PHY 352K

Classical Electromagnetism

an upper-division undergraduate level lecture course given by

Richard Fitzpatrick

Assistant Professor of Physics

The University of Texas at Austin

Fall 1997

Email: rfitzp@farside.ph.utexas.edu, Tel.: 512-471-9439
Homepage: http://farside.ph.utexas.edu/em1/em.html

1 Introduction

1.1 Major sources

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

Introduction to electrodynamics: D.J. Griffiths, 2nd edition (Prentice Hall, Englewood Cliffs NJ, 1989).

Electromagnetism: I.S. Grant and W.R. Phillips (John Wiley & Sons, Chichester, 1975).

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

The Feynman lectures on physics: R.P. Feynman, R.B. Leighton, and M. Sands, Vol. II (Addison-Wesley, Reading MA, 1964).

1.2 Outline of course

The main topic of this course is *Maxwell's equations*. These are a set of *eight* first order partial differential equations which constitute a *complete* description of electric and magnetic phenomena. To be more exact, Maxwell's equations constitute a complete description of the behaviour of electric and magnetic *fields*. You are all, no doubt, quite familiar with the concepts of electric and magnetic fields, but I wonder how many of you can answer the following question. "Do electric and magnetic fields have a real physical existence or are they just theoretical constructs which we use to calculate the electric and magnetic forces exerted by charged particles on one another?" In trying to formulate an answer to this question we shall, hopefully, come to a better understanding of the nature of electric and magnetic fields and the reasons why it is necessary to use these concepts in order to fully describe electric and magnetic phenomena.

At any given point in space an electric or magnetic field possesses two properties, a *magnitude* and a *direction*. In general, these properties vary from point to point. It is conventional to represent such a field in terms of its components measured with respect to some conveniently chosen set of Cartesian axes (*i.e.*, x, y, and z axes). Of course, the orientation of these axes is *arbitrary*. In other words, different observers may well choose different coordinate axes to describe the same field. Consequently, electric and magnetic fields may have different components according to different observers. We can see that any description of electric and magnetic fields is going to depend on two different things. Firstly, the nature of the fields themselves and, secondly, our arbitrary choice of the coordinate axes with respect to which we measure these fields. Likewise, Maxwell's equations, the equations which describe the behaviour of electric and magnetic fields, depend on two different things. Firstly, the fundamental laws of physics which govern the behaviour of electric and magnetic fields and, secondly, our arbitrary choice of coordinate axes. It would be nice if we could easily distinguish those elements of Maxwell's equations which depend on physics from those which only depend on coordinates. In fact, we can achieve this using what mathematicians call *vector field theory*. This enables us to write Maxwell's equations in a manner which is *completely independent* of our choice of coordinate axes. As an added bonus, Maxwell's equations look a lot simpler when written in a coordinate free manner.

In fact, instead of *eight* first order partial differential equations, we only require four such equations using vector field theory. It should be clear, by now, that we are going to be using a lot of vector field theory in this course. In order to help you with this, I have decided to devote the first few lectures of this course to a review of the basic results of vector field theory. I know that most of you have already taken a course on this topic. However, that course was taught by somebody from the mathematics department. Mathematicians have their own agenda when it comes to discussing vectors. They like to think of vector operations as a sort of algebra which takes place in an abstract "vector space." This is all very well, but it is not always particularly useful. So, when I come to review this topic I shall emphasize those aspects of vectors which make them of particular interest to physicists; namely, the fact that we can use them to write the laws of physics in a coordinate free fashion.

Traditionally, an upper division college level course on electromagnetic theory is organized as follows. First, there is a lengthy discussion of electrostatics (*i.e.*, electric fields generated by stationary charge distributions) and all of its applications. Next, there is a discussion of magnetostatics (*i.e.*, magnetic fields generated by steady current distributions) and all of its applications. At this point, there is usually some mention of the interaction of steady electric and magnetic fields with matter. Next, there is an investigation of induction (*i.e.*, electric and magnetic fields generated by time varying magnetic and electric fields, respectively) and its many applications. Only at this rather late stage in the course is it possible to write down the full set of Maxwell's equations. The course ends with a discussion of electromagnetic waves.

The organization of my course is somewhat different to that described above. There are two reasons for this. Firstly, I do not think that the traditional course emphasizes Maxwell's equations sufficiently. After all, they are only written down in their full glory more than three quarters of the way through the course. I find this a problem because, as I have already mentioned, I think that Maxwell's equations should be the principal topic of an upper division course on electromagnetic theory. Secondly, in the traditional course it is very easy for the lecturer to fall into the trap of dwelling too long on the relatively uninteresting subject matter at the beginning of the course (*i.e.*, electrostatics and magnetostatics) at the expense of the really interesting material towards the end of the course (*i.e.*, induction,

Maxwell's equations, and electromagnetic waves). I vividly remember that this is exactly what happened when I took this course as an undergraduate. I was very disappointed! I had been looking forward to hearing all about Maxwell's equations and electromagnetic waves, and we were only able to cover these topics in a hurried and rather cursory fashion because the lecturer ran out of time at the end of the course.

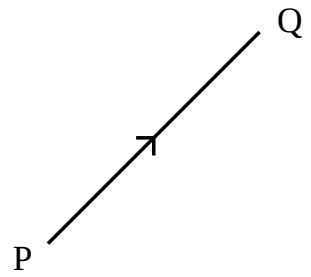
My course is organized as follows. The first section is devoted to *Maxwell's equations*. I shall describe how Maxwell's equations can be derived from the familiar laws of physics which govern electric and magnetic phenomena, such as Coulomb's law and Faraday's law. Next, I shall show that Maxwell's equations possess propagating wave like solutions, called electromagnetic waves, and, furthermore, that light, radio waves, and X-rays, are all different types of electromagnetic wave. Finally, I shall demonstrate that it is possible to write down a formal solution to Maxwell's equations, given a sensible choice of boundary conditions. The second section of my course is devoted to the *applications* of Maxwell's equations. We shall investigate electrostatic fields generated by stationary charge distributions, conductors, resistors, capacitors, inductors, the energy and momentum carried by electromagnetic fields, and the generation and transmission of electromagnetic radiation. This arrangement of material gives the proper emphasis to Maxwell's equations. It also reaches the right balance between the interesting and the more mundane aspects of electromagnetic theory. Finally, it ensures that even if I do run out of time towards the end of the course I shall still have covered Maxwell's equations and electromagnetic waves in adequate detail.

One topic which I am not going to mention at all in my course is the interaction of electromagnetic fields with matter. It is impossible to do justice to this topic at the college level, which is why I always prefer to leave it to graduate school.

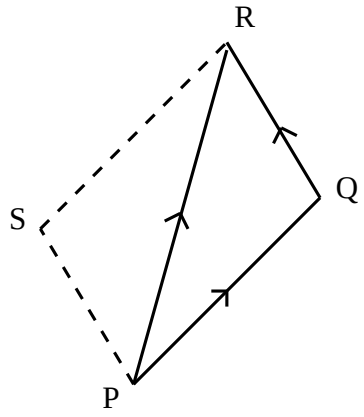
2 Vector assault course

2.1 Vector algebra

In applied mathematics physical quantities are represented by two distinct classes of objects. Some quantities, denoted *scalars*, are represented by *real numbers*. Others, denoted *vectors*, are represented by directed line elements: *e.g.* $\overrightarrow{PQ}$. Note



that line elements (and therefore vectors) are movable and do not carry intrinsic position information. In fact, vectors just possess a magnitude and a direction, whereas scalars possess a magnitude but no direction. By convention, vector quantities are denoted by bold-faced characters (*e.g.* $\boldsymbol{a}$) in typeset documents and by underlined characters (*e.g.* $\underline{a}$) in long-hand. Vectors can be added together but the *same units* must be used, like in scalar addition. Vector addition can be represented using a parallelogram: $\overrightarrow{PR} = \overrightarrow{PQ} + \overrightarrow{QR}$. Suppose that $\boldsymbol{a} \equiv \overrightarrow{PQ} \equiv \overrightarrow{SR}$,



$\boldsymbol{b} \equiv \overrightarrow{QR} \equiv \overrightarrow{PS}$, and $\boldsymbol{c} \equiv \overrightarrow{PR}$. It is clear from the diagram that vector addition is

commutative: *e.g.*, $\boldsymbol{a}+\boldsymbol{b}=\boldsymbol{b}+\boldsymbol{a}$. It can also be shown that the *associative* law holds: *e.g.*, $\boldsymbol{a}+(\boldsymbol{b}+\boldsymbol{c})=(\boldsymbol{a}+\boldsymbol{b})+\boldsymbol{c}$.

There are two approaches to vector analysis. The *geometric* approach is based on line elements in space. The *coordinate* approach assumes that space is defined by Cartesian coordinates and uses these to characterize vectors. In physics we adopt the second approach because we can generalize it to n-dimensional spaces without suffering brain failure. This is necessary in special relativity, where three-dimensional space and one-dimensional time combine to form four-dimensional space-time. The coordinate approach can also be generalized to curved spaces, as is necessary in general relativity.

In the coordinate approach a vector is denoted as the row matrix of its components along each of the Cartesian axes (the x, y, and z axes, say):

$$\boldsymbol{a}\equiv(a_x,a_y,a_z). \tag{2.1}$$

Here, a_x is the x-coordinate of the "head" of the vector minus the x-coordinate of its "tail." If $\boldsymbol{a}\equiv(a_x,a_y,a_z)$ and $\boldsymbol{b}\equiv(b_x,b_y,b_z)$ then vector addition is defined

$$\boldsymbol{a}+\boldsymbol{b}\equiv(a_x+b_x,a_y+b_y,a_z+b_z). \tag{2.2}$$

If $\boldsymbol{a}$ is a vector and n is a scalar then the product of a scalar and a vector is defined

$$n\boldsymbol{a}\equiv(na_x,na_y,na_z). \tag{2.3}$$

It is clear that vector algebra is *distributive* with respect to scalar multiplication: *e.g.*, $n(\boldsymbol{a}+\boldsymbol{b})=n\boldsymbol{a}+n\boldsymbol{b}$.

Unit vectors can be defined in the x, y, and z directions as $\boldsymbol{i}\equiv(1,0,0)$, $\boldsymbol{j}\equiv(0,1,0)$, and $\boldsymbol{k}\equiv(0,0,1)$. Any vector can be written in terms of these unit vectors

$$\boldsymbol{a}=a_x\boldsymbol{i}+a_y\boldsymbol{j}+a_z\boldsymbol{k}. \tag{2.4}$$

In mathematical terminology three vectors used in this manner form a *basis* of the vector space. If the three vectors are mutually perpendicular then they are termed orthogonal basis vectors. In fact, any set of three non-coplanar vectors can be used as basis vectors.

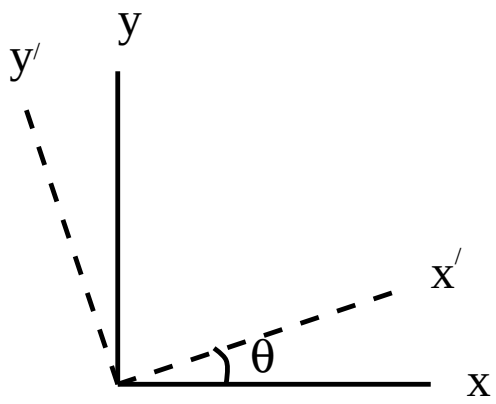
Examples of vectors in physics are displacements from an origin

$$\boldsymbol{r} = (x, y, z) \tag{2.5}$$

and velocities

$$\boldsymbol{v} = \frac{d\boldsymbol{r}}{dt} = \lim_{\delta t \rightarrow 0} \frac{\boldsymbol{r}(t + \delta t) - \boldsymbol{r}(t)}{\delta t}. \tag{2.6}$$

Suppose that we transform to new orthogonal basis, the x', y', and z' axes, which are related to the x, y, and z axes via rotation through an angle θ around the z-axis. In the new basis the coordinates of the general displacement $\boldsymbol{r}$ from the



origin are (x', y', z'). These coordinates are related to the previous coordinates via

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta, \\ y' &= -x\sin\theta + y\cos\theta, \\ z' &= z. \end{aligned} \tag{2.7}$$

We do not need to change our notation for the displacement in the new basis. It is still denoted $\boldsymbol{r}$. The reason for this is that the magnitude and direction of $\boldsymbol{r}$ are *independent* of the choice of basis vectors. The coordinates of $\boldsymbol{r}$ *do* depend on the choice of basis vectors. However, they must depend in a very specific manner [*i.e.*, Eq. (2.7)] which preserves the magnitude and direction of $\boldsymbol{r}$.

Since any vector can be represented as a displacement from an origin (this is just a special case of a directed line element) it follows that the components of a

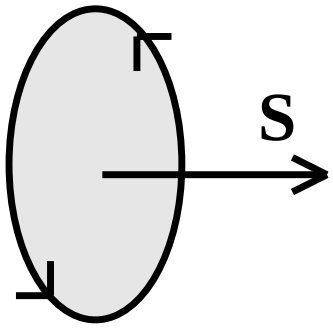
general vector $\boldsymbol{a}$ must transform in an analogous manner to Eq. (2.7). Thus,

$$\begin{aligned} a_{x'} &= a_x \cos\theta + a_y \sin\theta, \\ a_{y'} &= -a_x \sin\theta + a_y \cos\theta, \\ a_{z'} &= a_z, \end{aligned} \tag{2.8}$$

with similar transformation rules for rotation about the y- and z-axes. In the coordinate approach Eq. (2.8) is the definition of a vector. The three quantities (a_x, a_y, a_z) are the components of a vector provided that they transform under rotation like Eq. (2.8). Conversely, (a_x, a_y, a_z) cannot be the components of a vector if they do not transform like Eq. (2.8). Scalar quantities are *invariant* under transformation. Thus, the individual components of a vector (a_x, say) are real numbers but they are *not* scalars. Displacement vectors and all vectors derived from displacements automatically satisfy Eq. (2.8). There are, however, other physical quantities which have both magnitude and direction but which are not obviously related to displacements. We need to check carefully to see whether these quantities are vectors.

2.2 Vector areas

Suppose that we have planar surface of scalar area S. We can define a vector area $\boldsymbol{S}$ whose magnitude is S and whose direction is perpendicular to the plane, in the sense determined by the right-hand grip rule on the rim. This quantity



clearly possesses both magnitude and direction. But is it a true vector? We know that if the normal to the surface makes an angle α_x with the x-axis then the area

seen in the x-direction is $S\cos\alpha_x$. This is the x-component of $\boldsymbol{S}$. Similarly, if the normal makes an angle α_y with the y-axis then the area seen in the y-direction is $S\cos\alpha_y$. This is the y-component of $\boldsymbol{S}$. If we limit ourselves to a surface whose normal is perpendicular to the z-direction then $\alpha_x = \pi/2 - \alpha_y = \alpha$. It follows that $\boldsymbol{S} = S(\cos\alpha, \sin\alpha, 0)$. If we rotate the basis about the z-axis by θ degrees, which is equivalent to rotating the normal to the surface about the z-axis by $-\theta$ degrees, then

$$S_{x'} = S\cos(\alpha-\theta) = S\cos\alpha\cos\theta + S\sin\alpha\sin\theta = S_x\cos\theta + S_y\sin\theta, \qquad (2.9)$$

which is the correct transformation rule for the x-component of a vector. The other components transform correctly as well. This proves that a vector area is a true vector.

According to the vector addition theorem the projected area of two plane surfaces, joined together at a line, in the x direction (say) is the x-component of the sum of the vector areas. Likewise, for many joined up plane areas the projected area in the x-direction, which is the same as the projected area of the rim in the x-direction, is the x-component of the resultant of all the vector areas:

$$\boldsymbol{S} = \sum_i \boldsymbol{S}_i. \qquad (2.10)$$

If we approach a limit, by letting the number of plane facets increase and their area reduce, then we obtain a continuous surface denoted by the resultant vector area:

$$\boldsymbol{S} = \sum_i \delta\boldsymbol{S}_i. \qquad (2.11)$$

It is clear that the projected area of the rim in the x-direction is just S_x. Note that the rim of the surface determines the vector area rather than the nature of the surface. So, two different surfaces sharing the same rim both possess the same vector areas.

