(z^2)\right], \tag{7.14b}$$

where $(2l+1)!! = (2l+1)(2l-1)(2l-3)\cdots 5{\cdot}3{\cdot}1$. In the limit of large argument

$$j_l(z) \;\to\; \frac{\sin(z - l\pi/2)}{z}, \tag{7.15a}$$

$$y_l(z) \;\to\; -\frac{\cos(z - l\pi/2)}{z}, \tag{7.15b}$$

and

$$h_l^{(1)} \to (-\mathrm{i})^{l+1}\,\frac{\mathrm{e}^{\mathrm{i}\,z}}{z}. \tag{7.16}$$

The inhomogeneous Helmholtz equation is conveniently solved using the Green's function $G_\omega(\boldsymbol{r},\boldsymbol{r}')$, which satisfies (see Eq. (2.109))

$$(\nabla^2 + k^2)\, G_\omega(\boldsymbol{r},\boldsymbol{r}') = -\delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{7.17}$$

The solution of this equation, subject to the *Sommerfeld radiation condition*, which ensures that sources radiate waves instead of absorbing them, is written

(see Section 2.13)

$$G_\omega(\boldsymbol{r},\boldsymbol{r}') = \frac{\mathrm{e}^{\,\mathrm{i}\,k|\boldsymbol{r}-\boldsymbol{r}'|}}{4\pi\,|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{7.18}$$

The spherical harmonics satisfy the completeness relation

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y_{lm}^*(\theta',\varphi')\,Y_{lm}(\theta,\varphi) = \delta(\varphi-\varphi')\,\delta(\cos\theta-\cos\theta'). \tag{7.19}$$

Now the three dimensional delta function can be written

$$\delta(\boldsymbol{r}-\boldsymbol{r}') = \frac{\delta(r-r')}{r^2}\,\delta(\varphi-\varphi')\,\delta(\cos\theta-\cos\theta'). \tag{7.20}$$

It follows that

$$\delta(\boldsymbol{r}-\boldsymbol{r}') = \frac{\delta(r-r')}{r^2}\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y_{lm}^*(\theta',\varphi')\,Y_{lm}(\theta,\varphi). \tag{7.21}$$

Let us expand the Green's function in the form

$$G_\omega(\boldsymbol{r},\boldsymbol{r}') = \sum_{l,m} g_l(r,r')\,Y_{lm}^*(\theta',\varphi')\,Y_{lm}(\theta,\varphi). \tag{7.22}$$

Substitution of this expression into Eq. (7.17) yields

$$\left[\frac{d^2}{dr^2}+\frac{2}{r}\frac{d}{dr}+k^2-\frac{l(l+1)}{r^2}\right]g_l = -\frac{\delta(r-r')}{r^2}. \tag{7.23}$$

The appropriate boundary conditions are that g_l is finite at the origin and corresponds to an *outgoing* wave at infinity (*i.e.*, $g \propto \mathrm{e}^{\,\mathrm{i}\,kr}$ in the limit $r\rightarrow\infty$). The solution of the above equation which satisfies these boundary conditions is

$$g_l(r,r') = A\,j_l(kr_<)\,h_l^{(1)}(kr_>), \tag{7.24}$$

where $r_<$ and $r_>$ are the greater and the lesser of r and r', respectively. The correct discontinuity in slope at $r=r'$ is assured if $A = \mathrm{i}\,k$, since

$$\frac{dh_l^{(1)}(z)}{dz}\,j_l(z) - h_l^{(1)}(z)\,\frac{dj_l(z)}{dz} = \frac{\mathrm{i}}{z^2}. \tag{7.25}$$

Thus, the expansion of the Green's function is

$$\frac{\mathrm{e}^{\mathrm{i}\,k|\boldsymbol{r}-\boldsymbol{r}'|}}{4\pi\,|\boldsymbol{r}-\boldsymbol{r}'|} = \mathrm{i}\,k \sum_{l=0}^{\infty} j_l(kr_<)\,h_l^{(1)}(kr_>) \sum_{m=-l}^{l} Y_{lm}^*(\theta',\varphi')\,Y_{lm}(\theta,\varphi). \tag{7.26}$$

This is a particularly useful result, as we shall discover, since it easily allows us to express the general solution of the inhomogeneous wave equation as a multipole expansion.

It is well known in quantum mechanics that Eq. (7.6b) can be written in the form

$$L^2\,Y_{lm} = l(l+1)\,Y_{lm}. \tag{7.27}$$

The differential operator L^2 is given by

$$L^2 = L_x{}^2 + L_y{}^2 + L_z{}^2, \tag{7.28}$$

where

$$\boldsymbol{L} = -\mathrm{i}\,\boldsymbol{r}\wedge\nabla \tag{7.29}$$

is $1/\hbar$ times the orbital angular momentum operator of wave mechanics.

The components of $\boldsymbol{L}$ can be conveniently written in the combinations

$$L_+ = L_x + \mathrm{i}\,L_y = \mathrm{e}^{\mathrm{i}\,\varphi}\left(\frac{\partial}{\partial\theta} + \mathrm{i}\,\cot\theta\,\frac{\partial}{\partial\varphi}\right), \tag{7.30a}$$

$$L_- = L_x - \mathrm{i}\,L_y = \mathrm{e}^{-\mathrm{i}\,\varphi}\left(-\frac{\partial}{\partial\theta} + \mathrm{i}\,\cot\theta\,\frac{\partial}{\partial\varphi}\right), \tag{7.30b}$$

$$L_z = -\mathrm{i}\,\frac{\partial}{\partial\varphi}. \tag{7.30c}$$

We note that $\boldsymbol{L}$ operates only on angular variables and is independent of r. From the definition (7.29) it is evident that

$$\boldsymbol{r}\cdot\boldsymbol{L} = 0 \tag{7.31}$$

holds as an operator equation. It is easily demonstrated from Eqs. (7.30) that

$$L^2 = -\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\,\sin\theta\,\frac{\partial}{\partial\theta} - \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{7.32}$$

The following results are well known in quantum mechanics:

$$L_+ Y_{lm} = \sqrt{(l-m)(l+m+1)}\, Y_{l,m+1}, \tag{7.33a}$$

$$L_- Y_{lm} = \sqrt{(l+m)(l-m+1)}\, Y_{l,m-1}, \tag{7.33b}$$

$$L_z Y_{lm} = m\, Y_{lm}. \tag{7.33c}$$

In addition,

$$L^2\, \boldsymbol{L} = \boldsymbol{L}\, L^2, \tag{7.34a}$$

$$\boldsymbol{L} \wedge \boldsymbol{L} = \mathrm{i}\, \boldsymbol{L}, \tag{7.34b}$$

$$L_j \nabla^2 = \nabla^2 L_j, \tag{7.34c}$$

where

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\, r^2 \frac{\partial}{\partial r} - \frac{L^2}{r^2}. \tag{7.35}$$

7.2 Multipole expansion of the vector wave equation

Maxwell's equations in free space reduce to

$$\nabla\cdot\boldsymbol{E} = 0, \tag{7.36a}$$

$$\nabla\cdot c\boldsymbol{B} = 0, \tag{7.36b}$$

$$\nabla \wedge \boldsymbol{E} = \mathrm{i}\, k\, c\boldsymbol{B}, \tag{7.36c}$$

$$\nabla \wedge c\boldsymbol{B} = -\mathrm{i}\, k\, \boldsymbol{E}, \tag{7.36d}$$

assuming an $\mathrm{e}^{-\mathrm{i}\,\omega t}$ time dependence of all field quantities. Here, $k = \omega/c$. Eliminating $\boldsymbol{E}$ between Eqs. (7.36c) and (7.36d), we obtain the following equations for $\boldsymbol{B}$:

$$(\nabla^2 + k^2)\boldsymbol{B} = 0, \tag{7.37a}$$

$$\nabla\cdot\boldsymbol{B} = 0, \tag{7.37b}$$

with $\boldsymbol{E}$ given by

$$\boldsymbol{E} = \frac{\mathrm{i}}{k}\nabla \wedge c\boldsymbol{B}. \tag{7.38}$$

Alternatively, $\boldsymbol{B}$ can be eliminated to give

$$\begin{aligned}(\nabla^2 + k^2)\boldsymbol{E} &= 0, \qquad (7.39a)\\ \nabla\cdot\boldsymbol{E} &= 0, \qquad (7.39b)\end{aligned}$$

with $\boldsymbol{B}$ given by

$$c\boldsymbol{B} = -\frac{\mathrm{i}}{k}\nabla\wedge\boldsymbol{E}. \tag{7.40}$$

It is clear that each Cartesian component of $\boldsymbol{B}$ and $\boldsymbol{E}$ satisfies the Helmholtz wave equation (7.3). Hence, these components can be written in a general expansion of the form

$$\psi(\boldsymbol{r}) = \sum_{l,m}\left[A_{lm}^{(1)}\,h_l^{(1)}(kr) + A_{lm}^{(2)}\,h_l^{(2)}(kr)\right]Y_{lm}(\theta,\varphi), \tag{7.41}$$

where ψ stands for any Cartesian component of $\boldsymbol{E}$ or $c\boldsymbol{B}$. Note, however, that the three Cartesian components of $\boldsymbol{E}$ or $\boldsymbol{B}$ are not entirely independent, since they must also satisfy the constraints $\nabla\cdot\boldsymbol{E} = 0$ and $\nabla\cdot\boldsymbol{B} = 0$. Let us examine how these constraints can be satisfied with the minimum labour.