In conclusion, a loop (not all in one plane) has a vector area $\boldsymbol{S}$ which is the resultant of the vector areas of any surface ending on the loop. The components of $\boldsymbol{S}$ are the projected areas of the loop in the directions of the basis vectors. As a corollary, a closed surface has $\boldsymbol{S} = \boldsymbol{0}$ since it does not possess a rim.

2.3 The scalar product

A scalar quantity is invariant under all possible rotational transformations. The individual components of a vector are not scalars because they change under transformation. Can we form a scalar out of some combination of the components of one, or more, vectors? Suppose that we were to define the "ampersand" product

$$\boldsymbol{a}\&\boldsymbol{b} = a_x b_y + a_y b_z + a_z b_x = \text{scalar number} \tag{2.12}$$

for general vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Is $\boldsymbol{a}\&\boldsymbol{b}$ invariant under transformation, as must be the case if it is a scalar number? Let us consider an example. Suppose that $\boldsymbol{a} = (1, 0, 0)$ and $\boldsymbol{b} = (0, 1, 0)$. It is easily seen that $\boldsymbol{a}\&\boldsymbol{b} = 1$. Let us now rotate the basis through 45° about the z-axis. In the new basis, $\boldsymbol{a} = (1/\sqrt{2}, -1/\sqrt{2}, 0)$ and $\boldsymbol{b} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$, giving $\boldsymbol{a}\&\boldsymbol{b} = 1/2$. Clearly, $\boldsymbol{a}\&\boldsymbol{b}$ is not invariant under rotational transformation, so the above definition is a bad one.

Consider, now, the *dot product* or scalar product:

$$\boldsymbol{a}\cdot\boldsymbol{b} = a_x b_x + a_y b_y + a_z b_z = \text{scalar number}. \tag{2.13}$$

Let us rotate the basis though θ degrees about the z-axis. According to Eq. (2.8), in the new basis $\boldsymbol{a}\cdot\boldsymbol{b}$ takes the form

$$\begin{aligned} \boldsymbol{a}\cdot\boldsymbol{b} &= (a_x\cos\theta + a_y\sin\theta)(b_x\cos\theta + b_y\sin\theta) \\ &\quad +(-a_x\sin\theta + a_y\cos\theta)(-b_x\sin\theta + b_y\cos\theta) + a_z b_z \\ &= a_x b_x + a_y b_y + a_z b_z. \end{aligned} \tag{2.14}$$

Thus, $\boldsymbol{a}\cdot\boldsymbol{b}$ is invariant under rotation about the z-axis. It can easily be shown that it is also invariant under rotation about the x- and y-axes. Clearly, $\boldsymbol{a}\cdot\boldsymbol{b}$ is a true scalar, so the above definition is a good one. Incidentally, $\boldsymbol{a}\cdot\boldsymbol{b}$ is the only simple combination of the components of two vectors which transforms like a scalar. It is easily shown that the dot product is commutative and distributive:

$$\begin{aligned} \boldsymbol{a}\cdot\boldsymbol{b} &= \boldsymbol{b}\cdot\boldsymbol{a}, \\ \boldsymbol{a}\cdot(\boldsymbol{b}+\boldsymbol{c}) &= \boldsymbol{a}\cdot\boldsymbol{b} + \boldsymbol{a}\cdot\boldsymbol{c}. \end{aligned} \tag{2.15}$$

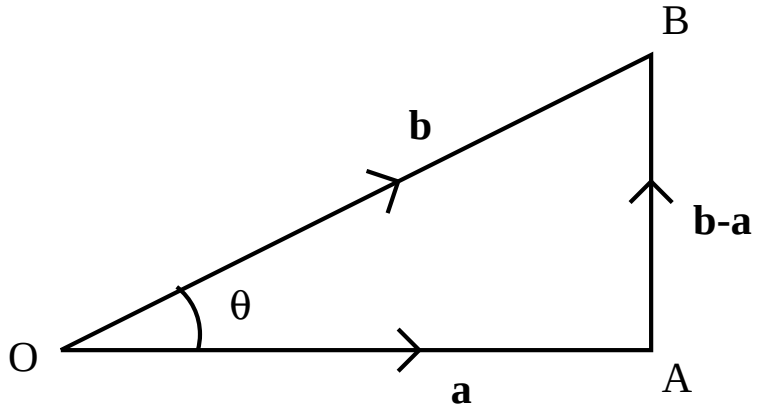
The associative property is meaningless for the dot product because we cannot have $(\boldsymbol{a}\cdot\boldsymbol{b})\cdot\boldsymbol{c}$ since $\boldsymbol{a}\cdot\boldsymbol{b}$ is scalar.

We have shown that the dot product $\boldsymbol{a}\cdot\boldsymbol{b}$ is coordinate independent. But what is the physical significance of this? Consider the special case where $\boldsymbol{a}=\boldsymbol{b}$. Clearly,

$$\boldsymbol{a}\cdot\boldsymbol{b} = {a_x}^2 + {a_y}^2 + {a_z}^2 = \text{Length } (OP)^2, \tag{2.16}$$

if $\boldsymbol{a}$ is the position vector of P relative to the origin O. So, the invariance of $\boldsymbol{a}\cdot\boldsymbol{a}$ is equivalent to the invariance of the length, or magnitude, of vector $\boldsymbol{a}$ under transformation. The length of vector $\boldsymbol{a}$ is usually denoted $|a|$ ("the modulus of a") or sometimes just a, so

$$\boldsymbol{a}\cdot\boldsymbol{a} = |a|^2 = a^2. \tag{2.17}$$



Let us now investigate the general case. The length squared of AB is

$$(\boldsymbol{b}-\boldsymbol{a})\cdot(\boldsymbol{b}-\boldsymbol{a}) = |a|^2 + |b|^2 - 2\,\boldsymbol{a}\cdot\boldsymbol{b}. \tag{2.18}$$

However, according to the "cosine rule" of trigonometry

$$(AB)^2 = (OA)^2 + (OB)^2 - 2(OA)(OB)\cos\theta, \tag{2.19}$$

where (AB) denotes the length of side AB. It follows that

$$\boldsymbol{a}\cdot\boldsymbol{b} = |a||b|\cos\theta. \tag{2.20}$$

Clearly, the invariance of $\boldsymbol{a}\cdot\boldsymbol{b}$ under transformation is equivalent to the invariance of the angle subtended between the two vectors. Note that if $\boldsymbol{a}\cdot\boldsymbol{b}=0$ then either

$|a| = 0$, $|b| = 0$, or the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are perpendicular. The angle subtended between two vectors can easily be obtained from the dot product:

$$\cos\theta = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{|a||b|}. \tag{2.21}$$

The work W performed by a force $\boldsymbol{F}$ moving through a displacement $\boldsymbol{r}$ is the product of the magnitude of $\boldsymbol{F}$ times the displacement in the direction of $\boldsymbol{F}$. If the angle subtended between $\boldsymbol{F}$ and $\boldsymbol{r}$ is θ then

$$W = |F|(|r|\cos\theta) = \boldsymbol{F}\cdot\boldsymbol{r}. \tag{2.22}$$

The rate of flow of liquid of constant velocity $\boldsymbol{v}$ through a loop of vector area $\boldsymbol{S}$ is the product of the magnitude of the area times the component of the velocity perpendicular to the loop. Thus,

$$\text{Rate of flow} = \boldsymbol{v}\cdot\boldsymbol{S}. \tag{2.23}$$

2.4 The vector product

We have discovered how to construct a scalar from the components of two general vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Can we also construct a vector which is not just a linear combination of $\boldsymbol{a}$ and $\boldsymbol{b}$? Consider the following definition:

$$\boldsymbol{a}\mathrm{x}\boldsymbol{b} = (a_x b_x, a_y b_y, a_z b_z). \tag{2.24}$$

Is $\boldsymbol{a}\mathrm{x}\boldsymbol{b}$ a proper vector? Suppose that $\boldsymbol{a} = (1,0,0)$, $\boldsymbol{b} = (0,1,0)$. Clearly, $\boldsymbol{a}\mathrm{x}\boldsymbol{b} = \boldsymbol{0}$. However, if we rotate the basis through 45° about the z-axis then $\boldsymbol{a} = (1/\sqrt{2}, -1/\sqrt{2}, 0)$, $\boldsymbol{b} = (1/\sqrt{2}, 1/\sqrt{2}, 0)$, and $\boldsymbol{a}\mathrm{x}\boldsymbol{b} = (1/2, -1/2, 0)$. Thus, $\boldsymbol{a}\mathrm{x}\boldsymbol{b}$ does not transform like a vector because its magnitude depends on the choice of axis. So, above definition is a bad one.

Consider, now, the *cross product* or vector product:

$$\boldsymbol{a}\wedge\boldsymbol{b} = (a_y b_z - a_z b_y, a_z b_x - a_x b_z, a_x b_y - a_y b_x) = \boldsymbol{c}. \tag{2.25}$$

Does this rather unlikely combination transform like a vector? Let us try rotating the basis through θ degrees about the z-axis using Eq. (2.8). In the new basis

$$\begin{aligned} c_{x'} &= (-a_x \sin\theta + a_y \cos\theta) b_z - a_z(-b_x \sin\theta + b_y \cos\theta) \\ &= (a_y b_z - a_z b_y)\cos\theta + (a_z b_x - a_x b_z)\sin\theta \\ &= c_x \cos\theta + c_y \sin\theta. \end{aligned} \tag{2.26}$$

Thus, the x-component of $\boldsymbol{a}\wedge\boldsymbol{b}$ transforms correctly. It can easily be shown that the other components transform correctly as well. Thus, $\boldsymbol{a}\wedge\boldsymbol{b}$ is a proper vector. The cross product is *anticommutative*:

$$\boldsymbol{a}\wedge\boldsymbol{b} = -\boldsymbol{b}\wedge\boldsymbol{a}, \tag{2.27}$$

distributive:

$$\boldsymbol{a}\wedge(\boldsymbol{b}+\boldsymbol{c}) = \boldsymbol{a}\wedge\boldsymbol{b} + \boldsymbol{a}\wedge\boldsymbol{c}, \tag{2.28}$$

but is *not* associative:

$$\boldsymbol{a}\wedge(\boldsymbol{b}\wedge\boldsymbol{c}) \neq (\boldsymbol{a}\wedge\boldsymbol{b})\wedge\boldsymbol{c}. \tag{2.29}$$

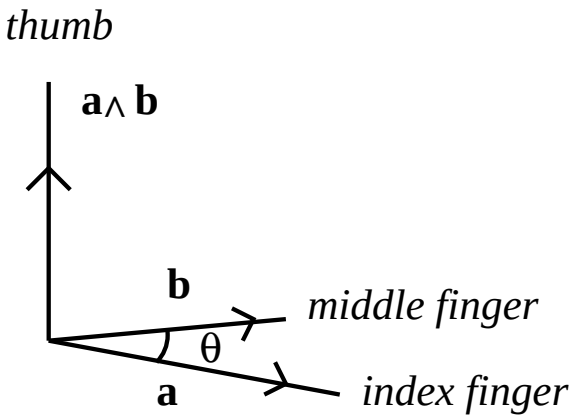
The cross product transforms like a vector, which means that it must have a well defined direction and magnitude. We can show that $\boldsymbol{a}\wedge\boldsymbol{b}$ is *perpendicular* to both $\boldsymbol{a}$ and $\boldsymbol{b}$. Consider $\boldsymbol{a}\cdot\boldsymbol{a}\wedge\boldsymbol{b}$. If this is zero then the cross product must be perpendicular to $\boldsymbol{a}$. Now

$$\begin{aligned} \boldsymbol{a}\cdot\boldsymbol{a}\wedge\boldsymbol{b} &= a_x(a_y b_z - a_z b_y) + a_y(a_z b_x - a_x b_z) + a_z(a_x b_y - a_y b_x) \\ &= 0. \end{aligned} \tag{2.30}$$

Therefore, $\boldsymbol{a}\wedge\boldsymbol{b}$ is perpendicular to $\boldsymbol{a}$. Likewise, it can be demonstrated that $\boldsymbol{a}\wedge\boldsymbol{b}$ is perpendicular to $\boldsymbol{b}$. The vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{a}\wedge\boldsymbol{b}$ form a *right-handed* set like the unit vectors $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$: $\boldsymbol{i}\wedge\boldsymbol{j} = \boldsymbol{k}$. This defines a unique direction for $\boldsymbol{a}\wedge\boldsymbol{b}$, which is obtained from the right-hand rule.

Let us now evaluate the magnitude of $\boldsymbol{a}\wedge\boldsymbol{b}$. We have

$$\begin{aligned} (\boldsymbol{a}\wedge\boldsymbol{b})^2 &= (a_y b_z - a_z b_y)^2 + (a_z b_x - a_x b_z)^2 + (a_x b_z - a_y b_x)^2 \\ &= (a_x^{\,2} + a_y^{\,2} + a_z^{\,2})(b_x^{\,2} + b_y^{\,2} + b_z^{\,2}) - (a_x b_x + a_y b_y + a_z b_z)^2 \\ &= |a|^2|b|^2 - (\boldsymbol{a}\cdot\boldsymbol{b})^2 \\ &= |a|^2|b|^2 - |a|^2|b|^2\cos^2\theta = |a|^2|b|^2\sin^2\theta. \end{aligned} \tag{2.31}$$

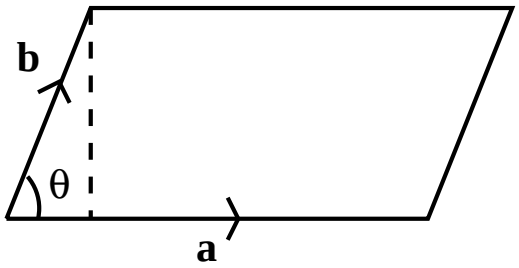


Thus,

$$|\boldsymbol{a} \wedge \boldsymbol{b}| = |a||b| \sin\theta. \tag{2.32}$$

Clearly, $\boldsymbol{a} \wedge \boldsymbol{a} = \boldsymbol{0}$ for any vector, since θ is always zero in this case. Also, if $\boldsymbol{a} \wedge \boldsymbol{b} = \boldsymbol{0}$ then either $|a| = 0$, $|b| = 0$, or $\boldsymbol{b}$ is parallel (or antiparallel) to $\boldsymbol{a}$.

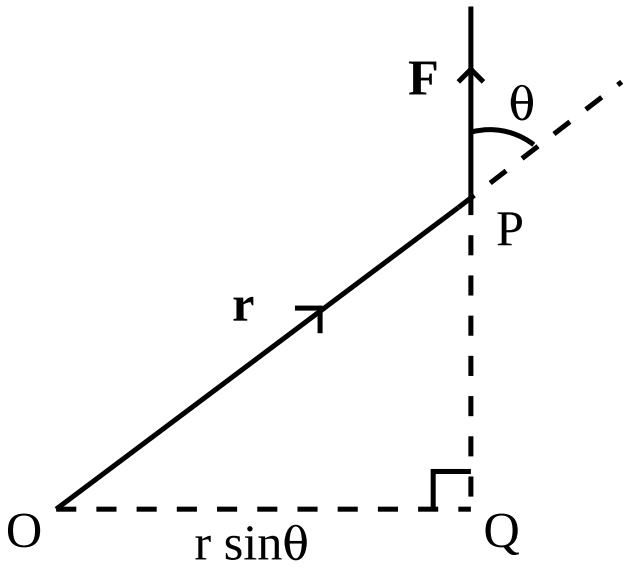
Consider the parallelogram defined by vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The scalar area is $ab\sin\theta$. The vector area has the magnitude of the scalar area and is normal to the plane of the parallelogram, which means that it is perpendicular to both $\boldsymbol{a}$ and $\boldsymbol{b}$. Clearly, the vector area is given by



$$\boldsymbol{S} = \boldsymbol{a} \wedge \boldsymbol{b}, \tag{2.33}$$

with the sense obtained from the right-hand grip rule by rotating $\boldsymbol{a}$ on to $\boldsymbol{b}$.