Consider the scalar $\boldsymbol{r}\cdot\boldsymbol{A}$, where $\boldsymbol{A}$ is a well behaved vector field. It is easily verified that

$$\nabla^2(\boldsymbol{r}\cdot\boldsymbol{A}) = \boldsymbol{r}\cdot(\nabla^2\boldsymbol{A}) + 2\nabla\cdot\boldsymbol{A}. \tag{7.42}$$

It follows from Eqs. (7.37) and (7.39) that the scalars $\boldsymbol{r}\cdot\boldsymbol{E}$ and $\boldsymbol{r}\cdot\boldsymbol{B}$ both satisfy the Helmholtz wave equation:

$$\begin{aligned}(\nabla^2 + k^2)(\boldsymbol{r}\cdot\boldsymbol{E}) &= 0, \qquad (7.43a)\\ (\nabla^2 + k^2)(\boldsymbol{r}\cdot\boldsymbol{B}) &= 0. \qquad (7.43b)\end{aligned}$$

Thus, the general solutions for $\boldsymbol{r}\cdot\boldsymbol{E}$ and $\boldsymbol{r}\cdot c\boldsymbol{B}$ can be written in the form (7.41).

Let us define a *magnetic multipole* field of order (l, m) by the conditions

$$\begin{aligned}\boldsymbol{r}\cdot c\boldsymbol{B}_{lm}^{(M)} &= \frac{l(l+1)}{k}\,g_l(kr)\,Y_{lm}(\theta,\varphi), \qquad (7.44a)\\ \boldsymbol{r}\cdot\boldsymbol{E}_{lm}^{(M)} &= 0, \qquad (7.44b)\end{aligned}$$

where

$$g_l(kr) = A_l^{(1)}\, h_l^{(1)}(kr) + A_l^{(2)}\, h_l^{(2)}(kr). \tag{7.45}$$

The presence of the factor $l(l+1)/k$ is for later convenience. Equation (7.40) yields

$$k\,\boldsymbol{r}\cdot c\boldsymbol{B} = -\mathrm{i}\,\boldsymbol{r}\cdot(\nabla\wedge\boldsymbol{E}) = -\mathrm{i}\,(\boldsymbol{r}\wedge\nabla)\cdot\boldsymbol{E} = \boldsymbol{L}\cdot\boldsymbol{E}, \tag{7.46}$$

where $\boldsymbol{L}$ is given by Eq. (7.29). With $\boldsymbol{r}\cdot\boldsymbol{B}$ given by Eq. (7.44a), the electric field associated with a magnetic multipole must satisfy

$$\boldsymbol{L}\cdot\boldsymbol{E}_{lm}^{(M)}(r,\theta,\varphi) = l(l+1)\, g_l(kr)\, Y_{lm}(\theta,\varphi) \tag{7.47}$$

and $\boldsymbol{r}\cdot\boldsymbol{E}_{lm}^{(M)} = 0$. Note that the operator $\boldsymbol{L}$ acts only on the angular variables (θ,φ). This means that the radial dependence of $\boldsymbol{E}_{lm}^{(M)}$ must be given by $g_l(kr)$. Note also, from Eqs. (7.33), that the operator $\boldsymbol{L}$ acting on Y_{lm} transforms the m value but does not change the l value. It is easily seen from Eqs. (7.27) and (7.31) that the solution to Eqs. (7.44b) and (7.47) can be written in the form

$$\boldsymbol{E}_{lm}^{(M)} = g_l(kr)\,\boldsymbol{L}\,Y_{lm}(\theta,\varphi). \tag{7.48}$$

Thus, the angular dependence of $\boldsymbol{E}_{lm}^{(M)}$ consists of some linear combination of $Y_{l,m-1}$, Y_{lm}, and $Y_{l,m+1}$. Equation (7.48), together with

$$c\boldsymbol{B}_{lm}^{(M)} = -\frac{\mathrm{i}}{k}\,\nabla\wedge\boldsymbol{E}_{lm}^{(M)}, \tag{7.49}$$

specifies the electromagnetic fields of a *magnetic* multipole of order (l,m). Note from Eq. (7.31) that the electric field given by Eq. (7.48) is transverse to the radius vector. Thus, magnetic multipole fields are sometimes termed *transverse electric* (TE) multipole fields.

The fields of an *electric* or *transverse magnetic* (TM) multipole of order (l,m) are specified by the conditions

$$\boldsymbol{r}\cdot\boldsymbol{E}_{lm}^{(E)} = -\frac{l(l+1)}{k}\, f_l(kr)\, Y_{lm}(\theta,\varphi), \tag{7.50a}$$

$$\boldsymbol{r}\cdot\boldsymbol{B}_{lm}^{(E)} = 0. \tag{7.50b}$$

It follows that the fields of an electric multipole are given by

$$c\boldsymbol{B}_{lm}^{(E)} = f_l(kr)\,\boldsymbol{L}\,Y_{lm}(\theta,\varphi), \tag{7.51a}$$

$$\boldsymbol{E}_{lm}^{(E)} = \frac{\mathrm{i}}{k}\nabla\wedge c\boldsymbol{B}_{lm}^{(E)}. \tag{7.51b}$$

The radial function $f_l(kr)$ is given by an expression like (7.45).

The two sets of multipole fields (7.48), (7.49), and (7.51), form a complete set of vector solutions to Maxwell's equations in free space. Since the vector spherical harmonic $\boldsymbol{L}\,Y_{lm}$ plays an important role in multipole fields, it is convenient to introduce the normalized form

$$\boldsymbol{X}_{lm}(\theta,\varphi) = \frac{1}{\sqrt{l(l+1)}}\,\boldsymbol{L}\,Y_{lm}(\theta,\varphi). \tag{7.52}$$

It can be demonstrated that the vector spherical harmonics possess the orthogonality properties

$$\int \boldsymbol{X}_{l'm'}^{*}\cdot\boldsymbol{X}_{lm}\,d\Omega = \delta_{ll'}\,\delta_{mm'}, \tag{7.53a}$$

$$\int \boldsymbol{X}_{l'm'}^{*}\cdot(\boldsymbol{r}\wedge\boldsymbol{X}_{lm})\,d\Omega = 0, \tag{7.53b}$$

for all l, l', m, and m'.

By combining the two types of fields we can write the general solution to Maxwell's equations in free space in the form

$$c\boldsymbol{B} = \sum_{l,m}\left[a_E(l,m)\,f_l(kr)\,\boldsymbol{X}_{lm} - \frac{\mathrm{i}}{k}\,a_M(l,m)\,\nabla\wedge g_l(kr)\boldsymbol{X}_{lm}\right], \tag{7.54a}$$

$$\boldsymbol{E} = \sum_{l,m}\left[\frac{\mathrm{i}}{k}\,a_E(l,m)\,\nabla\wedge f_l(kr)\boldsymbol{X}_{lm} + a_M(l,m)\,g_l(kr)\,\boldsymbol{X}_{lm}\right], \tag{7.54b}$$

where the coefficients $a_E(l,m)$ and $a_M(l,m)$ specify the amounts of electric (l,m) and magnetic (l,m) multipole fields. The radial functions $f_l(kr)$ and $g_l(kr)$ are of the form (7.45). The coefficients $a_E(l,m)$ and $a_M(l,m)$, as well as the relative proportions in (7.45), are determined by the sources and the boundary conditions.

Equations (7.54) yield

$$\begin{aligned} \boldsymbol{r}\cdot c\boldsymbol{B} &= \frac{1}{k}\sum_{l,m} a_M(l,m)\, g_l(kr)\, \boldsymbol{L}\,\boldsymbol{X}_{lm} \\ &= \frac{1}{k}\sum_{l,m} a_M(l,m)\, g_l(kr)\sqrt{l(l+1)}\, Y_{lm}, \end{aligned} \tag{7.55}$$

and

$$\begin{aligned} \boldsymbol{r}\cdot\boldsymbol{E} &= -\frac{1}{k}\sum_{l,m} a_E(l,m)\, f_l(kr)\, \boldsymbol{L}\,\boldsymbol{X}_{lm} \\ &= -\frac{1}{k}\sum_{l,m} a_E(l,m)\, f_l(kr)\sqrt{l(l+1)}\, Y_{lm}, \end{aligned} \tag{7.56}$$

where use has been made of Eqs. (7.27), (7.29), and (7.31). It follows from the well known orthogonality property of the spherical harmonics that

$$a_M(l,m)\, g_l(kr) = \frac{k}{\sqrt{l(l+1)}}\int Y_{lm}^*\, \boldsymbol{r}\cdot c\boldsymbol{B}\, d\Omega, \tag{7.57a}$$

$$a_E(l,m)\, f_l(kr) = -\frac{k}{\sqrt{l(l+1)}}\int Y_{lm}^*\, \boldsymbol{r}\cdot\boldsymbol{E}\, d\Omega. \tag{7.57b}$$

Thus, knowledge of $\boldsymbol{r}\cdot\boldsymbol{B}$ and $\boldsymbol{r}\cdot\boldsymbol{E}$ at two different radii in a source free region permits a complete specification of the fields, including the relative proportions of $h_l^{(1)}$ and $h_l^{(2)}$ in f_l and g_l.

7.3 Properties of multipole fields

Let us examine some of the properties of the multipole fields (7.48), (7.49), and (7.51). Consider, first of all, the so-called *near zone*, for which $kr \ll 1$. In this region $f_l(kr)$ is proportional to $y_l(kr)$, given by the asymptotic expansion (7.14b), unless its coefficient vanishes identically. Excluding this possibility, the limiting behaviour of the magnetic field for an electric (l, m) multipole is

$$c\boldsymbol{B}_{lm}^{(E)} \to -\frac{k}{l}\,\boldsymbol{L}\,\frac{Y_{lm}}{r^{l+1}}, \tag{7.58}$$

where the proportionality coefficient is chosen for later convenience. To find the electric field we must take the curl of the right-hand side. The following operator identity is useful

$$\mathrm{i}\,\nabla\wedge\boldsymbol{L} = \boldsymbol{r}\,\nabla^2 - \nabla\left(1 + r\frac{\partial}{\partial r}\right). \tag{7.59}$$

The electric field (7.51b) is

$$\boldsymbol{E}_{lm}^{(E)} \to \frac{-\mathrm{i}}{l}\,\nabla\wedge\boldsymbol{L}\left(\frac{Y_{lm}}{r^{l+1}}\right). \tag{7.60}$$

Since Y_{lm}/r^{l+1} is a solution of Laplace's equation, the first term in (7.59) vanishes. Consequently, the electric field at close distances for an electric (l, m) multipole is

$$\boldsymbol{E}_{lm}^{(E)} \to -\nabla\left(\frac{Y_{lm}}{r^{l+1}}\right). \tag{7.61}$$

This, of course, is an electrostatic multipole field. Such a field is obtained in a more straightforward manner by observing that $\boldsymbol{E} \to -\nabla\phi$, where $\nabla^2\phi = 0$, in the near zone. Solving Laplace's equation by separation of variables in spherical polar coordinates, and demanding that ϕ be well behaved as $|\boldsymbol{r}| \to \infty$, yields

$$\phi(r, \theta, \varphi) = \sum_{l,m}\frac{Y_{lm}(\theta, \varphi)}{r^{l+1}}. \tag{7.62}$$

Note that the magnetic field (7.58) (normalized with respect to c^{-1}) is smaller than the electric field (7.61) by a factor of order kr. Thus, in the near zone the magnetic field associated with an electric multipole is always much smaller than the corresponding electric field. For magnetic multipole fields it is evident from Eqs. (7.48), (7.49), and (7.51) that the roles of $\boldsymbol{E}$ and $\boldsymbol{B}$ are interchanged according to the transformation

$$\boldsymbol{E}^{(E)} \rightarrow -c\boldsymbol{B}^{(M)}, \tag{7.63a}$$

$$c\boldsymbol{B}^{(E)} \rightarrow \boldsymbol{E}^{(M)}. \tag{7.63b}$$

In the so-called *far zone* or *radiation zone*, for which $kr \gg 1$, the multipole fields depend on the boundary conditions imposed at infinity. For definiteness, let us consider the case of outgoing waves at infinity, which is appropriate to radiation by a localized source. For this case, the radial function $f_l(kr)$ is proportional to the spherical Hankel function $h_l^{(1)}(kr)$. From the asymptotic form (7.16), it is clear that in the radiation zone the magnetic field of an electric (l, m) multipole goes as

$$c\boldsymbol{B}_{lm}^{(E)} \rightarrow (-\mathrm{i})^{l+1} \frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{kr}\, \boldsymbol{L}\, Y_{lm}. \tag{7.64}$$

Using Eq. (7.51b), the electric field can be written

$$\boldsymbol{E}_{lm}^{(E)} = \frac{(-\mathrm{i})^l}{k^2} \left[\nabla\left(\frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{r}\right) \wedge \boldsymbol{L}\, Y_{lm} + \frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{r}\, \nabla \wedge \boldsymbol{L}\, Y_{lm}\right]. \tag{7.65}$$

Neglecting terms which fall off faster than $1/r$, the above expression reduces to

$$\boldsymbol{E}_{lm}^{(E)} = -(-\mathrm{i})^{l+1} \frac{\mathrm{e}^{\,\mathrm{i}\,kr}}{kr} \left[\boldsymbol{n} \wedge \boldsymbol{L}\, Y_{lm} - \frac{1}{k}(\boldsymbol{r}\nabla^2 - \nabla) Y_{lm}\right], \tag{7.66}$$

where use has been made of the identity (7.59), and $\boldsymbol{n} = \boldsymbol{r}/r$ is a unit vector pointing in the radial direction. The second term in square brackets is smaller than the first term by a factor of order $1/kr$, and can therefore be neglected in the limit $kr \gg 1$. Thus, we find that the electric field in the radiation zone is given by

$$\boldsymbol{E}_{lm}^{(E)} = c\boldsymbol{B}_{lm}^{(E)} \wedge \boldsymbol{n}, \tag{7.67}$$

where $\boldsymbol{B}_{lm}^{(E)}$ is given by Eq. (7.64). These fields are typical radiation fields; *i.e.*, they are transverse to the radius vector, mutually orthogonal, and fall off like $1/r$. For magnetic multipoles we merely make the transformation (7.63).

Consider a linear superposition of electric (l, m) multipoles with different m values, but all possessing a common l value. It follows from Eqs. (7.54) that

$$c\boldsymbol{B}_l = \sum_l a_E(l,m)\, \boldsymbol{X}_{lm}\, h_l^{(1)}(kr)\, \mathrm{e}^{-\mathrm{i}\,\omega t}, \tag{7.68a}$$

$$\boldsymbol{E}_l = \frac{\mathrm{i}}{k} \nabla \wedge c\boldsymbol{B}_l. \tag{7.68b}$$

For harmonically varying fields the time averaged energy density is given by

$$u = \frac{\epsilon_0}{4}\,(\boldsymbol{E}\cdot\boldsymbol{E}^* + c\boldsymbol{B}\cdot c\boldsymbol{B}^*). \tag{7.69}$$

In the radiation zone the two terms are equal. It follows that the energy density contained in a spherical shell between radii r and $r + dr$ is

$$dU = \frac{\epsilon_0\, dr}{2k^2} \sum_{m,m'} a_E^*(l,m')\, a_E(l,m) \int \boldsymbol{X}_{lm'}^* \cdot \boldsymbol{X}_{lm}\, d\Omega, \tag{7.70}$$

where the asymptotic form (7.16) of the spherical Hankel function has been used. Making use of the orthogonality relation (7.53a), we obtain

$$\frac{dU}{dr} = \frac{\epsilon_0}{2k^2} \sum_m |a_E(l,m)|^2, \tag{7.71}$$

which is clearly independent of the radius. For a general superposition of electric and magnetic multipoles the sum over m becomes a sum over l and m, and $|a_E|^2$ becomes $|a_E|^2 + |a_M|^2$. Thus, the total energy in a spherical shell in the radiation zone is an *incoherent sum* over all multipoles.

The time averaged angular momentum density of harmonically varying electromagnetic fields is given by

$$\boldsymbol{m} = \frac{\epsilon_0}{2}\,\mathrm{Re}\,[\boldsymbol{r} \wedge (\boldsymbol{E} \wedge \boldsymbol{B}^*)]. \tag{7.72}$$

For a superposition of electric multipoles the triple product can be expanded and the electric field (7.68b) substituted, to give

$$\boldsymbol{m} = \frac{\epsilon_0 c}{2k}\,\mathrm{Re}\,[\boldsymbol{B}^*(\boldsymbol{L}\cdot\boldsymbol{B})]. \tag{7.73}$$

Thus, the angular momentum in a spherical shell lying between radii r and $r+dr$ in the radiation zone is

$$d\boldsymbol{M} = \frac{\epsilon_0 c\,dr}{2k^3}\,\mathrm{Re}\sum_{m,m'} a_E^*(l,m')\,a_E(l,m)\int(\boldsymbol{L}\cdot\boldsymbol{X}_{lm'})^*\,\boldsymbol{X}_{lm}\,d\Omega. \tag{7.74}$$

It follows from Eqs. (7.27) and (7.52) that

$$\frac{d\boldsymbol{M}}{dr} = \frac{\epsilon_0 c}{2k^3}\,\mathrm{Re}\sum_{m,m'} a_E^*(l,m')\,a_E(l,m)\int Y_{lm'}^*\,\boldsymbol{L}\,Y_{lm}\,d\Omega. \tag{7.75}$$

According to Eqs. (7.33), the Cartesian components of $d\boldsymbol{M}/dr$ can be written:

$$\begin{aligned}\frac{dM_x}{dr} &= \frac{\epsilon_0 c}{4k^3}\,\mathrm{Re}\sum_m\Big[\sqrt{(l-m)(l+m+1)}\,a_E^*(l,m+1)\\ &\quad+\sqrt{(l+m)(l-m+1)}\,a_E^*(l,m-1)\Big]\,a_E(l,m), \qquad (7.76\text{a})\\ \frac{dM_y}{dr} &= \frac{\epsilon_0 c}{4k^3}\,\mathrm{Im}\sum_m\Big[\sqrt{(l-m)(l+m+1)}\,a_E^*(l,m+1)\\ &\quad-\sqrt{(l+m)(l-m+1)}\,a_E^*(l,m-1)\Big]\,a_E(l,m), \qquad (7.76\text{b})\\ \frac{dM_z}{dr} &= \frac{\epsilon_0 c}{2k^3}\sum_m m\,|a_E(l,m)|^2. \qquad (7.76\text{c})\end{aligned}$$

Thus, for a general lth order electric multipole that consists of a superposition of different m values, only the z component of the angular momentum takes a relatively simple form.