Suppose that a force $\boldsymbol{F}$ is applied at position $\boldsymbol{r}$. The moment about the origin O is the product of the magnitude of the force and the length of the lever arm OQ. Thus, the magnitude of the moment is $|F||r|\sin\theta$. The direction of a moment is conventionally the direction of the axis through O about which the force tries



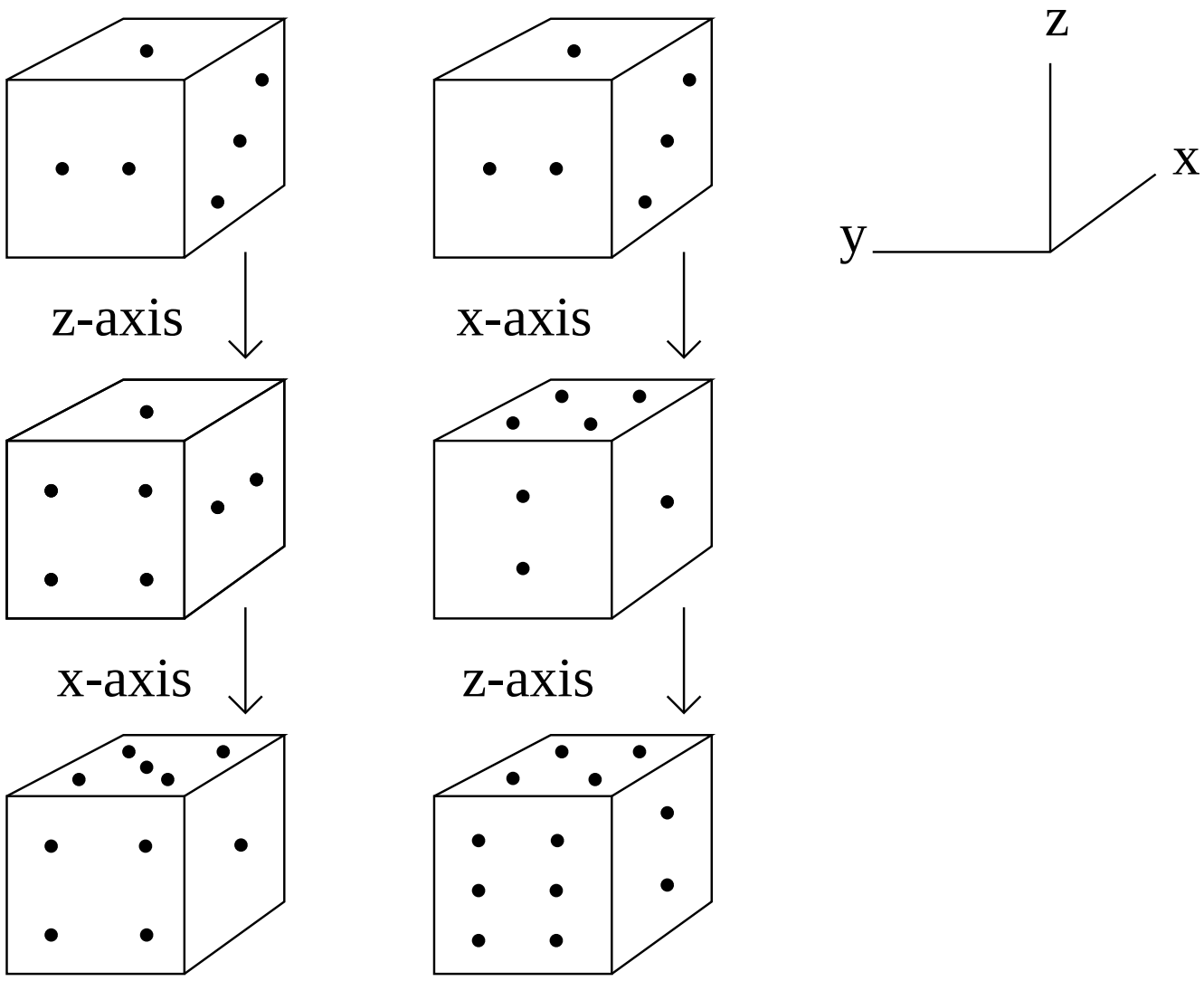
to rotate objects, in the sense determined by the right-hand grip rule. It follows that the vector moment is given by

$$\boldsymbol{M} = \boldsymbol{r} \wedge \boldsymbol{F}. \tag{2.34}$$

2.5 Rotation

Let us try to define a rotation vector $\boldsymbol{\theta}$ whose magnitude is the angle of the rotation, θ, and whose direction is the axis of the rotation, in the sense determined by the right-hand grip rule. Is this a good vector? The short answer is, no. The problem is that the addition of rotations is not commutative, whereas vector addition is. The diagram shows the effect of applying two successive 90° rotations, one about x-axis, and the other about the z-axis, to a six-sided die. In the left-hand case the z-rotation is applied before the x-rotation, and *vice versa* in the right-hand case. It can be seen that the die ends up in two completely different states. Clearly, the z-rotation plus the x-rotation does not equal the x-rotation plus the z-rotation. This non-commuting algebra cannot be represented by vectors. So, although rotations have a well defined magnitude and direction they are not vector quantities.

But, this is not quite the end of the story. Suppose that we take a general



vector $\boldsymbol{a}$ and rotate it about the z-axis by a *small* angle $\delta\theta_z$. This is equivalent to rotating the basis about the z-axis by $-\delta\theta_z$. According to Eq. (2.8) we have

$$\boldsymbol{a}' \simeq \boldsymbol{a} + \delta\theta_z \boldsymbol{k} \wedge \boldsymbol{a}, \tag{2.35}$$

where use has been made of the small angle expansions $\sin\theta \simeq \theta$ and $\cos\theta \simeq 1$. The above equation can easily be generalized to allow small rotations about the x- and y-axes by $\delta\theta_x$ and $\delta\theta_y$, respectively. We find that

$$\boldsymbol{a}' \simeq \boldsymbol{a} + \delta\boldsymbol{\theta} \wedge \boldsymbol{a}, \tag{2.36}$$

where

$$\delta\boldsymbol{\theta} = \delta\theta_x \boldsymbol{i} + \delta\theta_y \boldsymbol{j} + \delta\theta_z \boldsymbol{k}. \tag{2.37}$$

Clearly, we can define a rotation vector $\delta\boldsymbol{\theta}$, but it only works for *small* angle rotations (*i.e.*, sufficiently small that the small angle expansions of sine and cosine are good). According to the above equation, a small z-rotation plus a small x-rotation is (approximately) equal to the two rotations applied in the opposite order. The fact that infinitesimal rotation is a vector implies that angular velocity,

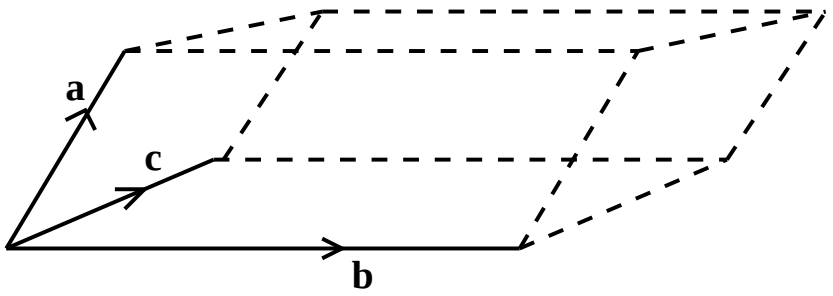
$$\boldsymbol{\omega} = \lim_{\delta t \rightarrow 0} \frac{\delta\boldsymbol{\theta}}{\delta t}, \tag{2.38}$$

must be a vector as well. If $\boldsymbol{a}'$ is interpreted as $\boldsymbol{a}(t+\delta t)$ in the above equation then it is clear that the equation of motion of a vector precessing about the origin with angular velocity $\boldsymbol{\omega}$ is

$$\frac{d\boldsymbol{a}}{dt} = \boldsymbol{\omega} \wedge \boldsymbol{a}. \tag{2.39}$$

2.6 The scalar triple product

Consider three vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$. The scalar triple product is defined $\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}$. Now, $\boldsymbol{b} \wedge \boldsymbol{c}$ is the vector area of the parallelogram defined by $\boldsymbol{b}$ and $\boldsymbol{c}$. So, $\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}$ is the scalar area of this parallelogram times the component of $\boldsymbol{a}$ in the direction of its normal. It follows that $\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}$ is the *volume* of the parallelepiped defined by vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$. This volume is independent of how the triple product is



formed from $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, except that

$$\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c} = -\boldsymbol{a} \cdot \boldsymbol{c} \wedge \boldsymbol{b}. \tag{2.40}$$

So, the "volume" is positive if $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ form a right-handed set (*i.e.*, if $\boldsymbol{a}$ lies above the plane of $\boldsymbol{b}$ and $\boldsymbol{c}$, in the sense determined from the right-hand grip rule by rotating $\boldsymbol{b}$ on to $\boldsymbol{c}$) and negative if they form a left-handed set. The triple product is unchanged if the dot and cross product operators are interchanged:

$$\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c} = \boldsymbol{a} \wedge \boldsymbol{b} \cdot \boldsymbol{c}. \tag{2.41}$$

The triple product is also invariant under any cyclic permutation of $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$,

$$\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c} = \boldsymbol{b} \cdot \boldsymbol{c} \wedge \boldsymbol{a} = \boldsymbol{c} \cdot \boldsymbol{a} \wedge \boldsymbol{b}, \tag{2.42}$$

but any anti-cyclic permutation causes it to change sign,

$$\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c} = -\boldsymbol{b} \cdot \boldsymbol{a} \wedge \boldsymbol{c}. \tag{2.43}$$

The scalar triple product is zero if any two of $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are parallel, or if $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are co-planar.

If $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are non-coplanar, then any vector $\boldsymbol{r}$ can be written in terms of them:

$$\boldsymbol{r} = \alpha\boldsymbol{a} + \beta\boldsymbol{b} + \gamma\boldsymbol{c}. \tag{2.44}$$

Forming the dot product of this equation with $\boldsymbol{b} \wedge \boldsymbol{c}$ then we obtain

$$\boldsymbol{r} \cdot \boldsymbol{b} \wedge \boldsymbol{c} = \alpha\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}, \tag{2.45}$$

so

$$\alpha = \frac{\boldsymbol{r} \cdot \boldsymbol{b} \wedge \boldsymbol{c}}{\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}}. \tag{2.46}$$

Analogous expressions can be written for β and γ. The parameters α, β, and γ are uniquely determined provided $\boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c} \neq 0$; *i.e.*, provided that the three basis vectors are not co-planar.

2.7 The vector triple product

For three vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ the vector triple product is defined $\boldsymbol{a} \wedge (\boldsymbol{b} \wedge \boldsymbol{c})$. The brackets are important because $\boldsymbol{a} \wedge (\boldsymbol{b} \wedge \boldsymbol{c}) \neq (\boldsymbol{a} \wedge \boldsymbol{b}) \wedge \boldsymbol{c}$. In fact, it can be demonstrated that

$$\boldsymbol{a} \wedge (\boldsymbol{b} \wedge \boldsymbol{c}) \equiv (\boldsymbol{a} \cdot \boldsymbol{c})\boldsymbol{b} - (\boldsymbol{a} \cdot \boldsymbol{b})\boldsymbol{c} \tag{2.47}$$

and

$$(\boldsymbol{a}\wedge\boldsymbol{b})\wedge\boldsymbol{c}\equiv(\boldsymbol{a}\cdot\boldsymbol{c})\boldsymbol{b}-(\boldsymbol{b}\cdot\boldsymbol{c})\boldsymbol{a}. \tag{2.48}$$

Let us try to prove the first of the above theorems. The left-hand side and the right-hand side are both proper vectors, so if we can prove this result in one particular coordinate system then it must be true in general. Let us take convenient axes such that the x-axis lies along $\boldsymbol{b}$, and $\boldsymbol{c}$ lies in the x-y plane. It follows that $\boldsymbol{b}=(b_x,0,0)$, $\boldsymbol{c}=(c_x,c_y,0)$, and $\boldsymbol{a}=(a_x,a_y,a_z)$. The vector $\boldsymbol{b}\wedge\boldsymbol{c}$ is directed along the z-axis: $\boldsymbol{b}\wedge\boldsymbol{c}=(0,0,b_xc_y)$. It follows that $\boldsymbol{a}\wedge(\boldsymbol{b}\wedge\boldsymbol{c})$ lies in the x-y plane: $\boldsymbol{a}\wedge(\boldsymbol{b}\wedge\boldsymbol{c})=(a_xb_xc_y,-a_xb_xc_y,0)$. This is the left-hand side of Eq. (2.47) in our convenient axes. To evaluate the right-hand side we need $\boldsymbol{a}\cdot\boldsymbol{c}=a_xc_x+a_yc_y$ and $\boldsymbol{a}\cdot\boldsymbol{b}=a_xb_x$. It follows that the right-hand side is

$$\begin{aligned}\text{RHS} &= (\,(a_xc_x+a_yc_y)b_x,0,0)-(a_xb_xc_x,a_xb_xc_y,0)\\ &= (a_yc_yb_x,-a_xb_xc_y,0)=\text{LHS},\end{aligned} \tag{2.49}$$

which proves the theorem.

2.8 Vector calculus

Suppose that vector $\boldsymbol{a}$ varies with time, so that $\boldsymbol{a}=\boldsymbol{a}(t)$. The time derivative of the vector is defined

$$\frac{d\boldsymbol{a}}{dt}=\lim_{\delta t\rightarrow 0}\left[\frac{\boldsymbol{a}(t+\delta t)-\boldsymbol{a}(t)}{\delta t}\right]. \tag{2.50}$$

When written out in component form this becomes

$$\frac{d\boldsymbol{a}}{dt}=\left(\frac{da_x}{dt},\frac{da_y}{dt},\frac{da_z}{dt}\right). \tag{2.51}$$

Note that $d\boldsymbol{a}/dt$ is often written in shorthand as $\dot{\boldsymbol{a}}$.

Suppose that $\boldsymbol{a}$ is, in fact, the product of a scalar $\phi(t)$ and another vector $\boldsymbol{b}(t)$. What now is the time derivative of $\boldsymbol{a}$? We have

$$\frac{da_x}{dt}=\frac{d}{dt}\,(\phi b_x)=\frac{d\phi}{dt}\,b_x+\phi\,\frac{db_x}{dt}, \tag{2.52}$$

which implies that

$$\frac{d\boldsymbol{a}}{dt} = \frac{d\phi}{dt}\,\boldsymbol{b} + \phi\,\frac{d\boldsymbol{b}}{dt}. \tag{2.53}$$

It is easily demonstrated that

$$\frac{d}{dt}\,(\boldsymbol{a}\cdot\boldsymbol{b}) = \dot{\boldsymbol{a}}\cdot\boldsymbol{b} + \boldsymbol{a}\cdot\dot{\boldsymbol{b}}. \tag{2.54}$$

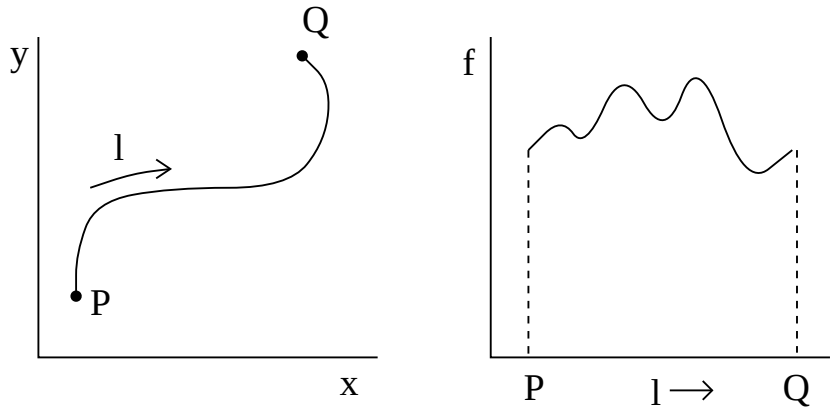
Likewise,

$$\frac{d}{dt}\,(\boldsymbol{a}\wedge\boldsymbol{b}) = \dot{\boldsymbol{a}}\wedge\boldsymbol{b} + \boldsymbol{a}\wedge\dot{\boldsymbol{b}}. \tag{2.55}$$

It can be seen that the laws of vector differentiation are analogous to those in conventional calculus.