7.4 Sources of multipole radiation

Let us now examine the connection between multipole fields and their sources. Suppose that there exist localized distributions of electric change $\rho(\boldsymbol{r}, t)$, true current $\boldsymbol{j}(\boldsymbol{r}, t)$, and magnetization $\boldsymbol{M}(\boldsymbol{r}, t)$. We assume that the time dependence can be analyzed into its Fourier components, and we therefore only consider harmonically varying sources, $\rho(\boldsymbol{r})\,\mathrm{e}^{-\mathrm{i}\,\omega t}$, $\boldsymbol{j}(\boldsymbol{r})\,\mathrm{e}^{-\mathrm{i}\,\omega t}$, and $\boldsymbol{M}(\boldsymbol{r})\,\mathrm{e}^{-\mathrm{i}\,\omega t}$, where it is understood that we take the real parts of complex quantities.

Maxwell's equations can be written

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{7.77a}$$

$$\nabla\cdot\boldsymbol{B} = 0, \tag{7.77b}$$

$$\nabla\wedge\boldsymbol{E} - \mathrm{i}\,k\,c\boldsymbol{B} = 0, \tag{7.77c}$$

$$\nabla\wedge c\boldsymbol{B} + \mathrm{i}\,k\,\boldsymbol{E} = \mu_0 c\,(\boldsymbol{j} + \nabla\wedge\boldsymbol{M}), \tag{7.77d}$$

with the continuity equation

$$\mathrm{i}\,\omega\rho = \nabla\cdot\boldsymbol{j}. \tag{7.78}$$

It is convenient to deal with divergenceless fields. Thus, we use as the field variables, $\boldsymbol{B}$ and

$$\boldsymbol{E}' = \boldsymbol{E} + \frac{\mathrm{i}}{\epsilon_0\omega}\,\boldsymbol{j}. \tag{7.79}$$

In the region outside the sources $\boldsymbol{E}'$ reduces to $\boldsymbol{E}$. When expressed in terms of these fields, Maxwell's equations become

$$\nabla\cdot\boldsymbol{E}' = 0, \tag{7.80a}$$

$$\nabla\cdot\boldsymbol{B} = 0, \tag{7.80b}$$

$$\nabla\wedge\boldsymbol{E}' - \mathrm{i}\,k\,c\boldsymbol{B} = \frac{\mathrm{i}}{\epsilon_0\omega}\nabla\wedge\boldsymbol{j}, \tag{7.80c}$$

$$\nabla\wedge c\boldsymbol{B} + \mathrm{i}\,k\,\boldsymbol{E}' = \mu_0 c\,\nabla\wedge\boldsymbol{M}. \tag{7.80d}$$

The curl equations can be combined to give two inhomogeneous Helmholtz wave equations:

$$(\nabla^2 + k^2)c\boldsymbol{B} = -\mu_0 c\,\nabla\wedge(\boldsymbol{j} + \nabla\wedge\boldsymbol{M}), \tag{7.81}$$

and

$$(\nabla^2 + k^2)\boldsymbol{E}' = -\mathrm{i}\,k\,\mu_0 c\,\nabla \wedge \left(\boldsymbol{M} + \frac{\nabla \wedge \boldsymbol{j}}{k^2}\right). \tag{7.82}$$

These equations, together with $\nabla\cdot\boldsymbol{B} = 0$, and $\nabla\cdot\boldsymbol{E}' = 0$, and the curl equations giving $\boldsymbol{E}'$ in terms of $\boldsymbol{B}$ and *vice versa*, are the analogues to Eqs. (7.37)–(7.40) when sources are present.

Since the multipole coefficients in Eqs. (7.54) are determined according to Eqs. (7.57) from the scalars $\boldsymbol{r}\cdot\boldsymbol{B}$ and $\boldsymbol{r}\cdot\boldsymbol{E}'$, it is sufficient to consider wave equations for these quantities, rather than the vector fields $\boldsymbol{B}$ and $\boldsymbol{E}'$. From Eqs. (7.42), (7.81), (7.82), and the identity

$$\boldsymbol{r}\cdot(\nabla \wedge \boldsymbol{A}) = (\boldsymbol{r} \wedge \nabla)\cdot\boldsymbol{A} = \mathrm{i}\,\boldsymbol{L}\cdot\boldsymbol{A} \tag{7.83}$$

for any vector field $\boldsymbol{A}$, we obtain the inhomogeneous wave equations

$$(\nabla^2 + k^2)\,\boldsymbol{r}\cdot c\boldsymbol{B} \;=\; -\mathrm{i}\,\mu_0 c\,\boldsymbol{L}\cdot(\boldsymbol{j} + \nabla \wedge \boldsymbol{M}), \tag{7.84a}$$

$$(\nabla^2 + k^2)\,\boldsymbol{r}\cdot\boldsymbol{E}' \;=\; k\,\mu_0 c\,\boldsymbol{L}\cdot\left(\boldsymbol{M} + \frac{\nabla \wedge \boldsymbol{j}}{k^2}\right). \tag{7.84b}$$

Now the Green's function for the inhomogeneous Helmholtz equation (defined by Eq. (7.17)), subject to the boundary condition of outgoing waves at infinity, is given by Eq. (7.18). It follows that Eqs. (7.84) can be inverted to give

$$\boldsymbol{r}\cdot c\boldsymbol{B}(\boldsymbol{r}) \;=\; \frac{\mathrm{i}\,\mu_0 c}{4\pi}\int \frac{\mathrm{e}^{\,\mathrm{i}\,k\,|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,\boldsymbol{L}'\cdot[\boldsymbol{j}(\boldsymbol{r}') + \nabla' \wedge \boldsymbol{M}(\boldsymbol{r}')]\;d^3\boldsymbol{r}', \tag{7.85a}$$

$$\boldsymbol{r}\cdot\boldsymbol{E}'(\boldsymbol{r}) \;=\; -\frac{k\,\mu_0 c}{4\pi}\int \frac{\mathrm{e}^{\,\mathrm{i}\,k\,|\boldsymbol{r}-\boldsymbol{r}'|}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,\boldsymbol{L}'\cdot\left[\boldsymbol{M}(\boldsymbol{r}') + \frac{\nabla' \wedge \boldsymbol{j}(\boldsymbol{r}')}{k^2}\right]\;d^3\boldsymbol{r}'. \tag{7.85b}$$

In order to evaluate the multipole coefficients by means of Eqs. (7.57), we first observe that the requirement of outgoing waves at infinity makes $A_l^{(2)} = 0$ in Eq. (7.45). Thus, we choose $f_l(kr) = g_l(kr) = h_l^{(1)}(kr)$ in Eqs. (7.54) as the radial eigenfunctions of $\boldsymbol{E}$ and $\boldsymbol{B}$ in the source free region. Next, let us consider the expansion (7.26) of the Green's function for the Helmholtz equation in terms of spherical harmonics. We assume that the point $\boldsymbol{r}$ lies outside some spherical shell which completely encloses the sources. It follows that $r_< = r'$ and $r_> = r$ in all of the integrations. Making use of the orthogonality property of the spherical harmonics, it follows from Eq. (7.26) that

$$\int Y_{lm}^*(\theta,\varphi)\,\frac{\mathrm{e}^{\,\mathrm{i}\,k\,|\boldsymbol{r}-\boldsymbol{r}'|}}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|}\,d\Omega = \mathrm{i}\,k\,h_l^{(1)}(kr)\,j_l(kr')\,Y_{lm}^*(\theta',\varphi'). \tag{7.86}$$

Finally, Eqs. (7.57), (7.85), and (7.86) yield

$$a_E(l,m) = \frac{\mu_0 c\,\mathrm{i}\,k^3}{\sqrt{l(l+1)}}\int j_l(kr)\,Y_{lm}^*\,\boldsymbol{L}\cdot\left(\boldsymbol{M}+\frac{\nabla\wedge\boldsymbol{j}}{k^2}\right)d^3\boldsymbol{r}, \tag{7.87a}$$

$$a_M(l,m) = -\frac{\mu_0 c\,k^2}{\sqrt{l(l+1)}}\int j_l(kr)\,Y_{lm}^*\,\boldsymbol{L}\cdot(\boldsymbol{j}+\nabla\wedge\boldsymbol{M})\,d^3\boldsymbol{r}. \tag{7.87b}$$

The expressions (7.87) give the strengths of the various multipole fields outside the source in terms of integrals over the source densities $\boldsymbol{j}$ and $\boldsymbol{M}$. They can be transformed into more useful forms by means of the following arguments. The results

$$\boldsymbol{L}\cdot\boldsymbol{A} = \mathrm{i}\,\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{A}), \tag{7.88a}$$

$$\boldsymbol{L}\cdot(\nabla\wedge\boldsymbol{A}) = \mathrm{i}\,\nabla^2(\boldsymbol{r}\cdot\boldsymbol{A}) - \mathrm{i}\,\frac{1}{r}\frac{\partial(r^2\,\nabla\cdot\boldsymbol{A})}{\partial r} \tag{7.88b}$$

follow from the definition (7.29) of $\boldsymbol{L}$, and simple vector identities. Substituting into Eq. (7.87a), we obtain

$$a_E(l,m) = -\frac{\mu_0 c\;k^3}{\sqrt{l(l+1)}}\int j_l(kr)\,Y_{lm}^*\left[\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{M}) + \frac{\nabla^2(\boldsymbol{r}\cdot\boldsymbol{j})}{k^2} - \mathrm{i}\,\frac{c}{k\,r}\frac{\partial(r^2\rho)}{\partial r}\right]d^3\boldsymbol{r}, \tag{7.89}$$

where use has been made of Eq. (7.78). Use of Green's theorem on the second term replaces ∇^2 by $-k^2$ (since we can neglect the surface terms, and $j_l(kr)\,Y_{lm}^*$

is a solution of the Helmholtz equation). A radial integration by part on the third term (again neglecting surface terms) casts the radial derivative over onto the spherical Bessel function. The result for the *electric multipole coefficient* is

$$a_E(l,m) = \frac{\mu_0 c\,k^2}{\mathrm{i}\sqrt{l(l+1)}}\int Y_{lm}^*\left[c\rho\,\frac{d[r\,j_l(kr)]}{dr} + \mathrm{i}\,k\,(\boldsymbol{r}\cdot\boldsymbol{j})\,j_l(kr) - \mathrm{i}\,k\,\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{M})\,j_l(kr)\right]d^3\boldsymbol{r}. \tag{7.90}$$

The analogous set of manipulations using Eq. (7.87b) leads to an expression for the *magnetic multipole coefficient*:

$$a_M(l,m) = \frac{\mu_0 c\,k^2}{\mathrm{i}\sqrt{l(l+1)}}\int Y_{lm}^*\left[\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{j})\,j_l(kr) + \nabla\cdot\boldsymbol{M}\,\frac{d[r\,j_l(kr)]}{dr} - k^2\,(\boldsymbol{r}\cdot\boldsymbol{M})\,j_l(kr)\right]d^3\boldsymbol{r}. \tag{7.91}$$

Both the above results are exact, and are valid for arbitrary wavelength and source size.

In the limit in which the source dimensions are very small compared to a wavelength (*i.e.*, $kr \ll 1$) the expressions for the multipole coefficients can be considerably simplified. Using the asymptotic form (7.14a), and keeping only lowest powers in kr for terms involving ρ, $\boldsymbol{j}$, and $\boldsymbol{M}$, we obtain the approximate electric multipole coefficient

$$a_E(l,m) \simeq \frac{\mu_0 c\,k^{l+2}}{\mathrm{i}\,(2l+1)!!}\left(\frac{l+1}{l}\right)^{1/2}(Q_{lm} + Q_{lm}'), \tag{7.92}$$

where the multipole moments are

$$Q_{lm} = \int r^l\,Y_{lm}^*\,c\rho\,d^3\boldsymbol{r}, \tag{7.93a}$$

$$Q_{lm}' = \frac{-\mathrm{i}\,k}{l+1}\int r^l\,Y_{lm}^*\,\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{M})\,d^3\boldsymbol{r}. \tag{7.93b}$$

The moment Q_{lm} has the same form as a conventional electrostatic multipole moment. The moment Q_{lm}' is an induced electric multipole moment due to the magnetization. It is generally a factor kr smaller than the normal moment Q_{lm}. For the magnetic multipole coefficient $a_M(l,m)$ the corresponding long wavelength approximation is

$$a_M(l,m) \simeq \frac{\mu_0 c\,\mathrm{i}\,k^{l+2}}{(2l+1)!!}\left(\frac{l+1}{l}\right)^{1/2}(\mathcal{M}_{lm}+\mathcal{M}'_{lm}), \tag{7.94}$$

where the magnetic multipole moments are

$$\mathcal{M}_{lm} = -\frac{1}{l+1}\int r^l\,Y_{lm}^*\,\nabla\cdot(\boldsymbol{r}\wedge\boldsymbol{j})\,d^3\boldsymbol{r}, \tag{7.95a}$$

$$\mathcal{M}'_{lm} = -\int r^l\,Y_{lm}^*\,\nabla\cdot\boldsymbol{M}\,d^3\boldsymbol{r} \tag{7.95b}$$

Note that for a system with intrinsic magnetization the magnetic moments $\mathcal{M}_{lm}$ and $\mathcal{M}'_{lm}$ are generally of the same order of magnitude.

Thus, in the long wavelength limit the electric multipole fields are determined by the charge density ρ, whereas the magnetic multipole fields are determined by the magnetic moment densities $\boldsymbol{r}\wedge\boldsymbol{j}/2$ and $\boldsymbol{M}$.

7.5 Radiation from a linear centre-fed antenna

As an illustration of the use of a multipole expansion for a source whose dimensions are comparable to a wavelength, consider the radiation from a linear centre-fed antenna. We assume that the antenna lies along the z-axis, and extends from $z=-d/2$ to $z=d/2$. The current flowing along the antenna vanishes at the end points, and is an even function of z. Thus, we can write

$$I(z,t) = I(|z|)\,\mathrm{e}^{-\mathrm{i}\,\omega\,t}, \tag{7.96}$$

where $I(d/2)=0$. Since the current flows radially, $\boldsymbol{r}\wedge\boldsymbol{j}=0$. Furthermore, there is no intrinsic magnetization. Thus, according to Eq. (7.91), all of the magnetic multipole coefficients $a_M(l,m)$ vanish. In order to calculate the electric multipole coefficients $a_E(l,m)$, we need expressions for the charge and current densities. In spherical polar coordinates the current density $\boldsymbol{j}$ can be written in the form

$$\boldsymbol{j}(\boldsymbol{r}) = \hat{\boldsymbol{r}}\,\frac{I(r)}{2\pi r^2}\,\left[\delta(\cos\theta - 1) - \delta(\cos\theta + 1)\right], \tag{7.97}$$

for $r < d/2$, where the delta functions cause the current to flow only upwards and downwards along the z-axis. From the continuity equation (7.78), the charge density is given by

$$\rho(\boldsymbol{r}) = \frac{1}{\mathrm{i}\,\omega}\,\frac{dI(r)}{dr}\left[\frac{\delta(\cos\theta - 1) - \delta(\cos\theta + 1)}{2\pi r^2}\right], \tag{7.98}$$

for $r < d/2$.

These expressions for $\boldsymbol{j}$ and ρ can be substituted into Eq. (7.90) to give

$$\begin{aligned} a_E(l,m) &= \frac{\mu_0 c\, k^2}{2\pi\sqrt{l(l+1)}}\int_0^{d/2} dr\left\{kr\, j_l(kr)\, I(r) - \frac{1}{k}\frac{dI(r)}{dr}\frac{d[r\, j_l(kr)]}{dr}\right\} \\ &\qquad \int d\Omega\, Y_{lm}^*\left[\delta(\cos\theta - 1) - \delta(\cos\theta + 1)\right]. \end{aligned} \tag{7.99}$$

The angular integral has the value

$$\int d\Omega\, Y_{lm}^*\left[\delta(\cos\theta - 1) - \delta(\cos\theta + 1)\right] = 2\pi\,\delta_{m,0}\,\left[Y_{l0}(0) - Y_{l0}(\pi)\right], \tag{7.100}$$

showing that only $m = 0$ multipoles occur. This is hardly surprising given the cylindrical symmetry of the antenna. The $m = 0$ spherical harmonics are even (odd) about $\theta = \pi/2$ for l even (odd). Hence, the only nonvanishing multipoles have l odd. So, the angular integral takes the value

$$\int d\Omega\, Y_{lm}^*\left[\delta(\cos\theta - 1) - \delta(\cos\theta + 1)\right] = \sqrt{4\pi(2l+1)}, \tag{7.101}$$

for l odd and $m = 0$. After some slight rearrangement, Eq. (7.99) can be written

$$\begin{aligned} a_E(l,0) &= \frac{\mu_0 c\, k}{2\pi}\left[\frac{4\pi(2l+1)}{l(l+1)}\right]^{1/2}\int_0^{d/2}\left\{-\frac{d}{dr}\left[r j_l(kr)\,\frac{dI}{dr}\right]\right. \\ &\qquad\qquad \left. + r j_l(kr)\left(\frac{d^2 I}{dr^2} + k^2\, I\right)\right\}\, dr. \end{aligned} \tag{7.102}$$

kd	$a_E(1,0)$	$a_E(3,0)/a_E(1,0)$	$a_E(5,0)/a_E(1,0)$
π	$4\sqrt{6\pi}\,(\mu_0 c\,I/4\pi d)$	4.95×10^{-2}	1.02×10^{-3}
2π	$4\pi\sqrt{6\pi}\,(\mu_0 c\,I/4\pi d)$	0.325	3.09×10^{-2}

Table 3: The first few electric multipole coefficients for a half-wave and a full-wave antenna

In order to evaluate the integral (7.102) we need to specify the current $I(z)$ along the antenna. In the absence of radiation, the sinusoidal time variation at frequency ω implies a sinusoidal space variation with wavenumber $k = \omega/c$. However, the emission of radiation generally modifies the current distribution. The correct current $I(z)$ can only be found be solving a complicated boundary value problem. For the sake of simplicity, we assume that $I(z)$ is a known function; specifically,

$$I(z) = I\,\sin(kd/2 - k\,|z|), \tag{7.103}$$

for $z < d/2$, where I is the peak current. With a sinusoidal current the second term in curly brackets in Eq. (7.102) vanishes. The first term is a perfect differential. Consequently, Eqs. (7.102) and (7.103) yield

$$a_E(l,0) = \frac{\mu_0 c\,I}{\pi d}\left[\frac{4\pi(2l+1)}{l(l+1)}\right]^{1/2}\left(\frac{kd}{2}\right)^2 j_l(kd/2), \tag{7.104}$$

for l odd.