2.9 Line integrals

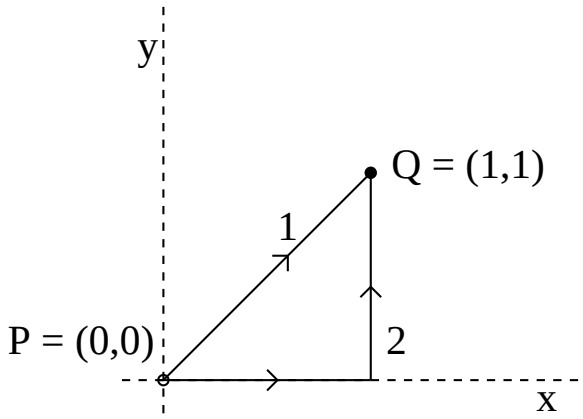
Consider a two-dimensional function $f(x,y)$ which is defined for all x and y. What is meant by the integral of f along a given curve from P to Q in the x-y



plane? We first draw out f as a function of length l along the path. The integral is then simply given by

$$\int_P^Q f(x,y)\,dl = \text{Area under the curve.} \tag{2.56}$$

As an example of this, consider the integral of $f(x, y) = xy$ between P and Q along the two routes indicated in the diagram below. Along route 1 we have



$x = y$, so $dl = \sqrt{2}\, dx$. Thus,

$$\int_P^Q xy\, dl = \int_0^1 x^2\, \sqrt{2}\, dx = \frac{\sqrt{2}}{3}. \tag{2.57}$$

The integration along route 2 gives

$$\begin{aligned} \int_P^Q xy\, dl &= \int_0^1 xy\, dx\Big|_{y=0} + \int_0^1 xy\, dy\Big|_{x=1} \\ &= 0 + \int_0^1 y\, dy = \frac{1}{2}. \end{aligned} \tag{2.58}$$

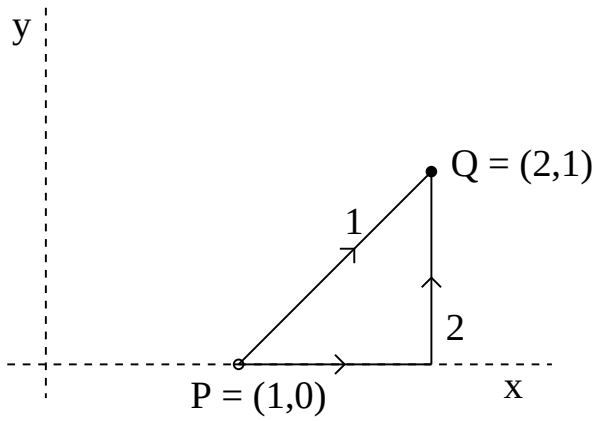
Note that the integral depends on the route taken between the initial and final points.

The most common type of line integral is where the contributions from dx and dy are evaluated separately, rather that through the path length dl;

$$\int_P^Q \left[f(x, y)\, dx + g(x, y)\, dy \right]. \tag{2.59}$$

As an example of this consider the integral

$$\int_P^Q \left[y^3\, dx + x\, dy \right] \tag{2.60}$$



along the two routes indicated in the diagram below. Along route 1 we have $x = y + 1$ and $dx = dy$, so

$$\int_P^Q = \int_0^1 \left(y^3\,dy + (y+1)\,dy\right) = \frac{7}{4}. \tag{2.61}$$

Along route 2

$$\int_P^Q = \int_1^2 y^3\,dx\Big|_{y=0} + \int_0^1 x\,dy\Big|_{x=2} = 2. \tag{2.62}$$

Again, the integral depends on the path of integration.

Suppose that we have a line integral which does not depend on the path of integration. It follows that

$$\int_P^Q (f\,dx + g\,dy) = F(Q) - F(P) \tag{2.63}$$

for some function F. Given $F(P)$ for one point P in the x-y plane, then

$$F(Q) = F(P) + \int_P^Q (f\,dx + g\,dy) \tag{2.64}$$

defines $F(Q)$ for all other points in the plane. We can then draw a contour map of $F(x, y)$. The line integral between points P and Q is simply the change in height in the contour map between these two points:

$$\int_P^Q (f\,dx + g\,dy) = \int_P^Q dF(x, y) = F(Q) - F(P). \tag{2.65}$$

Thus,

$$dF(x,y) = f(x,y)\,dx + g(x,y)\,dy. \tag{2.66}$$

For instance, if $F = xy^3$ then $dF = y^3\,dx + 3xy^2\,dy$ and

$$\int_P^Q \left(y^3\,dx + 3xy^2\,dy\right) = \left[xy^3\right]_P^Q \tag{2.67}$$

is independent of the path of integration.

It is clear that there are two distinct types of line integral. Those that depend only on their endpoints and not on the path of integration, and those which depend both on their endpoints and the integration path. Later on, we shall learn how to distinguish between these two types.

2.10 Vector line integrals

A *vector field* is defined as a set of vectors associated with each point in space. For instance, the velocity $\boldsymbol{v}(\boldsymbol{r})$ in a moving liquid (*e.g.*, a whirlpool) constitutes a vector field. By analogy, a *scalar field* is a set of scalars associated with each point in space. An example of a scalar field is the temperature distribution $T(\boldsymbol{r})$ in a furnace.

Consider a general vector field $\boldsymbol{A}(\boldsymbol{r})$. Let $d\boldsymbol{l} = (dx, dy, dz)$ be the vector element of line length. Vector line integrals often arise as

$$\int_P^Q \boldsymbol{A}\cdot d\boldsymbol{l} = \int_P^Q (A_x\,dx + A_y\,dy + A_z\,dz). \tag{2.68}$$

For instance, if $\boldsymbol{A}$ is a force then the line integral is the work done in going from P to Q.

As an example, consider the work done in a repulsive, inverse square law, central field $\boldsymbol{F} = -\boldsymbol{r}/|r^3|$. The element of work done is $dW = \boldsymbol{F}\cdot d\boldsymbol{l}$. Take $P = (\infty, 0, 0)$ and $Q = (a, 0, 0)$. Route 1 is along the x-axis, so

$$W = \int_\infty^a \left(-\frac{1}{x^2}\right) dx = \left[\frac{1}{x}\right]_\infty^a = \frac{1}{a}. \tag{2.69}$$

The second route is, firstly, around a large circle (r = constant) to the point (a, ∞, 0) and then parallel to the y-axis. In the first part no work is done since $\boldsymbol{F}$ is perpendicular to $d\boldsymbol{l}$. In the second part

$$W = \int_{\infty}^{0} \frac{-y\,dy}{(a^2+y^2)^{3/2}} = \left[\frac{1}{(y^2+a^2)^{1/2}}\right]_0^{\infty} = \frac{1}{a}. \tag{2.70}$$

In this case the integral is independent of path (which is just as well!).

2.11 Surface integrals

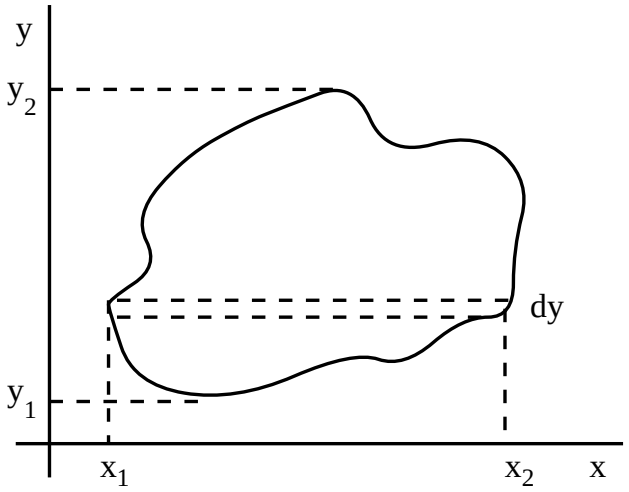
Let us take a surface S, which is not necessarily co-planar, and divide in up into (scalar) elements δS_i. Then

$$\int\int_S f(x,y,z)\,dS = \lim_{\delta S_i \to 0} \sum_i f(x,y,z)\,\delta S_i \tag{2.71}$$

is a surface integral. For instance, the volume of water in a lake of depth $D(x,y)$ is

$$V = \int\int D(x,y)\,dS. \tag{2.72}$$

To evaluate this integral we must split the calculation into two ordinary integrals.



The volume in the strip shown in the diagram is

$$\left[\int_{x_1}^{x_2} D(x,y)\,dx\right] dy. \tag{2.73}$$

Note that the limits x_1 and x_2 depend on y. The total volume is the sum over all strips:

$$V = \int_{y_1}^{y_2} dy \left[\int_{x_1(y)}^{x_2(y)} D(x,y)\,dx\right] \equiv \int\int_S D(x,y)\,dx\,dy. \tag{2.74}$$

Of course, the integral can be evaluated by taking the strips the other way around:

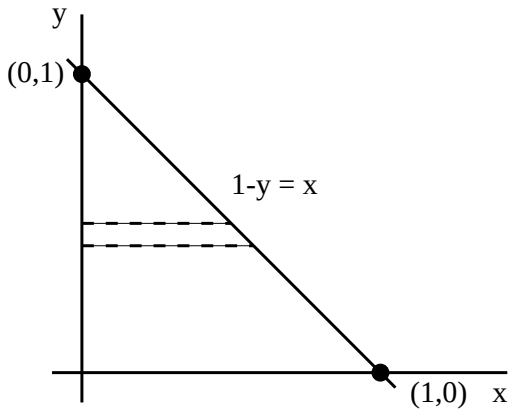
$$V = \int_{x_1}^{x_2} dx \int_{y_1(x)}^{y_2(x)} D(x,y)\,dy. \tag{2.75}$$

Interchanging the order of integration is a very powerful and useful trick. But great care must be taken when evaluating the limits.

As an example, consider

$$\int\int_S x^2 y\,dx\,dy, \tag{2.76}$$

where S is shown in the diagram below. Suppose that we evaluate the x integral



first:

$$dy\left(\int_0^{1-y} x^2 y\,dx\right) = y\,dy\left[\frac{x^3}{3}\right]_0^{1-y} = \frac{y}{3}(1-y)^3\,dy. \tag{2.77}$$

Let us now evaluate the y integral:

$$\int_0^1 \left(\frac{y}{3} - y^2 + y^3 - \frac{y^4}{3} \right) dy = \frac{1}{60}. \tag{2.78}$$

We can also evaluate the integral by interchanging the order of integration:

$$\int_0^1 x^2 \, dx \int_0^{1-x} y \, dy = \int_0^1 \frac{x^2}{2} (1-x)^2 \, dx = \frac{1}{60}. \tag{2.79}$$

In some cases a surface integral is just the product of two separate integrals. For instance,

$$\int\int_S x^2 y \, dx \, dy \tag{2.80}$$

where S is a unit square. This integral can be written

$$\int_0^1 dx \int_0^1 x^2 y \, dy = \left(\int_0^1 x^2 \, dx \right) \left(\int_0^1 y \, dy \right) = \frac{1}{3} \frac{1}{2} = \frac{1}{6}, \tag{2.81}$$

since the limits are both independent of the other variable.

In general, when interchanging the order of integration the most important part of the whole problem is getting the limits of integration right. The only foolproof way of doing this is to *draw a diagram*.

2.12 Vector surface integrals

Surface integrals often occur during vector analysis. For instance, the rate of flow of a liquid of velocity $\boldsymbol{v}$ through an infinitesimal surface of vector area $d\boldsymbol{S}$ is $\boldsymbol{v} \cdot d\boldsymbol{S}$. The net rate of flow of a surface $\boldsymbol{S}$ made up of lots of infinitesimal surfaces is

$$\int\int_S \boldsymbol{v} \cdot d\boldsymbol{S} = \lim_{dS \to 0} \left[\sum v \cos\theta \, dS \right], \tag{2.82}$$

where θ is the angle subtended between the normal to the surface and the flow velocity.

As with line integrals, most surface integrals depend both on the surface and the rim. But some (very important) integrals depend only on the rim, and not on the nature of the surface which spans it. As an example of this, consider incompressible fluid flow between two surfaces S_1 and S_2 which end on the same rim. The volume between the surfaces is constant, so what goes in must come out, and

$$\int\int_{S_1} \boldsymbol{v} \cdot d\boldsymbol{S} = \int\int_{S_2} \boldsymbol{v} \cdot d\boldsymbol{S}. \tag{2.83}$$

It follows that

$$\int\int \boldsymbol{v} \cdot d\boldsymbol{S} \tag{2.84}$$

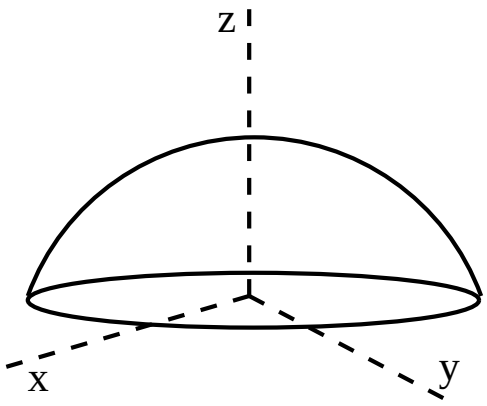
depends only on the rim, and not on the form of surfaces S_1 and S_2.