Let us consider the special cases of a half-wave antenna ($kd = \pi$; *i.e.*, the length of the antenna is half a wavelength) and a full-wave antenna ($kd = 2\pi$). For these two values of kd the $l = 1$ coefficient is tabulated in Table 3, along with the relative values for $l = 3$, 5. It is clear from the table that the coefficients decrease rapidly in magnitude as l increases, and that higher l coefficients are more important the larger the source dimensions. However, even for a full-wave antenna it is generally adequate to retain only the $l = 1$ and $l = 3$ coefficients in order to calculate the angular distribution of the radiation. It is certainly adequate to keep only these two harmonics in order to calculate the total power radiated (which depends on the sum of the squares of the coefficients).

In the radiation zone the multipole fields (7.54) reduce to

$$\begin{aligned} c\boldsymbol{B} &\simeq \frac{\mathrm{e}^{\,\mathrm{i}\,(kr-\omega t)}}{kr}\sum_{l,m}(-\mathrm{i})^{l+1}\left[a_E(l,m)\,\boldsymbol{X}_{lm}\right.\\ &\qquad\left.+a_M(l,m)\,\boldsymbol{n}\wedge\boldsymbol{X}_{lm}\right], &(7.105a)\\ \boldsymbol{E} &\simeq c\boldsymbol{B}\wedge\boldsymbol{n}, &(7.105b)\end{aligned}$$

where use has been made of the asymptotic form (7.16). The time-averaged power radiated per unit solid angle is given by

$$\frac{dP}{d\Omega}=\frac{\mathrm{Re}\,(\boldsymbol{n}\cdot\boldsymbol{E}\wedge\boldsymbol{B}^*)\,r^2}{2\mu_0}, \tag{7.106}$$

or

$$\frac{dP}{d\Omega}=\frac{1}{2\,\mu_0 c\,k^2}\left|\sum_{l,m}(-\mathrm{i})^{l+1}\left[a_E(l,m)\,\boldsymbol{X}_{lm}+a_M(l,m)\,\boldsymbol{n}\wedge\boldsymbol{X}_{lm}\right]\right|^2. \tag{7.107}$$

Retaining only the $l=1$ and $l=3$ electric multipole coefficients, the angular distribution of the radiation from the antenna is given by

$$\frac{dP}{d\Omega}=\frac{|a_E(l,0)|^2}{4\,\mu_0 c\,k^2}\left|\boldsymbol{L}Y_{1,0}-\frac{a_E(3,0)}{\sqrt{6}\,a_E(1,0)}\,\boldsymbol{L}Y_{3,0}\right|^2, \tag{7.108}$$

where use has been made of Eq. (7.52). The various factors in the absolute square are

$$\begin{aligned} |\boldsymbol{L}Y_{1,0}|^2 &= \frac{3}{4\pi}\,\sin^2\theta, &(7.109a)\\ |\boldsymbol{L}Y_{3,0}|^2 &= \frac{63}{16\pi}\,\sin^2\theta\,(5\cos^2\theta-1)^2, &(7.109b)\\ (\boldsymbol{L}Y_{1,0})^*\cdot(\boldsymbol{L}Y_{3,0}) &= \frac{3\sqrt{21}}{8\pi}\,\sin^2\theta\,(5\cos^2\theta-1). &(7.109c)\end{aligned}$$

With these angular factors, Eq. (7.108) becomes

$$\frac{dP}{d\Omega}=\lambda\,\frac{3\,\mu_0 c\,I^2}{\pi^3}\,\frac{3\,\sin^2\theta}{8\pi}\left|1-\sqrt{\frac{7}{8}}\frac{a_E(3,0)}{a_E(1,0)}\,(5\cos^2\theta-1)\right|^2, \tag{7.110}$$

where λ equals 1 for a half-wave antenna and $\pi^2/4$ for a full-wave antenna. The coefficient in front of $(5\cos^2\theta - 1)$ is 0.0463 and 0.304 for the half-wave and full-wave antenna, respectively. It turns out that the radiation pattern from the two-term multipole expansion given above is almost indistinguishable from the exact result for the case of a half-wave antenna. For the case of a full-wave antenna the two-term expansion yields a radiation pattern which differs from the exact result by less than 5%.

The total power radiated by the antenna is

$$P = \frac{1}{2\,\mu_0 c\, k^2} \sum_{l\,\mathrm{odd}} |a_E(l,0)|^2, \tag{7.111}$$

where use has been made of Eq. (7.71). It is evident from Table 3 that a two-term multipole expansion gives an accurate expression for the radiated power for both a half-wave and a full-wave antenna. In fact, a one-term multipole expansion gives a fairly accurate result for the case of a half-wave antenna.

It is clear from the above analysis that the multipole expansion converges rapidly when the source dimensions are of order the wavelength of the radiation. It is also clear that if the source dimensions are much less than the wavelength then the multipole expansion is likely to be completely dominated by the term corresponding to the lowest value of l.

7.6 Spherical wave expansion of a vector plane wave

In discussing the scattering or absorption of electromagnetic radiation by localized systems, it is useful to be able to express a plane electromagnetic wave as a superposition of spherical waves.

Consider, first of all, the expansion of a scalar plane wave as a set of scalar spherical waves. This expansion is conveniently obtained from the expansion (7.26) for the Green's function of the scalar Helmholtz equation. Let us take the limit $\boldsymbol{r}' \to \infty$ of this equation. We can make the substitution $|\boldsymbol{r} - \boldsymbol{r}'| \simeq r' - \boldsymbol{n}\cdot\boldsymbol{r}$ on the left-hand-side, where $\boldsymbol{n}$ is a unit vector pointing in the direction of $\boldsymbol{r}'$. On the right-hand side, $r_< = r$ and $r_> = r'$. Furthermore, we can use the asymptotic form (7.16) for $h_l^{(1)}(kr)$. Thus, we obtain

$$\frac{\mathrm{e}^{\,\mathrm{i}\,kr'}}{4\pi\,r'}\,\mathrm{e}^{-\mathrm{i}\,k\boldsymbol{n}\cdot\boldsymbol{r}} = \mathrm{i}\,k\,\frac{\mathrm{e}^{\,\mathrm{i}\,kr'}}{kr'}\sum_{l,m}(-\mathrm{i})^{l+1} j_l(kr)\,Y^*_{lm}(\theta',\varphi')\,Y_{lm}(\theta,\varphi). \tag{7.112}$$

Canceling the factor $\mathrm{e}^{\,\mathrm{i}\,kr'}/r'$ on either side, and taking the complex conjugate, we get the following expansion for a scalar plane wave,

$$\mathrm{e}^{\,\mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{r}} = 4\pi\sum_{l=0}^{\infty}\mathrm{i}^l\,j_l(kr)\sum_{m=-l}^{+l} Y^*_{lm}(\theta,\varphi)\,Y_{lm}(\theta',\varphi'), \tag{7.113}$$

where $\boldsymbol{k}$ is the wave vector with the spherical coordinates k, θ', φ'. The well known *addition theorem* for the spherical harmonics states that

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1}\sum_{m=-1}^{+l} Y^*_{lm}(\theta,\varphi)\,Y_{lm}(\theta',\varphi'), \tag{7.114}$$

where γ is the angle subtended between the vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$. Consequently,

$$\cos\gamma = \cos\theta\,\cos\theta' + \sin\theta\,\sin\theta'\,\cos(\varphi-\varphi'). \tag{7.115}$$

It follows from Eqs. (7.113) and (7.114) that

$$\mathrm{e}^{\,\mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{r}} = \sum_{l=0}^{\infty}\mathrm{i}^l\,(2l+1)\,j_l(kr)\,P_l(\cos\gamma), \tag{7.116}$$

or

$$\mathrm{e}^{\,\mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{r}} = \sum_{l=0}^{\infty}\mathrm{i}^l\sqrt{4\pi\,(2l+1)}\,j_l(kr)\,Y_{l,0}(\gamma), \tag{7.117}$$

since

$$Y_{l,0}(\theta) = \sqrt{\frac{2l+1}{4\pi}}\,P_l(\cos\theta). \tag{7.118}$$

Let us now make an equivalent expansion for a circularly polarized plane wave incident along the z-axis:

$$\boldsymbol{E}(\boldsymbol{r}) = (\hat{\boldsymbol{x}} \pm \mathrm{i}\,\hat{\boldsymbol{y}})\,\mathrm{e}^{\,\mathrm{i}\,kz}, \tag{7.119a}$$

$$c\boldsymbol{B}(\boldsymbol{r}) = \hat{\boldsymbol{z}}\wedge\boldsymbol{E} = \mp\mathrm{i}\,\boldsymbol{E}. \tag{7.119b}$$

Since the plane wave is finite everywhere (including the origin), its multipole expansion (7.54) can only involve the well behaved radial eigenfunctions $j_l(kr)$. Thus,

$$\boldsymbol{E} = \sum_{l,m}\left[a_\pm(l,m)\,j_l(kr)\,\boldsymbol{X}_{lm} + \frac{\mathrm{i}}{k}\,b_\pm(l,m)\,\nabla\wedge j_l(kr)\boldsymbol{X}_{lm}\right], \tag{7.120a}$$

$$c\boldsymbol{B} = \sum_{l,m}\left[\frac{-\mathrm{i}}{k}\,a_\pm(l,m)\,\nabla\wedge j_l(kr)\boldsymbol{X}_{lm} + b_\pm(l,m)\,j_l(kr)\,\boldsymbol{X}_{lm}\right]. \tag{7.120b}$$