2.13 Volume integrals

A volume integral takes the form

$$\int\int\int_V f(x, y, z)\, dV \tag{2.85}$$

where V is some volume and $dV = dx\, dy\, dz$ is a small volume element. The volume element is sometimes written $d^3\boldsymbol{r}$, or even $d\tau$. As an example of a volume integral, let us evaluate the centre of gravity of a solid hemisphere of radius a. The height of the centre of gravity is given by



$$\overline{z} = \int\int\int z\, dV \Big/ \int\int\int dV. \tag{2.86}$$

The bottom integral is simply the volume of the hemisphere, which is $2\pi a^3/3$. The top integral is most easily evaluated in spherical polar coordinates, for which $z = r\cos\theta$ and $dV = r^2 \sin\theta\, dr\, d\theta\, d\phi$. Thus,

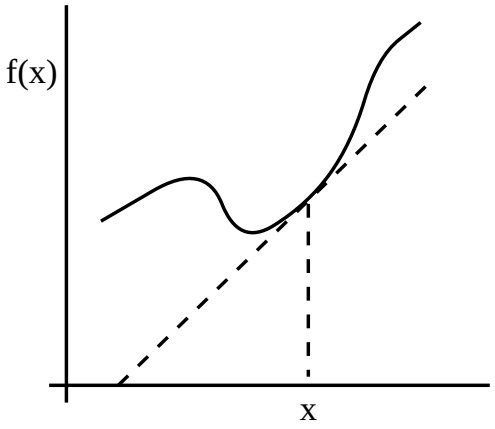
$$\begin{aligned} \int\int\int z\, dV &= \int_0^a dr \int_0^{\pi/2} d\theta \int_0^{2\pi} d\phi\ r\cos\theta\ r^2 \sin\theta \\ &= \int_0^a r^3\, dr \int_0^{\pi/2} \sin\theta\cos\theta\, d\theta \int_0^{2\pi} d\phi = \frac{\pi a^4}{4}, \end{aligned} \tag{2.87}$$

giving

$$\overline{z} = \frac{\pi a^4}{4}\frac{3}{2\pi a^3} = \frac{3a}{8}. \tag{2.88}$$

2.14 Gradient

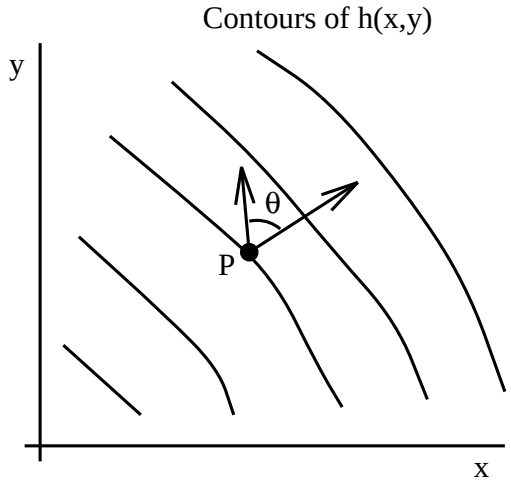
A one-dimensional function $f(x)$ has a gradient df/dx which is defined as the slope of the tangent to the curve at x. We wish to extend this idea to cover scalar



fields in two and three dimensions.

Consider a two-dimensional scalar field $h(x, y)$ which is (say) the height of a hill. Let $d\boldsymbol{l} = (dx, dy)$ be an element of horizontal distance. Consider dh/dl,

where dh is the change in height after moving an infinitesimal distance $d\boldsymbol{l}$. This quantity is somewhat like the one-dimensional gradient, except that dh depends on the *direction* of $d\boldsymbol{l}$, as well as its magnitude. In the immediate vicinity of some



point P the slope reduces to an inclined plane. The largest value of dh/dl is straight up the slope. For any other direction

$$\frac{dh}{dl} = \left(\frac{dh}{dl}\right)_{\max} \cos\theta. \tag{2.89}$$

Let us define a two-dimensional vector $\boldsymbol{grad}\, h$, called the *gradient* of h, whose magnitude is $(dh/dl)_{\max}$ and whose direction is the direction of the steepest slope. Because of the $\cos\theta$ property, the component of $\boldsymbol{grad}\, h$ in any direction equals dh/dl for that direction. [The argument, here, is analogous to that used for vector areas in Section 2.2. See, in particular, Eq. (2.9).]

The component of dh/dl in the x-direction can be obtained by plotting out the profile of h at constant y, and then finding the slope of the tangent to the curve at given x. This quantity is known as the *partial derivative* of h with respect to x at constant y, and is denoted $(\partial h/\partial x)_y$. Likewise, the gradient of the profile at constant x is written $(\partial h/\partial y)_x$. Note that the subscripts denoting constant-x and constant-y are usually omitted, unless there is any ambiguity. If follows that in component form

$$\boldsymbol{grad}\, h = \left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right). \tag{2.90}$$

Now, the equation of the tangent plane at $P = (x_0, y_0)$ is

$$h_T(x, y) = h(x_0, y_0) + \alpha(x - x_0) + \beta(y - y_0). \tag{2.91}$$

This has the same local gradients as $h(x, y)$, so

$$\alpha = \frac{\partial h}{\partial x}, \qquad \beta = \frac{\partial h}{\partial y} \tag{2.92}$$

by differentiation of the above. For small $dx = x - x_0$ and $dy = y - y_0$ the function h is coincident with the tangent plane. We have

$$dh = \frac{\partial h}{\partial x}\, dx + \frac{\partial h}{\partial y}\, dy, \tag{2.93}$$

but $\boldsymbol{grad}\, h = (\partial h/\partial x, \partial h/\partial y)$ and $d\boldsymbol{l} = (dx, dy)$, so

$$dh = \boldsymbol{grad}\, h \cdot d\boldsymbol{l}. \tag{2.94}$$

Incidentally, the above equation demonstrates that $\boldsymbol{grad}\, h$ is a proper vector, since the left-hand side is a scalar and, according to the properties of the dot product, the right-hand side is also a scalar provided that $d\boldsymbol{l}$ and $\boldsymbol{grad}\, h$ are both proper vectors ($d\boldsymbol{l}$ is an obvious vector because it is directly derived from displacements).

Consider, now, a three-dimensional temperature distribution $T(x, y, z)$ in (say) a reaction vessel. Let us define $\boldsymbol{grad}\, T$, as before, as a vector whose magnitude is $(dT/dl)_{\max}$ and whose direction is the direction of the maximum gradient. This vector is written in component form

$$\boldsymbol{grad}\, T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z} \right). \tag{2.95}$$

Here, $\partial T/\partial x \equiv (\partial T/\partial x)_{y,z}$ is the gradient of the one-dimensional temperature profile at constant y and z. The change in T in going from point P to a neighbouring point offset by $d\boldsymbol{l} = (dx, dy, dz)$ is

$$dT = \frac{\partial T}{\partial x}\, dx + \frac{\partial T}{\partial y}\, dy + \frac{\partial T}{\partial z}\, dz. \tag{2.96}$$

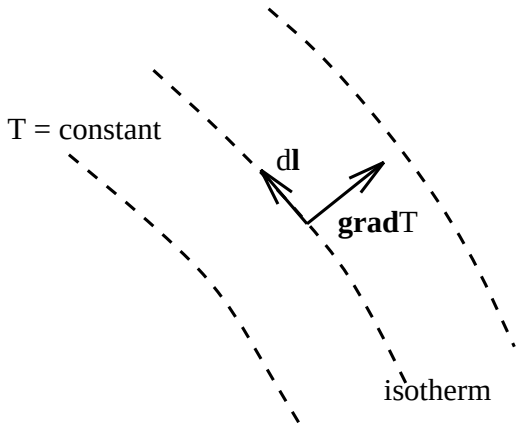
In vector form this becomes

$$dT = \boldsymbol{grad}\, T \cdot d\boldsymbol{l}. \tag{2.97}$$

Suppose that $dT = 0$ for some $d\boldsymbol{l}$. It follows that

$$dT = \boldsymbol{grad}\, T \cdot d\boldsymbol{l} = 0, \tag{2.98}$$

so $d\boldsymbol{l}$ is perpendicular to $\boldsymbol{grad}\, T$. Since $dT = 0$ along so-called "isotherms" (*i.e.*, contours of the temperature) we conclude that the isotherms (contours) are everywhere perpendicular to $\boldsymbol{grad}\, T$.



It is, of course, possible to integrate dT. The line integral from point P to point Q is written

$$\int_P^Q dT = \int_P^Q \boldsymbol{grad}\, T \cdot d\boldsymbol{l} = T(Q) - T(P). \tag{2.99}$$

This integral is clearly independent of the path taken between P and Q, so $\int_P^Q \boldsymbol{grad}\, T \cdot d\boldsymbol{l}$ must be path independent.

In general, $\int_P^Q \boldsymbol{A} \cdot d\boldsymbol{l}$ depends on path, but for some special vector fields the integral is path independent. Such fields are called *conservative* fields. It can be shown that if $\boldsymbol{A}$ is a conservative field then $\boldsymbol{A} = \boldsymbol{grad}\, \phi$ for some scalar field ϕ. The proof of this is straightforward. Keeping P fixed we have

$$\int_P^Q \boldsymbol{A} \cdot d\boldsymbol{l} = V(Q), \tag{2.100}$$

where $V(Q)$ is a well-defined function due to the path independent nature of the line integral. Consider moving the position of the end point by an infinitesimal amount dx in the x-direction. We have

$$V(Q+dx) = V(Q) + \int_Q^{Q+dx} \boldsymbol{A}\cdot d\boldsymbol{l} = V(Q) + A_x\, dx. \tag{2.101}$$

Hence,

$$\frac{\partial V}{\partial x} = A_x, \tag{2.102}$$

with analogous relations for the other components of $\boldsymbol{A}$. It follows that

$$\boldsymbol{A} = \boldsymbol{grad}\, V. \tag{2.103}$$

In physics, the force due to gravity is a good example of a conservative field. If $\boldsymbol{A}$ is a force, then $\int \boldsymbol{A}\cdot d\boldsymbol{l}$ is the work done in traversing some path. If $\boldsymbol{A}$ is conservative then

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = 0, \tag{2.104}$$

where $\oint$ corresponds to the line integral around some closed loop. The fact that zero net work is done in going around a closed loop is equivalent to the conservation of energy (this is why conservative fields are called "conservative"). A good example of a non-conservative field is the force due to friction. Clearly, a frictional system loses energy in going around a closed cycle, so $\oint \boldsymbol{A}\cdot d\boldsymbol{l} \neq 0$.

It is useful to define the vector *operator*

$$\nabla \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right), \tag{2.105}$$

which is usually called the "grad" or "del" operator. This operator acts on everything to its right in a expression until the end of the expression or a closing bracket is reached. For instance,

$$\boldsymbol{grad}\, f = \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right). \tag{2.106}$$

For two scalar fields ϕ and ψ,

$$\boldsymbol{grad}(\phi\psi) = \phi\,\boldsymbol{grad}\,\psi + \psi\,\boldsymbol{grad}\,\phi \tag{2.107}$$

can be written more succinctly as

$$\nabla(\phi\psi) = \phi\nabla\psi + \psi\nabla\phi. \tag{2.108}$$

Suppose that we rotate the basis about the z-axis by θ degrees. By analogy with Eq. (2.7), the old coordinates (x, y, z) are related to the new ones (x', y', z') via

$$\begin{aligned} x &= x'\cos\theta - y'\sin\theta, \\ y &= x'\sin\theta + y'\cos\theta, \\ z &= z'. \end{aligned} \tag{2.109}$$

Now,

$$\frac{\partial}{\partial x'} = \left(\frac{\partial x}{\partial x'}\right)_{y',z'}\frac{\partial}{\partial x} + \left(\frac{\partial y}{\partial x'}\right)_{y',z'}\frac{\partial}{\partial y} + \left(\frac{\partial z}{\partial x'}\right)_{y',z'}\frac{\partial}{\partial z}, \tag{2.110}$$

giving

$$\frac{\partial}{\partial x'} = \cos\theta\,\frac{\partial}{\partial x} + \sin\theta\,\frac{\partial}{\partial y}, \tag{2.111}$$

and

$$\nabla_{x'} = \cos\theta\,\nabla_x + \sin\theta\,\nabla_y. \tag{2.112}$$

It can be seen that the differential operator ∇ transforms like a proper vector, according to Eq. (2.8). This is another proof that ∇f is a good vector.

2.15 Divergence

Let us start with a vector field $\boldsymbol{A}$. Consider $\oint_S \boldsymbol{A}\cdot d\boldsymbol{S}$ over some closed surface S, where $d\boldsymbol{S}$ denotes an *outward* pointing surface element. This surface integral

is usually called the *flux* of $\boldsymbol{A}$ out of S. If $\boldsymbol{A}$ is the velocity of some fluid, then $\oint_S \boldsymbol{A}\cdot d\boldsymbol{S}$ is the rate of flow of material out of S.

If $\boldsymbol{A}$ is constant in space then it is easily demonstrated that the net flux out of S is zero:

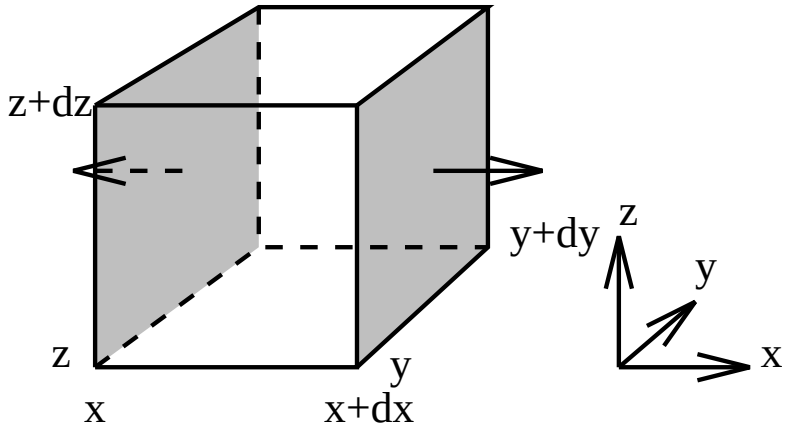
$$\oint \boldsymbol{A}\cdot d\boldsymbol{S} = \boldsymbol{A}\cdot\oint d\boldsymbol{S} = \boldsymbol{A}\cdot\boldsymbol{S} = 0, \tag{2.113}$$

since the vector area $\boldsymbol{S}$ of a closed surface is zero.

Suppose, now, that $\boldsymbol{A}$ is not uniform in space. Consider a very small rectangular volume over which $\boldsymbol{A}$ hardly varies. The contribution to $\oint \boldsymbol{A}\cdot d\boldsymbol{S}$ from the two faces normal to the x-axis is

$$A_x(x+dx)\,dy\,dz - A_x(x)\,dy\,dz = \frac{\partial A_x}{\partial x}\,dx\,dy\,dz = \frac{\partial A_x}{\partial x}\,dV, \tag{2.114}$$

where $dV = dx\,dy\,dz$ is the volume element. There are analogous contributions



from the sides normal to the y and z-axes, so the total of all the contributions is

$$\oint \boldsymbol{A}\cdot d\boldsymbol{S} = \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) dV. \tag{2.115}$$

The *divergence* of a vector field is defined

$$div\,\boldsymbol{A} = \nabla\cdot\boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}. \tag{2.116}$$

Divergence is a good scalar (*i.e.*, it is coordinate independent), since it is the dot product of the vector operator ∇ with $\boldsymbol{A}$. The formal definition of $div\,\boldsymbol{A}$ is

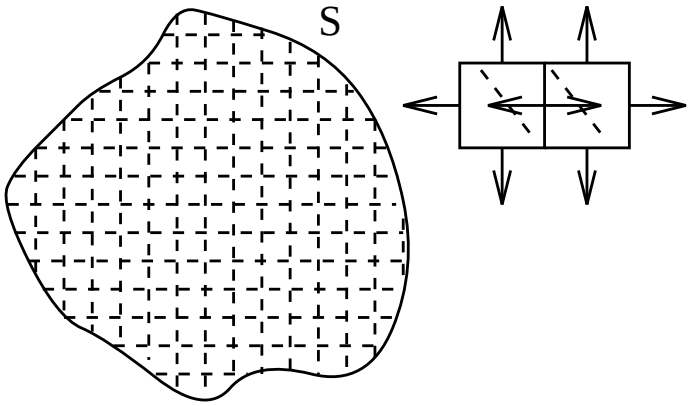
$$div\,\boldsymbol{A} = \lim_{dV\to 0} \frac{\oint \boldsymbol{A}\cdot d\boldsymbol{S}}{dV}. \tag{2.117}$$

This definition is independent of the shape of the infinitesimal volume element.

One of the most important results in vector field theory is the so-called *divergence theorem* or Gauss' theorem. This states that for any volume V surrounded by a closed surface S,

$$\oint_S \boldsymbol{A}\cdot d\boldsymbol{S} = \int_V div\,\boldsymbol{A}\,dV, \tag{2.118}$$

where $d\boldsymbol{S}$ is an outward pointing volume element. The proof is very straightfor-



ward. We divide up the volume into lots of very small cubes and sum $\int \boldsymbol{A}\cdot d\boldsymbol{S}$ over all of the surfaces. The contributions from the interior surfaces cancel out, leaving just the contribution from the outer surface. We can use Eq. (2.115) for each cube individually. This tells us that the summation is equivalent to $\int div\,\boldsymbol{A}\,dV$ over the whole volume. Thus, the integral of $\boldsymbol{A}\cdot d\boldsymbol{S}$ over the outer surface is equal to the integral of $div\,\boldsymbol{A}$ over the whole volume, which proves the divergence theorem.