To determine the coefficients $a_\pm(l,m)$ and $b_\pm(l,m)$ we make use of a slight generalization of the standard orthogonality properties (7.53) of the vector spherical harmonics:

$$\int [f_l(r)\boldsymbol{X}_{l'm'}]^*\cdot[g_l(r)\boldsymbol{X}_{lm}]\,d\Omega = f_l^*\,g_l\,\delta_{ll'}\,\delta_{mm'}, \tag{7.121a}$$

$$\int [f_l(r)\boldsymbol{X}_{l'm'}]^*\cdot[\nabla\wedge g_l(r)\boldsymbol{X}_{lm}]\,d\Omega = 0. \tag{7.121b}$$

The first of these follows directly from Eq. (7.53a). The second follows from Eqs. (7.31), (7.53b), (7.59), and the identity

$$\nabla = \frac{\boldsymbol{r}}{r}\frac{\partial}{\partial r} - \frac{\mathrm{i}}{r^2}\,\boldsymbol{r}\wedge\boldsymbol{L}. \tag{7.122}$$

The coefficients $a_\pm(l,m)$ and $b_\pm(l,m)$ are obtained by taking the scalar product of Eqs. (7.120) with $\boldsymbol{X}_{lm}^*$ and integrating over all solid angle, making use of the orthogonality relations (7.121). This yields

$$a_\pm(l,m)\,j_l(kr) = \int \boldsymbol{X}_{lm}^*\cdot\boldsymbol{E}\,d\Omega, \tag{7.123a}$$

$$b_\pm(l,m)\,j_l(kr) = \int \boldsymbol{X}_{lm}^*\cdot c\boldsymbol{B}\,d\Omega. \tag{7.123b}$$

Substitution of Eqs. (7.52) and (7.120a) into Eq. (7.123a) gives

$$a_\pm(l,m)\,j_l(kr) = \int \frac{(L_\mp\,Y_{lm})^*}{\sqrt{l(l+1)}}\,\mathrm{e}^{\,\mathrm{i}\,kz}\,d\Omega, \tag{7.124}$$

where the operators $L_\pm$ are defined in Eqs. (7.30). Making use of Eqs. (7.33), the above expression reduces to

$$a_\pm(l,m)\,j_l(kr) = \frac{\sqrt{(l\pm m)(l\mp m+1)}}{\sqrt{l(l+1)}}\int Y^*_{l,m\mp 1}\,\mathrm{e}^{\,\mathrm{i}\,kz}\,d\Omega. \tag{7.125}$$

If the expansion (7.117) is substituted for $\mathrm{e}^{\,\mathrm{i}\,kz}$, and use is made of the orthogonality properties of the spherical harmonics, then we obtain the result

$$a_\pm(l,m) = \mathrm{i}^l\,\sqrt{4\pi\,(2l+1)}\,\delta_{m,\pm 1}. \tag{7.126}$$

It is clear from Eqs. (7.119b) and (7.123b) that

$$b_\pm(l,m) = \mp\,\mathrm{i}\,a_\pm(l,m). \tag{7.127}$$

Thus, the general expansion of a circularly polarized plane wave takes the form

$$\boldsymbol{E}(\boldsymbol{r}) = \sum_{l=1}^{\infty}\mathrm{i}^l\sqrt{4\pi(2l+1)}\left[j_l(kr)\boldsymbol{X}_{l,\pm 1}\pm\frac{1}{k}\,\nabla\wedge j_l(kr)\boldsymbol{X}_{l,\pm 1}\right], \tag{7.128a}$$

$$\boldsymbol{B}(\boldsymbol{r}) = \sum_{l=1}^{\infty}\mathrm{i}^l\sqrt{4\pi(2l+1)}\left[\frac{-\mathrm{i}}{k}\,\nabla\wedge j_l(kr)\boldsymbol{X}_{l,\pm 1}\mp\mathrm{i}\,j_l(kr)\boldsymbol{X}_{l,\pm 1}\right]. \tag{7.128b}$$

The expansion for a linearly polarized plane wave is easily obtained by taking the appropriate linear combination of the above two expansions.

7.7 Mie scattering

Consider a plane electromagnetic wave incident on a spherical obstacle. In general, the wave is scattered, to some extent, by the obstacle. Thus, far away from the sphere the electromagnetic fields can be expressed as the sum of a plane wave and a set of outgoing spherical waves. There may be absorption by the obstacle, as well as scattering. In this case, the energy flow away from the obstacle is less than the total energy flow towards it: the difference represents the absorbed energy.

The fields outside the sphere can be written as the sum of incident and scattered waves:

$$\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{E}_{\text{inc}} + \boldsymbol{E}_{sc}, \tag{7.129a}$$

$$\boldsymbol{B}(\boldsymbol{r}) = \boldsymbol{B}_{\text{inc}} + \boldsymbol{B}_{sc}, \tag{7.129b}$$

where $\boldsymbol{E}_{\text{inc}}$ and $\boldsymbol{B}_{\text{inc}}$ are given by (7.128). Since the scattered fields are outgoing waves at infinity, their expansions must be of the form

$$\begin{aligned}\boldsymbol{E}_{\text{sc}} &= \frac{1}{2}\sum_{l=1}^{\infty} \mathrm{i}^l \sqrt{4\pi\,(2l+1)} \qquad\qquad (7.130\text{a})\\ &\qquad \left[\alpha_{\pm}(l)\, h_l^{(1)}(kr)\, \boldsymbol{X}_{l,\pm 1} \pm \frac{\beta_{\pm}(l)}{k}\, \nabla \wedge h_l^{(1)}(kr)\, \boldsymbol{X}_{l,\pm 1}\right],\\ c\boldsymbol{B}_{\text{sc}} &= \frac{1}{2}\sum_{l=1}^{\infty} \mathrm{i}^l \sqrt{4\pi\,(2l+1)} \qquad\qquad (7.130\text{b})\\ &\qquad \left[\frac{-\mathrm{i}\,\alpha_{\pm}(l)}{k}\, \nabla \wedge h_l^{(1)}(kr)\boldsymbol{X}_{l,\pm 1} \mp \mathrm{i}\,\beta_{\pm}(l)\, h_l^{(1)}(kr)\, \boldsymbol{X}_{l,\pm 1}\right].\end{aligned}$$

The coefficients $\alpha_{\pm}(l)$ and $\beta_{\pm}(l)$ are determined by the boundary conditions on the surface of the sphere. In general, it is necessary to sum over all m harmonics in the above expressions. However, for the restricted class of spherically symmetric scatterers only $m = \pm 1$ harmonics need be retained (since only these harmonics occur in the spherical wave expansion of the incident plane wave (see Eqs. (7.128)), and a spherically symmetric scatterer does not couple different m harmonics).

The angular distribution of the scattered power can be written in terms of the coefficients $\alpha(l)$ and $\beta(l)$ using the scattered electromagnetic fields evaluated on the surface of a sphere of radius a surrounding the scatterer. In fact, it is easily demonstrated that

$$\begin{aligned}\frac{dP_{\rm sc}}{d\Omega} &= \frac{a^2}{2\mu_0}\,{\rm Re}\,[\boldsymbol{n}\cdot\boldsymbol{E}_{\rm sc}\wedge\boldsymbol{B}_{\rm sc}^\ast]_{r=a}\\ &= -\frac{a^2}{2\mu_0}\,{\rm Re}\,[\boldsymbol{E}_{\rm sc}\cdot(\boldsymbol{n}\wedge\boldsymbol{B}_{\rm sc}^\ast)]_{r=a},\end{aligned}\tag{7.131}$$

where $\boldsymbol{n}$ is a radially directed outward normal. The differential scattering cross section is defined as the ratio of $dP_{\rm sc}/d\Omega$ to the incident flux $1/\mu_0 c$. Hence,

$$\frac{d\sigma_{\rm sc}}{d\Omega} = -\frac{a^2}{2}\,{\rm Re}\,[\boldsymbol{E}_{\rm sc}\cdot(\boldsymbol{n}\wedge c\boldsymbol{B}_{\rm sc}^\ast)]_{r=a}.\tag{7.132}$$

We need to evaluate this expression using the electromagnetic fields specified in Eqs. (7.128), (7.129), and (7.130). The following identity, which can be established with the aid of Eqs. (7.29), (7.52), and (7.59), is helpful in this regard:

$$\nabla\wedge f(r)\boldsymbol{X}_{lm} = \boldsymbol{n}\,\frac{{\rm i}\sqrt{l(l+1)}}{r}\,f(r)\,Y_{lm} + \frac{1}{r}\frac{d[r f(r)]}{dr}\,\boldsymbol{n}\wedge\boldsymbol{X}_{lm}.\tag{7.133}$$