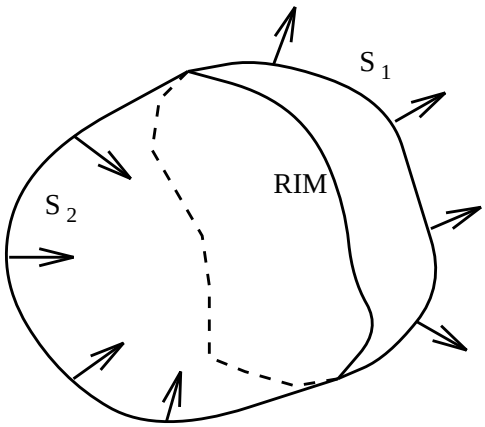
Now, for a vector field with $div\,\boldsymbol{A} = 0$,

$$\oint_S \boldsymbol{A}\cdot d\boldsymbol{S} = 0 \tag{2.119}$$

for any closed surface S. So, for two surfaces on the same rim,

$$\int_{S_1} \boldsymbol{A} \cdot d\boldsymbol{S} = \int_{S_2} \boldsymbol{A} \cdot d\boldsymbol{S}. \tag{2.120}$$

Thus, if $div\, \boldsymbol{A} = 0$ then the surface integral depends on the rim but not the nature of the surface which spans it. On the other hand, if $div\, \boldsymbol{A} \neq 0$ then the integral depends on both the rim and the surface.



Consider an incompressible fluid whose velocity field is $\boldsymbol{v}$. It is clear that $\oint \boldsymbol{v} \cdot d\boldsymbol{S} = 0$ for any closed surface, since what flows into the surface must flow out again. Thus, according to the divergence theorem, $\int div\, \boldsymbol{v}\, dV = 0$ for any volume. The only way in which this is possible is if $div\, \boldsymbol{v}$ is everywhere zero. Thus, the velocity components of an incompressible fluid satisfy the following differential relation:

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} = 0. \tag{2.121}$$

Consider, now, a compressible fluid of density ρ and velocity $\boldsymbol{v}$. The surface integral $\oint_S \rho \boldsymbol{v} \cdot d\boldsymbol{S}$ is the net rate of mass flow out of the closed surface S. This must be equal to the rate of decrease of mass inside the volume V enclosed by S, which is written $-(\partial/\partial t)(\int_V \rho\, dV)$. Thus,

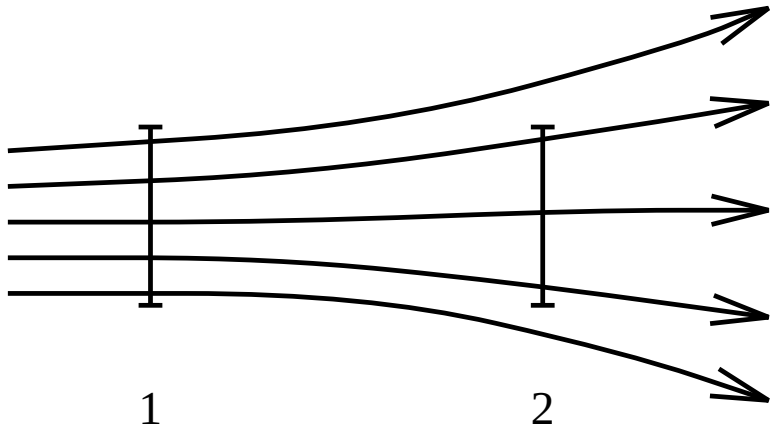
$$\oint_S \rho \boldsymbol{v} \cdot d\boldsymbol{S} = -\frac{\partial}{\partial t}\left(\int_V \rho\, dV\right) \tag{2.122}$$

for any volume. It follows from the divergence theorem that

$$div(\rho \boldsymbol{v}) = -\frac{\partial \rho}{\partial t}. \tag{2.123}$$

This is called the *equation of continuity* of the fluid, since it ensures that fluid is neither created nor destroyed as it flows from place to place. If ρ is constant then the equation of continuity reduces to the previous incompressible result $div\, \boldsymbol{v} = 0$.

It is sometimes helpful to represent a vector field $\boldsymbol{A}$ by "lines of force" or "field lines." The direction of a line of force at any point is the same as the direction of $\boldsymbol{A}$. The density of lines (*i.e.*, the number of lines crossing a unit surface perpendicular to $\boldsymbol{A}$) is equal to $|A|$. In the diagram, $|A|$ is larger at point



1 than at point 2. The number of lines crossing a surface element $d\boldsymbol{S}$ is $\boldsymbol{A} \cdot d\boldsymbol{S}$. So, the net number of lines leaving a closed surface is

$$\oint_S \boldsymbol{A} \cdot d\boldsymbol{S} = \int_V div\, \boldsymbol{A}\, dV. \tag{2.124}$$

If $div\, \boldsymbol{A} = 0$ then there is no net flux of lines out of any surface, which mean that the lines of force must form *closed loops*. Such a field is called a *solenoidal* vector field.

2.16 The Laplacian

So far we have encountered

$$\boldsymbol{grad}\,\phi = \left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, \frac{\partial\phi}{\partial z}\right), \tag{2.125}$$

which is a vector field formed from a scalar field, and

$$div\,\boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}, \tag{2.126}$$

which is a scalar field formed from a vector field. There are two ways in which we can combine $\boldsymbol{grad}$ and div. We can either form the vector field $\boldsymbol{grad}(div\,\boldsymbol{A})$ or the scalar field $div(\boldsymbol{grad}\,\phi)$. The former is not particularly interesting, but the scalar field $div(\boldsymbol{grad}\,\phi)$ turns up in a great many physics problems and is, therefore, worthy of discussion.

Let us introduce the heat flow vector $\boldsymbol{h}$ which is the rate of flow of heat energy per unit area across a surface perpendicular to the direction of $\boldsymbol{h}$. In many substances heat flows directly down the temperature gradient, so that we can write

$$\boldsymbol{h} = -\kappa\,\boldsymbol{grad}\,T, \tag{2.127}$$

where κ is the thermal conductivity. The net rate of heat flow $\oint_S \boldsymbol{h}\cdot d\boldsymbol{S}$ out of some closed surface S must be equal to the rate of decrease of heat energy in the volume V enclosed by S. Thus, we can write

$$\oint_S \boldsymbol{h}\cdot d\boldsymbol{S} = -\frac{\partial}{\partial t}\left(\int c\,T\,dV\right), \tag{2.128}$$

where c is the specific heat. It follows from the divergence theorem that

$$div\,\boldsymbol{h} = -c\,\frac{\partial T}{\partial t}. \tag{2.129}$$

Taking the divergence of both sides of Eq. (2.127), and making use of Eq. (2.129), we obtain

$$div\,(\kappa\,\boldsymbol{grad}\,T) = c\,\frac{\partial T}{\partial t}, \tag{2.130}$$

or

$$\nabla \cdot (\kappa \nabla T) = c \frac{\partial T}{\partial t}. \tag{2.131}$$

If κ is constant then the above equation can be written

$$div(\boldsymbol{grad}\, T) = \frac{c}{\kappa} \frac{\partial T}{\partial t}. \tag{2.132}$$

The scalar field $div(\boldsymbol{grad}\, T)$ takes the form

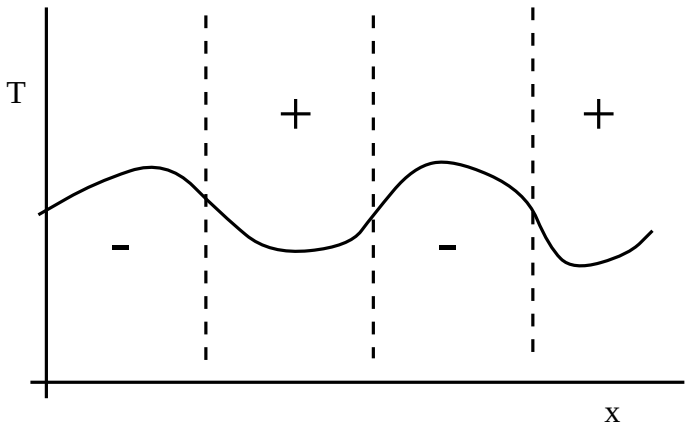
$$\begin{aligned} div(\boldsymbol{grad}\, T) &= \frac{\partial}{\partial x}\left(\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial T}{\partial z}\right) \\ &= \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \equiv \nabla^2 T. \end{aligned} \tag{2.133}$$

Here, the scalar differential operator

$$\nabla^2 \equiv \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \tag{2.134}$$

is called the *Laplacian*. The Laplacian is a good scalar operator (*i.e.*, it is coordinate independent) because it is formed from a combination of *div* (another good scalar operator) and ***grad*** (a good vector operator).

What is the physical significance of the Laplacian? In one-dimension $\nabla^2 T$ reduces to $\partial^2 T/\partial x^2$. Now, $\partial^2 T/\partial x^2$ is positive if $T(x)$ is concave (from above)



and negative if it is convex. So, if T is less than the average of T in its surroundings then $\nabla^2 T$ is positive, and *vice versa*.

In two dimensions

$$\nabla^2 T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}. \tag{2.135}$$

Consider a local minimum of the temperature. At the minimum the slope of T increases in all directions so $\nabla^2 T$ is positive. Likewise, $\nabla^2 T$ is negative at a local maximum. Consider, now, a steep-sided valley in T. Suppose that the bottom of the valley runs parallel to the x-axis. At the bottom of the valley $\partial^2 T/\partial y^2$ is large and positive, whereas $\partial^2 T/\partial x^2$ is small and may even be negative. Thus, $\nabla^2 T$ is positive, and this is associated with T being less than the average local value $\overline{T}$.

Let us now return to the heat conduction problem:

$$\nabla^2 T = \frac{c}{\kappa}\frac{\partial T}{\partial t}. \tag{2.136}$$

It is clear that if $\nabla^2 T$ is positive then T is locally less than the average value, so $\partial T/\partial t > 0$; *i.e.*, the region heats up. Likewise, if $\nabla^2 T$ is negative then T is locally greater than the average value and heat flows out of the region; *i.e.*, $\partial T/\partial t < 0$. Thus, the above heat conduction equation makes physical sense.

2.17 Curl

Consider a vector field $\boldsymbol{A}$ and a loop which lies in one plane. The integral of $\boldsymbol{A}$ around this loop is written $\oint \boldsymbol{A}\cdot d\boldsymbol{l}$, where $d\boldsymbol{l}$ is a line element of the loop. If $\boldsymbol{A}$ is a conservative field then $\boldsymbol{A} = \boldsymbol{grad}\,\phi$ and $\oint \boldsymbol{A}\cdot d\boldsymbol{l} = 0$ for all loops. For a non-conservative field $\oint \boldsymbol{A}\cdot d\boldsymbol{l} \neq 0$, in general.

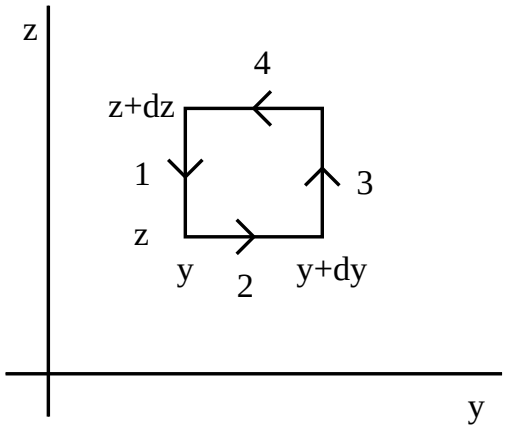
For a small loop we expect $\oint \boldsymbol{A}\cdot d\boldsymbol{l}$ to be proportional to the area of the loop. Moreover, for a fixed area loop we expect $\oint \boldsymbol{A}\cdot d\boldsymbol{l}$ to depend on the *orientation* of the loop. One particular orientation will give the maximum value: $\oint \boldsymbol{A}\cdot d\boldsymbol{l} = I_{\max}$. If the loop subtends an angle θ with this optimum orientation then we expect

$I = I_{\max} \cos\theta$. Let us introduce the vector field $\boldsymbol{curl\, A}$ whose magnitude is

$$|\boldsymbol{curl\, A}| = \lim_{dS \to 0} \frac{\oint \boldsymbol{A} \cdot d\boldsymbol{l}}{dS} \tag{2.137}$$

for the orientation giving $I_{\max}$. Here, dS is the area of the loop. The direction of $\boldsymbol{curl\, A}$ is perpendicular to the plane of the loop, when it is in the orientation giving $I_{\max}$, with the sense given by the right-hand grip rule assuming that the loop is right-handed.

Let us now express $\boldsymbol{curl\, A}$ in terms of the components of $\boldsymbol{A}$. First, we shall evaluate $\oint \boldsymbol{A} \cdot d\boldsymbol{l}$ around a small rectangle in the y-z plane. The contribution from



sides 1 and 3 is

$$A_z(y + dy)\, dz - A_z(y)\, dz = \frac{\partial A_z}{\partial y}\, dy\, dz. \tag{2.138}$$

The contribution from sides 2 and 4 is

$$-A_y(z + dz)\, dy + A_y(z)\, dy = -\frac{\partial A_y}{\partial y}\, dy\, dz. \tag{2.139}$$

So, the total of all contributions gives

$$\oint \boldsymbol{A} \cdot d\boldsymbol{l} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right) dS, \tag{2.140}$$

where $dS = dy\, dz$ is the area of the loop.

Consider a non-rectangular (but still small) loop in the y-z plane. We can divide it into rectangular elements and form $\oint \boldsymbol{A}\cdot d\boldsymbol{l}$ over all the resultant loops. The interior contributions cancel, so we are just left with the contribution from the outer loop. Also, the area of the outer loop is the sum of all the areas of the inner loops. We conclude that

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) dS_x \tag{2.141}$$

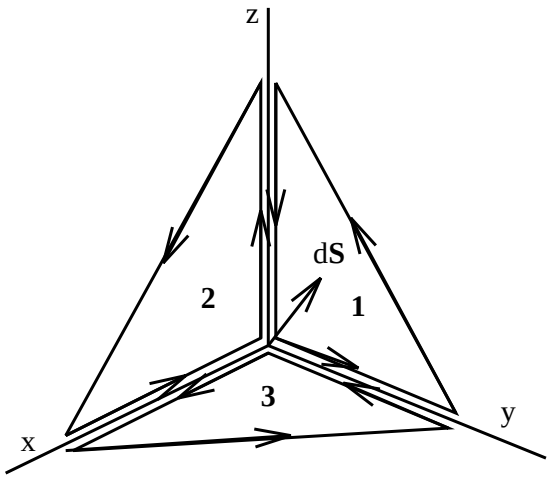
is valid for a small loop $d\boldsymbol{S} = (dS_x, 0, 0)$ of any shape in the y-z plane. Likewise, we can show that if the loop is in the x-z plane then $d\boldsymbol{S} = (0, dS_y, 0)$ and

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) dS_y. \tag{2.142}$$

Finally, if the loop is in the x-y plane then $d\boldsymbol{S} = (0, 0, dS_z)$ and

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) dS_z. \tag{2.143}$$

Imagine an arbitrary loop of vector area $d\boldsymbol{S} = (dS_x, dS_y, dS_z)$. We can construct this out of three loops in the x, y, and z directions, as indicated in the diagram below. If we form the line integral around all three loops then the interior contributions cancel and we are left with the line integral around the original



loop. Thus,

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = \oint \boldsymbol{A}\cdot d\boldsymbol{l}_1 + \oint \boldsymbol{A}\cdot d\boldsymbol{l}_2 + \oint \boldsymbol{A}\cdot d\boldsymbol{l}_3, \tag{2.144}$$

giving

$$\oint \boldsymbol{A}\cdot d\boldsymbol{l} = \boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S} = |\boldsymbol{curl}\,\boldsymbol{A}||dS|\cos\theta, \tag{2.145}$$

where

$$\boldsymbol{curl}\,\boldsymbol{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right). \tag{2.146}$$

Note that

$$\boldsymbol{curl}\,\boldsymbol{A} = \nabla\wedge\boldsymbol{A}. \tag{2.147}$$

This demonstrates that $\boldsymbol{curl}\,\boldsymbol{A}$ is a good vector field, since it is the cross product of the ∇ operator (a good vector operator) and the vector field $\boldsymbol{A}$.

Consider a solid body rotating about the z-axis. The angular velocity is given by $\boldsymbol{\omega} = (0, 0, \omega)$, so the rotation velocity at position $\boldsymbol{r}$ is

$$\boldsymbol{v} = \boldsymbol{\omega}\wedge\boldsymbol{r} \tag{2.148}$$

[see Eq. (2.39)]. Let us evaluate $\boldsymbol{curl}\,\boldsymbol{v}$ on the axis of rotation. The x-component is proportional to the integral $\oint \boldsymbol{v}\cdot d\boldsymbol{l}$ around a loop in the y-z plane. This is plainly zero. Likewise, the y-component is also zero. The z-component is $\oint \boldsymbol{v}\cdot d\boldsymbol{l}/dS$ around some loop in the x-y plane. Consider a circular loop. We have $\oint \boldsymbol{v}\cdot d\boldsymbol{l} = 2\pi r\,\omega r$ with $dS = \pi r^2$. Here, r is the radial distance from the rotation axis. It follows that $(\boldsymbol{curl}\,\boldsymbol{v})_z = 2\omega$, which is independent of r. So, on the axis $\boldsymbol{curl}\,\boldsymbol{v} = (0, 0, 2\omega)$. Off the axis, at position $\boldsymbol{r}_0$, we can write

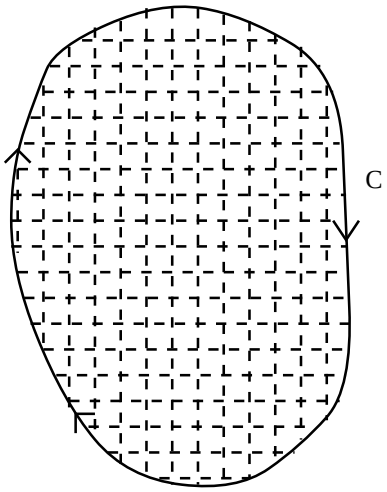
$$\boldsymbol{v} = \boldsymbol{\omega}\wedge(\boldsymbol{r}-\boldsymbol{r}_0) + \boldsymbol{\omega}\wedge\boldsymbol{r}_0. \tag{2.149}$$

The first part has the same curl as the velocity field on the axis, and the second part has zero curl since it is constant. Thus, $\boldsymbol{curl}\,\boldsymbol{v} = (0, 0, 2\omega)$ everywhere in the body. This allows us to form a physical picture of $\boldsymbol{curl}\,\boldsymbol{A}$. If we imagine $\boldsymbol{A}$ as the velocity field of some fluid then $\boldsymbol{curl}\,\boldsymbol{A}$ at any given point is equal to twice the local angular rotation velocity, *i.e.*, $2\boldsymbol{\omega}$. Hence, a vector field with $\boldsymbol{curl}\,\boldsymbol{A} = \boldsymbol{0}$ everywhere is said to be *irrotational*.

Another important result of vector field theory is the *curl theorem* or Stokes' theorem:

$$\oint_C \boldsymbol{A}\cdot d\boldsymbol{l} = \int_S \boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S}, \tag{2.150}$$

for some (non-planar) surface S bounded by a rim C. This theorem can easily be proved by splitting the loop up into many small rectangular loops and forming the integral around all of the resultant loops. All of the contributions from the interior loops cancel, leaving just the contribution from the outer rim. Making use of Eq. (2.145) for each of the small loops, we can see that the contribution from all of the loops is also equal to the integral of $\boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S}$ across the whole surface. This proves the theorem.



One immediate consequence of of Stokes' theorem is that $\boldsymbol{curl}\,\boldsymbol{A}$ is "incompressible." Consider two surfaces, S_1 and S_2, which share the same rim. It is clear from Stokes' theorem that $\int \boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S}$ is the same for both surfaces. Thus, it follows that $\oint \boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S} = 0$ for any closed surface. However, we have from the divergence theorem that $\oint \boldsymbol{curl}\,\boldsymbol{A}\cdot d\boldsymbol{S} = \int div(\boldsymbol{curl}\,\boldsymbol{A})\,dV = 0$ for any volume. Hence,

$$div(\boldsymbol{curl}\,\boldsymbol{A}) \equiv 0. \tag{2.151}$$

So, the field-lines of $\boldsymbol{curl}\,\boldsymbol{A}$ never begin or end. In other words, $\boldsymbol{curl}\,\boldsymbol{A}$ is a solenoidal field.

We have seen that for a conservative field $\oint \boldsymbol{A}\cdot d\boldsymbol{l} = 0$ for any loop. This is entirely equivalent to $\boldsymbol{A} = \boldsymbol{grad}\,\phi$. However, the magnitude of $\boldsymbol{curl}\,\boldsymbol{A}$ is

$\lim_{dS\rightarrow 0} \oint \boldsymbol{A}\cdot d\boldsymbol{l}/dS$ for some particular loop. It is clear then that $\boldsymbol{curl}\,\boldsymbol{A} = \boldsymbol{0}$ for a conservative field. In other words,

$$\boldsymbol{curl}(\boldsymbol{grad}\,\phi) \equiv \boldsymbol{0}. \tag{2.152}$$

Thus, a conservative field is also an irrotational one.

Finally, it can be shown that

$$\boldsymbol{curl}(\boldsymbol{curl}\,\boldsymbol{A}) = \boldsymbol{grad}(div\,\boldsymbol{A}) - \nabla^2\boldsymbol{A}, \tag{2.153}$$

where

$$\nabla^2\boldsymbol{A} = (\nabla^2 A_x, \nabla^2 A_y, \nabla^2 A_z). \tag{2.154}$$

It should be emphasized, however, that the above result is only valid in Cartesian coordinates.

2.18 Summary

Vector addition:

$$\boldsymbol{a}+\boldsymbol{b} \equiv (a_x+b_x, a_y+b_y, a_z+b_z)$$

Vector multiplication:

$$n\boldsymbol{a} \equiv (na_x, na_y, na_z)$$

Scalar product:

$$\boldsymbol{a}\cdot\boldsymbol{b} = a_x b_x + a_y b_y + a_z b_z$$

Vector product:

$$\boldsymbol{a}\wedge\boldsymbol{b} = (a_y b_z - a_z b_y, a_z b_x - a_x b_z, a_x b_y - a_y b_x)$$

Scalar triple product:

$$\boldsymbol{a}\cdot\boldsymbol{b}\wedge\boldsymbol{c} = \boldsymbol{a}\wedge\boldsymbol{b}\cdot\boldsymbol{c} = \boldsymbol{b}\cdot\boldsymbol{c}\wedge\boldsymbol{a} = -\boldsymbol{b}\cdot\boldsymbol{a}\wedge\boldsymbol{c}$$

Vector triple product:

$$\boldsymbol{a} \wedge (\boldsymbol{b} \wedge \boldsymbol{c}) = (\boldsymbol{a} \cdot \boldsymbol{c})\boldsymbol{b} - (\boldsymbol{a} \cdot \boldsymbol{b})\boldsymbol{c}$$

$$(\boldsymbol{a} \wedge \boldsymbol{b}) \wedge \boldsymbol{c} = (\boldsymbol{a} \cdot \boldsymbol{c})\boldsymbol{b} - (\boldsymbol{b} \cdot \boldsymbol{c})\boldsymbol{a}$$

Gradient:

$$\boldsymbol{grad}\, \phi = \left(\frac{\partial \phi}{\partial x}, \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial z} \right)$$

Divergence:

$$div\, \boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

Curl:

$$\boldsymbol{curl}\, \boldsymbol{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right)$$

Gauss' theorem:

$$\oint_S \boldsymbol{A} \cdot d\boldsymbol{S} = \int_V div\, \boldsymbol{A}\, dV$$

Stokes' theorem:

$$\oint_C \boldsymbol{A} \cdot d\boldsymbol{l} = \int_S \boldsymbol{curl}\, \boldsymbol{A} \cdot d\boldsymbol{S}$$

Del operator:

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

$$\boldsymbol{grad}\, \phi = \nabla \phi$$

$$div\, \boldsymbol{A} = \nabla \cdot \boldsymbol{A}$$

$$\boldsymbol{curl}\, \boldsymbol{A} = \nabla \wedge \boldsymbol{A}$$

Vector identities:

$$\nabla \cdot \nabla \phi = \nabla^2 \phi = \left(\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \right)$$

$$\nabla \cdot \nabla \wedge \boldsymbol{A} = 0$$

$$\nabla \wedge \nabla \phi = 0$$

$$\nabla^2 \boldsymbol{A} = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla \wedge \nabla \wedge \boldsymbol{A}$$

Other vector identities:

$$\nabla(\phi\psi) = \phi\nabla(\psi) + \psi\nabla(\phi)$$

$$\nabla \cdot (\phi \boldsymbol{A}) = \phi \nabla \cdot \boldsymbol{A} + \boldsymbol{A} \cdot \nabla \phi$$

$$\nabla \wedge (\phi \boldsymbol{A}) = \phi \nabla \wedge \boldsymbol{A} + \nabla \phi \wedge \boldsymbol{A}$$

$$\nabla \cdot (\boldsymbol{A} \wedge \boldsymbol{B}) = \boldsymbol{B} \cdot \nabla \wedge \boldsymbol{A} - \boldsymbol{A} \cdot \nabla \wedge \boldsymbol{B}$$

$$\nabla \wedge (\boldsymbol{A} \wedge \boldsymbol{B}) = \boldsymbol{A}(\nabla \cdot \boldsymbol{B}) - \boldsymbol{B}(\nabla \cdot \boldsymbol{A}) + (\boldsymbol{B} \cdot \nabla)\boldsymbol{A} - (\boldsymbol{A} \cdot \nabla)\boldsymbol{B}$$

$$\nabla(\boldsymbol{A} \cdot \boldsymbol{B}) = \boldsymbol{A} \wedge (\nabla \wedge \boldsymbol{B}) + \boldsymbol{B} \wedge (\nabla \wedge \boldsymbol{A}) + (\boldsymbol{A} \cdot \nabla)\boldsymbol{B} + (\boldsymbol{B} \cdot \nabla)\boldsymbol{A}$$

Acknowledgment

This section is almost entirely based on my undergraduate notes taken during a course of lectures given by Dr. Steven Gull of the Cavendish Laboratory, Cambridge.

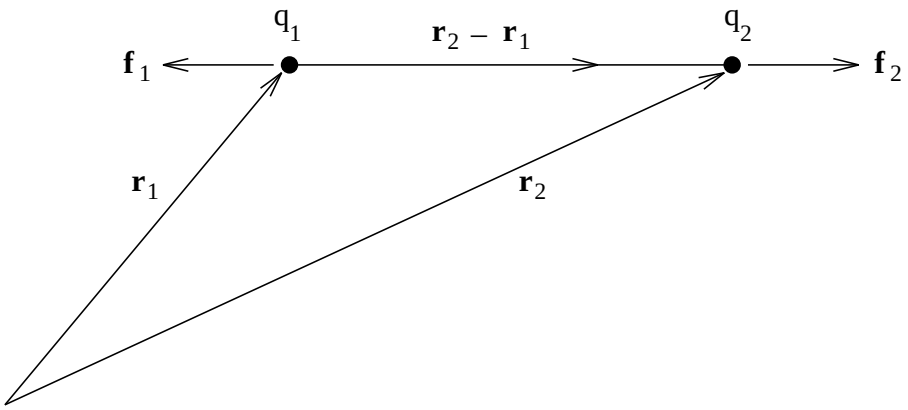
3 Maxwell's equations

3.1 Coulomb's law

Between 1785 and 1787 the French physicist Charles Augustine de Coulomb performed a series of experiments involving electric charges and eventually established what is nowadays known as "Coulomb's law." According to this law the force acting between two charges is radial, inverse-square, and proportional to the product of the charges. Two like charges repel one another whereas two unlike charges attract. Suppose that two charges, q_1 and q_2, are located at position vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. The electrical force acting on the second charge is written

$$\boldsymbol{f}_2 = \frac{q_1 q_2}{4\pi\epsilon_0}\frac{\boldsymbol{r}_2 - \boldsymbol{r}_1}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|^3} \tag{3.1}$$

in vector notation. An equal and opposite force acts on the first charge, in accordance with Newton's third law of motion. The SI unit of electric charge is



the coulomb (C). The charge of an electron is 1.6022×10^{-19} C. The universal constant ϵ_0 is called the "permittivity of free space" and takes the value

$$\epsilon_0 = 8.8542 \times 10^{-12}\ \mathrm{C^2\,N^{-1}m^{-2}}. \tag{3.2}$$

Coulomb's law has the same mathematical form as Newton's law of gravity. Suppose that two masses, m_1 and m_2, are located at position vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$.

The gravitational force acting on the second mass is written

$$\boldsymbol{f}_2 = -G m_1 m_2 \, \frac{\boldsymbol{r}_2 - \boldsymbol{r}_1}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|^3} \tag{3.3}$$

in vector notation. The gravitational constant G takes the value

$$G = 6.6726 \times 10^{-11} \ \mathrm{N\,m^2\,kg^{-2}}. \tag{3.4}$$

Coulomb's law and Newton's law are both "inverse-square"; *i.e.*

$$|\boldsymbol{f}_2| \propto \frac{1}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|^2}. \tag{3.5}$$

However, they differ in two crucial respects. Firstly, the force due to gravity is always *attractive* (there is no such thing as a negative mass!). Secondly, the magnitudes of the two forces are vastly different. Consider the ratio of the electrical and gravitational forces acting on two particles. This ratio is a constant, independent of the relative positions of the particles, and is given by

$$\frac{|\boldsymbol{f}_{\mathrm{electrical}}|}{|\boldsymbol{f}_{\mathrm{gravitational}}|} = \frac{q_1}{m_1} \frac{q_2}{m_2} \frac{1}{4\pi\epsilon_0 G}. \tag{3.6}$$

For electrons the charge to mass ratio $q/m = 1.759 \times 10^{11} \ \mathrm{C\,kg^{-1}}$, so

$$\frac{|\boldsymbol{f}_{\mathrm{electrical}}|}{|\boldsymbol{f}_{\mathrm{gravitational}}|} = 4.17 \times 10^{42}. \tag{3.7}$$

This is a colossal number! Suppose you had a homework problem involving the motion of particles in a box under the action of two forces with the same range but differing in magnitude by a factor 10^{42}. I think that most people would write on line one something like "it is a good approximation to neglect the weaker force in favour of the stronger one." In fact, most people would write this even if the forces differed in magnitude by a factor 10! Applying this reasoning to the motion of particles in the universe we would expect the universe to be governed entirely by electrical forces. However, this is not the case. The force which holds us to the surface of the Earth, and prevents us from floating off into space, is gravity. The force which causes the Earth to orbit the Sun is also gravity. In fact, on

astronomical length-scales gravity is the dominant force and electrical forces are largely irrelevant. The key to understanding this paradox is that there are both positive and negative electric charges whereas there are only positive gravitational "charges." This means that gravitational forces are always cumulative whereas electrical forces can cancel one another out. Suppose, for the sake of argument, that the universe starts out with randomly distributed electric charges. Initially, we expect electrical forces to completely dominate gravity. These forces try to make every positive charge get as far away as possible from other positive charges and as close as possible to other negative charges. After a bit we expect the positive and negative charges to form close pairs. Just how close is determined by quantum mechanics but, in general, it is pretty close; *i.e.*, about 10^{-10} m. The electrical forces due to the charges in each pair effectively cancel one another out on length-scales much larger than the mutual spacing of the pair. It is only possible for gravity to be the dominant long-range force if the number of positive charges in the universe is almost equal to the number of negative charges. In this situation every positive change can find a negative charge to team up with and there are virtually no charges left over. In order for the cancellation of long-range electrical forces to be effective the relative difference in the number of positive and negative charges in the universe must be incredibly small. In fact, positive and negative charges have to cancel each other out to such accuracy that most physicists believe that the net charge of the universe is *exactly* zero. But, it is not enough for the universe to start out with zero charge. Suppose there were some elementary particle process which did not conserve electric charge. Even if this were to go on at a very low rate it would not take long before the fine balance between positive and negative charges in the universe were wrecked. So, it is important that electric charge is a *conserved* quantity (*i.e.*, the charge of the universe can neither increase or decrease). As far as we know, this is the case. To date no elementary particle reactions have been discovered which create or destroy net electric charge.