For instance, using this result we can write $\boldsymbol{n}\wedge c\boldsymbol{B}_{sc}$ in the form

$$\begin{aligned}\boldsymbol{n}\wedge c\boldsymbol{B}_{\rm sc} &= \frac{1}{2}\sum_{l=1}^{\infty}{\rm i}^l\sqrt{4\pi\,(2l+1)} \qquad (7.134)\\ &\quad\left[\frac{{\rm i}\,\alpha_\pm(l)}{k}\,\frac{1}{r}\,\frac{d[r h_l^{(1)}(kr)]}{dr}\boldsymbol{X}_{l,\pm 1}\mp{\rm i}\,\beta_\pm(l)\,h_l^{(1)}(kr)\,\boldsymbol{n}\wedge\boldsymbol{X}_{l,\pm 1}\right].\end{aligned}$$

It can be demonstrated, after considerable algebra, that

$$\frac{d\sigma_{\rm sc}}{d\Omega} = \frac{\pi}{2k^2}\left|\sum_l\sqrt{2l+1}\,[\alpha_\pm(l)\,\boldsymbol{X}_{l,\pm 1}\pm{\rm i}\,\beta_\pm(l)\,\boldsymbol{n}\wedge\boldsymbol{X}_{l,\pm 1}]\right|^2.\tag{7.135}$$

In obtaining this formula, use has been made of the standard result

$$\frac{df_l(z)}{dz}\,f_{l'}^\ast(z) - \frac{df_{l'}^\ast(z)}{dz}\,f_l(z) = \frac{2\,{\rm i}}{z^2},\tag{7.136}$$

where $f_l(z) = \mathrm{i}^l\, h_l^{(1)}(z)$. The total scattering cross section is obtained by integrating Eq. (7.135) over all solid angle, making use of the following orthogonality relations for the vector spherical harmonics (see Eqs. (7.53)):

$$\int \boldsymbol{X}^*_{l'm'}\cdot\boldsymbol{X}_{lm}\,d\Omega = \delta_{ll'}\,\delta_{mm'}, \tag{7.137a}$$

$$\int \boldsymbol{X}^*_{l'm'}\cdot(\boldsymbol{n}\wedge\boldsymbol{X}_{lm})\,d\Omega = 0, \tag{7.137b}$$

$$\int (\boldsymbol{n}\wedge\boldsymbol{X}^*_{l'm'})\cdot(\boldsymbol{n}\wedge\boldsymbol{X}_{lm})\,d\Omega = \delta_{ll'}\,\delta_{mm'}. \tag{7.137c}$$

Thus,

$$\sigma_{\rm sc} = \frac{\pi}{2k^2}\sum_l (2l+1)\left[|\alpha_\pm(l)|^2+|\beta_\pm(l)|^2\right]. \tag{7.138}$$

According to Eqs. (7.135) and (7.138), the total scattering cross section is independent of the polarization of the incident radiation (*i.e.*, it is the same for both the $\pm$ signs). However, the differential scattering cross section in any particular direction is, in general, different for different circular polarizations of the incident radiation. This implies that if the incident radiation is linearly polarized then the scattered radiation is elliptically polarized. Furthermore, if the incident radiation is unpolarized then the scattered radiation exhibits partial polarization, with the degree of polarization depending on the angle of observation.

The total power absorbed by the sphere is given by

$$\begin{aligned} P_{\rm abs} &= -\frac{a^2}{2\mu_0}\,{\rm Re}\int[\boldsymbol{n}\cdot\boldsymbol{E}\wedge\boldsymbol{B}^*]_{r=a}\,d\Omega \\ &= \frac{a^2}{2\mu_0}\,{\rm Re}\int[\boldsymbol{E}\cdot(\boldsymbol{n}\wedge\boldsymbol{B}^*)]_{r=a}\,d\Omega. \end{aligned} \tag{7.139}$$

A similar calculation to that outlined above yields the following expression for the absorption cross section,

$$\sigma_{\rm abs} = \frac{\pi}{2k^2}\sum_l (2l+1)\left[2-|\alpha_\pm(l)+1|^2-|\beta_\pm(l)+1|^2\right]. \tag{7.140}$$

The total or extinction cross section is the sum of $\sigma_{\rm sc}$ and $\sigma_{\rm abs}$:

$$\sigma_{\rm t} = -\frac{\pi}{k^2}\sum_l (2l+1)\,{\rm Re}\,[\alpha_\pm(l)+\beta_\pm(l)]. \tag{7.141}$$

Not surprisingly, the above expressions for the cross sections closely resemble those obtained in quantum mechanics from partial wave expansions.

Let us now consider the boundary conditions at the surface of the sphere (whose radius is a, say). For the sake of simplicity, let us suppose that the sphere is a perfect conductor. In this case, the appropriate boundary condition is that the tangential electric field is zero at $r = a$. According to Eqs. (7.128), (7.129), and (7.133), the tangential electric field is given by

$$\begin{aligned}\boldsymbol{E}_{\rm tan} &= \sum_l {\rm i}^l\sqrt{4\pi(2l+1)}\left\{\left[j_l + \frac{\alpha_\pm(l)}{2}\,h_l^{(1)}\right]\boldsymbol{X}_{l,\pm 1}\right.\\ &\qquad \left.\pm\,\frac{1}{x}\frac{d}{dx}\left[x\left(j_l+\frac{\beta_\pm(l)}{2}\,h_l^{(1)}\right)\right]\boldsymbol{n}\wedge\boldsymbol{X}_{l,\pm 1}\right\},\end{aligned} \tag{7.142}$$

where $x = ka$, and all of the spherical Bessel functions have the argument x. Thus, the boundary condition yields

$$\alpha_\pm(l)+1 = -\frac{h_l^{(2)}(ka)}{h_l^{(1)}(ka)}, \tag{7.143a}$$

$$\beta_\pm(l)+1 = -\left[\frac{(x\,h_l^{(2)}(x)\,)'}{(x\,h_l^{(1)}(x)\,)'}\right]_{x=ka}, \tag{7.143b}$$

where $'$ denotes d/dx. Note that $\alpha_\pm(l)+1$ and $\beta_\pm(l)+1$ are both numbers of modulus unity. This implies, from Eq. (7.140), that there is no absorption for the case of a perfectly conducting sphere (in general, there is some absorption if the sphere has a finite conductivity). We can write $\alpha_\pm(l)$ and $\beta_\pm(l)$ in the form

$$\alpha_\pm(l) = {\rm e}^{2\,{\rm i}\,\delta_l}-1, \tag{7.144a}$$

$$\beta_\pm(l) = {\rm e}^{2\,{\rm i}\,\delta_l'}-1, \tag{7.144b}$$

where the phase angles δ_l and δ_l' are called *scattering phase shifts*. It follows from Eqs. (7.143) that

$$\tan\delta_l = \frac{j_l(ka)}{y_l(ka)}, \tag{7.145a}$$

$$\tan\delta_l' = \left[\frac{(x\,j_l(x))'}{(x\,y_l(x))'}\right]_{x=ka}. \tag{7.145b}$$

Let us specialize to the limit $ka \ll 1$, in which the wavelength of the radiation is much greater than the radius of the sphere. The asymptotic expansions (7.14) yield

$$\begin{aligned}\alpha_\pm(l) &\simeq -\frac{2\,\mathrm{i}\,(ka)^{2l+1}}{(2l+1)[(2l-1)!!]^2},\\ \beta_\pm(l) &\simeq -\frac{(l+1)}{l}\,\alpha_\pm(l),\end{aligned} \tag{7.146a}$$

for $l \geq 1$. It is clear that the scattering coefficients $\alpha_\pm(l)$ and $\beta_\pm(l)$ become small very rapidly as l increases. In the very long wavelength limit only the $l = 1$ coefficients need be retained. It is easily seen that

$$\alpha_\pm(1) = -\frac{\beta_\pm(1)}{2} \simeq -\frac{2\,\mathrm{i}}{3}\,(ka)^3. \tag{7.147}$$

In this limit, the differential scattering cross section (7.135) reduces to

$$\frac{d\sigma_{\rm sc}}{d\Omega} \simeq \frac{2\pi}{3}\,a^2(ka)^4\,|\boldsymbol{X}_{1,\pm1} \mp 2\,\mathrm{i}\,\boldsymbol{n}\wedge\boldsymbol{X}_{1,\pm1}|^2. \tag{7.148}$$

It can be demonstrated that

$$|\boldsymbol{n}\wedge\boldsymbol{X}_{1,\pm1}|^2 = |\boldsymbol{X}_{1,\pm1}|^2 = \frac{3}{16\pi}\,(1+\cos^2\theta), \tag{7.149}$$

and

$$[\pm\mathrm{i}\,(\boldsymbol{n}\wedge\boldsymbol{X}^*_{1,\pm1})\cdot\boldsymbol{X}_{1,\pm1}] = -\frac{3\pi}{8}\,\cos\theta. \tag{7.150}$$

Thus, in long wavelength limit the differential scattering cross section limits to

$$\frac{d\sigma_{\rm sc}}{d\Omega} \simeq a^2 (ka)^4 \left[\frac{5}{8}\,(1+\cos^2\theta) - \cos\theta\right]. \tag{7.151}$$

The scattering is predominately backwards, and is independent of the state of polarization of the incident radiation. The total scattering cross section is given by

$$\sigma_{\rm sc} = \frac{10\pi}{3}\,a^2\,(ka)^4. \tag{7.152}$$

This well known result was first obtained by Mie and Debye. Note that the cross section scales as the inverse fourth power of the wavelength of the incident radiation. This scaling is generic to all scatterers whose dimensions are much smaller than the wavelength. In fact, it was first derived by Rayleigh using dimensional analysis.