In summary, there are two long-range forces in the universe, electromagnetism and gravity. The former is enormously stronger than the latter, but is usually "hidden" away inside neutral atoms. The fine balance of forces due to negative and positive electric charges starts to break down on atomic scales. In fact, interatomic and intermolecular forces are electrical in nature. So, electrical forces are basically what prevent us from falling though the floor. But, this is electromag-

netism on the microscopic or atomic scale; what is usually known as "quantum electromagnetism." This course is about "classical electromagnetism"; that is, electromagnetism on length-scales much larger than the atomic scale. Classical electromagnetism generally describes phenomena in which some sort of "violence" is done to matter, so that the close pairing of negative and positive charges is disrupted. This allows electrical forces to manifest themselves on macroscopic length-scales. Of course, very little disruption is necessary before gigantic forces are generated. It is no coincidence that the vast majority of useful machines which mankind has devised during the last century are electromagnetic in nature.

Coulomb's law and Newton's law are both examples of what are usually referred to as "action at a distance" theories. According to Eqs. (3.1) and (3.3), if the first charge or mass is moved then the force acting on the second charge or mass immediately responds. In particular, equal and opposite forces act on the two charges or masses at all times. However, this cannot be correct according to Einstein's theory of relativity. The maximum speed with which information can propagate through the universe is the speed of light. So, if the first charge or mass is moved then there must always be time delay (*i.e.*, at least the time needed for a light signal to propagate between the two charges or masses) before the second charge or mass responds. Consider a rather extreme example. Suppose the first charge or mass is suddenly annihilated. The second charge or mass only finds out about this some time later. During this time interval the second charge or mass experiences an electrical or gravitational force which is as if the first charge or mass were still there. So, during this period there is an action but no reaction, which violates Newton's third law of motion. It is clear that "action at a distance" is not compatible with relativity and, consequently, that Newton's third law of motion is not strictly true. Of course, Newton's third law is intimately tied up with the conservation of momentum in the universe. A concept which most physicists are loath to abandon. It turns out that we can "rescue" momentum conservation by abandoning "action at a distance" theories and adopting so-called "field" theories in which there is a medium, called a field, which transmits the force from one particle to another. In electromagnetism there are, in fact, two fields; the electric field and the magnetic field. Electromagnetic forces are transmitted though these fields at the speed of light, which implies that the laws of relativity are never violated. Moreover, the fields can soak up energy and momentum. This means that even when the actions and reactions

acting on particles are not quite equal and opposite, momentum is still conserved. We can bypass some of the problematic aspects of "action at a distance" by only considering steady-state situations. For the moment, this is how we shall proceed.

Consider N charges, q_1 though q_N, which are located at position vectors $\boldsymbol{r}_1$ through $\boldsymbol{r}_N$. Electrical forces obey what is known as the *principle of superposition*. The electrical force acting on a test charge q at position vector $\boldsymbol{r}$ is simply the vector sum of all of the Coulomb law forces from each of the N charges taken in isolation. In other words, the electrical force exerted by the ith charge (say) on the test charge is the same as if all the other charges were not there. Thus, the force acting on the test charge is given by

$$\boldsymbol{f}(\boldsymbol{r}) = q \sum_{i=1}^{N} \frac{q_i}{4\pi\epsilon_0} \frac{\boldsymbol{r}-\boldsymbol{r}_i}{|\boldsymbol{r}-\boldsymbol{r}_i|^3}. \tag{3.8}$$

It is helpful to define a vector field $\boldsymbol{E}(\boldsymbol{r})$, called the electric field, which is the force exerted on a unit test charge located at position vector $\boldsymbol{r}$. So, the force on a test charge is written

$$\boldsymbol{f} = q\,\boldsymbol{E}, \tag{3.9}$$

and the electric field is given by

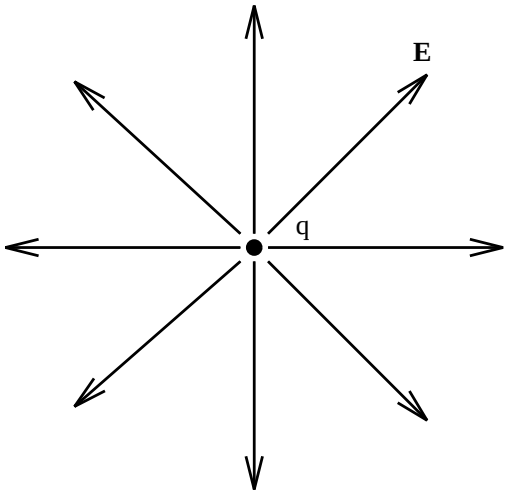
$$\boldsymbol{E}(\boldsymbol{r}) = \sum_{i=1}^{N} \frac{q_i}{4\pi\epsilon_0} \frac{\boldsymbol{r}-\boldsymbol{r}_i}{|\boldsymbol{r}-\boldsymbol{r}_i|^3}. \tag{3.10}$$

At this point, we have no reason to believe that the electric field has any real existence; it is just a useful device for calculating the force which acts on test charges placed at various locations.

The electric field from a single charge q located at the origin is purely radial, points outwards if the charge is positive, inwards if it is negative, and has magnitude

$$E_r(r) = \frac{q}{4\pi\epsilon_0\, r^2}, \tag{3.11}$$

where $r = |\boldsymbol{r}|$.



We can represent an electric field by "field-lines." The direction of the lines indicates the direction of the local electric field and the density of the lines perpendicular to this direction is proportional to the magnitude of the local electric field. Thus, the field of a point positive charge is represented by a group of equally spaced straight lines radiating from the charge.

The electric field from a collection of charges is simply the vector sum of the fields from each of the charges taken in isolation. In other words, electric fields are completely superposable. Suppose that, instead of having discrete charges, we have a continuous distribution of charge represented by a *charge density* $\rho(\boldsymbol{r})$. Thus, the charge at position vector $\boldsymbol{r}'$ is $\rho(\boldsymbol{r}')\,d^3\boldsymbol{r}'$, where $d^3\boldsymbol{r}'$ is the volume element at $\boldsymbol{r}'$. It follows from a simple extension of Eq. (3.10) that the electric field generated by this charge distribution is

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \rho(\boldsymbol{r}')\,\frac{\boldsymbol{r}-\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,d^3\boldsymbol{r}', \tag{3.12}$$

where the volume integral is over all space, or, at least, over all space for which $\rho(\boldsymbol{r}')$ is non-zero.

3.2 The electric scalar potential

Suppose that $\boldsymbol{r} = (x, y, z)$ and $\boldsymbol{r}' = (x', y', z')$ in Cartesian coordinates. The x component of $(\boldsymbol{r} - \boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|^3$ is written

$$\frac{x - x'}{[(x - x')^2 + (y - y')^2 + (z - z')^2]^{3/2}}. \tag{3.13}$$

However, it is easily demonstrated that

$$\frac{x - x'}{[(x - x')^2 + (y - y')^2 + (z - z')^2]^{3/2}} = - \frac{\partial}{\partial x} \frac{1}{[(x - x')^2 + (y - y')^2 + (z - z')^2]^{1/2}}. \tag{3.14}$$

Since there is nothing special about the x-axis we can write

$$\frac{\boldsymbol{r} - \boldsymbol{r}'}{|\boldsymbol{r} - \boldsymbol{r}'|^3} = -\nabla \left(\frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|} \right), \tag{3.15}$$

where $\nabla \equiv (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ is a differential operator which involves the components of $\boldsymbol{r}$ but not those of $\boldsymbol{r}'$. It follows from Eq. (3.12) that

$$\boldsymbol{E} = -\nabla \phi, \tag{3.16}$$

where

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \, d^3\boldsymbol{r}'. \tag{3.17}$$

Thus, the electric field generated by a collection of fixed charges can be written as the gradient of a scalar potential, and this potential can be expressed as a simple volume integral involving the charge distribution.

The scalar potential generated by a charge q located at the origin is

$$\phi(r) = \frac{q}{4\pi\epsilon_0 \, r}. \tag{3.18}$$

According to Eq. (3.10) the scalar potential generated by a set of N discrete charges q_i, located at $\boldsymbol{r}_i$, is

$$\phi(\boldsymbol{r}) = \sum_{i=1}^{N} \phi_i(\boldsymbol{r}), \tag{3.19}$$

where

$$\phi_i(\boldsymbol{r}) = \frac{q_i}{4\pi\epsilon_0\,|\boldsymbol{r}-\boldsymbol{r}_i|}. \tag{3.20}$$

Thus, the scalar potential is just the sum of the potentials generated by each of the charges taken in isolation.

Suppose that a particle of charge q is taken along some path from point P to point Q. The net work done on the particle by electrical forces is

$$W = \int_P^Q \boldsymbol{f}\cdot d\boldsymbol{l}, \tag{3.21}$$

where $\boldsymbol{f}$ is the electrical force and $d\boldsymbol{l}$ is a line element along the path. Making use of Eqs. (3.9) and (3.16) we obtain

$$W = q\int_P^Q \boldsymbol{E}\cdot d\boldsymbol{l} = -q\int_P^Q \nabla\phi\cdot d\boldsymbol{l} = -q\,(\,\phi(Q)-\phi(P)\,)\,. \tag{3.22}$$

Thus, the work done on the particle is simply minus its charge times the difference in electric potential between the end point and the beginning point. This quantity is clearly independent of the path taken from P to Q. So, an electric field generated by stationary charges is an example of a conservative field. In fact, this result follows immediately from vector field theory once we are told, in Eq. (3.16), that the electric field is the gradient of a scalar potential. The work done on the particle when it is taken around a closed path is zero, so

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = 0 \tag{3.23}$$

for any closed loop C. This implies from Stokes' theorem that

$$\nabla\wedge\boldsymbol{E} = \boldsymbol{0} \tag{3.24}$$

for any electric field generated by stationary charges. Equation (3.24) also follows directly from Eq. (3.16), since $\nabla \wedge \nabla\phi = \mathbf{0}$ for any scalar potential ϕ.

The SI unit of electric potential is the volt, which is equivalent to a joule per coulomb. Thus, according to Eq. (3.22) the electrical work done on a particle when it is taken between two points is the product of its charge and the voltage difference between the points.

We are familiar with the idea that a particle moving in a gravitational field possesses potential energy as well as kinetic energy. If the particle moves from point P to a lower point Q then the gravitational field does work on the particle causing its kinetic energy to increase. The increase in kinetic energy of the particle is balanced by an equal decrease in its potential energy so that the overall energy of the particle is a conserved quantity. Therefore, the work done on the particle as it moves from P to Q is *minus* the difference in its gravitational potential energy between points Q and P. Of course, it only makes sense to talk about gravitational potential energy because the gravitational field is conservative. Thus, the work done in taking a particle between two points is path independent and, therefore, well defined. This means that the difference in potential energy of the particle between the beginning and end points is also well defined. We have already seen that an electric field generated by stationary charges is a conservative field. In follows that we can define an electrical potential energy of a particle moving in such a field. By analogy with gravitational fields, the work done in taking a particle from point P to point Q is equal to minus the difference in potential energy of the particle between points Q and P. It follows from Eq. (3.22) that the potential energy of the particle at a general point Q, relative to some reference point P, is given by

$$\mathcal{E}(Q) = q\,\phi(Q). \tag{3.25}$$

Free particles try to move down gradients of potential energy in order to attain a minimum potential energy state. Thus, free particles in the Earth's gravitational field tend to fall downwards. Likewise, positive charges moving in an electric field tend to migrate towards regions with the most negative voltage and *vice versa* for negative charges.

The scalar electric potential is undefined to an additive constant. So, the

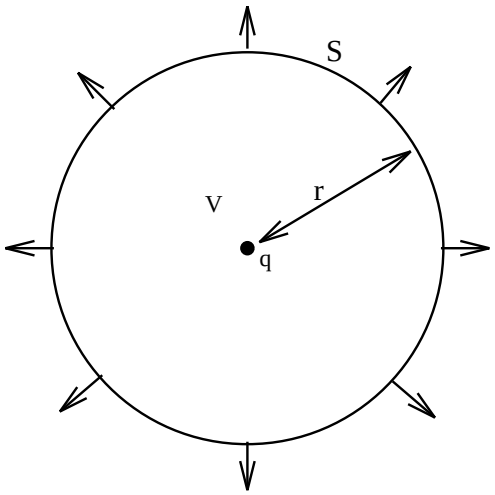
transformation

$$\phi(\boldsymbol{r}) \rightarrow \phi(\boldsymbol{r}) + c \tag{3.26}$$

leaves the electric field unchanged according to Eq. (3.16). The potential can be fixed unambiguously by specifying its value at a single point. The usual convention is to say that the potential is zero at infinity. This convention is implicit in Eq. (3.17), where it can be seen that $\phi \rightarrow 0$ as $|\boldsymbol{r}| \rightarrow \infty$ provided that the total charge $\int \rho(\boldsymbol{r}')\, d^3\boldsymbol{r}'$ is finite.

3.3 Gauss' law

Consider a single charge located at the origin. The electric field generated by such a charge is given by Eq. (3.11). Suppose that we surround the charge by a concentric spherical surface S of radius r. The flux of the electric field through this surface is given by



$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \oint_S E_r\, dS_r = E_r(r)\, 4\pi r^2 = \frac{q}{4\pi\epsilon_0\, r^2}\; 4\pi r^2 = \frac{q}{\epsilon_0}, \tag{3.27}$$

since the normal to the surface is always parallel to the local electric field. However, we also know from Gauss' theorem that

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \int_V \nabla\cdot\boldsymbol{E}\; d^3\boldsymbol{r}, \tag{3.28}$$

where V is the volume enclosed by surface S. Let us evaluate $\nabla \cdot \boldsymbol{E}$ directly. In Cartesian coordinates the field is written

$$\boldsymbol{E} = \frac{q}{4\pi\epsilon_0}\left(\frac{x}{r^3},\ \frac{y}{r^3},\ \frac{z}{r^3}\right), \tag{3.29}$$

where $r^2 = x^2 + y^2 + z^2$. So,

$$\frac{\partial E_x}{\partial x} = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{r^3} - \frac{3x}{r^4}\frac{x}{r}\right) = \frac{q}{4\pi\epsilon_0}\frac{r^2 - 3x^2}{r^5}. \tag{3.30}$$

Here, use has been made of

$$\frac{\partial r}{\partial x} = \frac{x}{r}. \tag{3.31}$$

Formulae analogous to Eq. (3.30) can be obtained for $\partial E_y/\partial y$ and $\partial E_z/\partial z$. The divergence of the field is given by

$$\nabla \cdot \boldsymbol{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \frac{q}{4\pi\epsilon_0}\frac{3r^2 - 3x^2 - 3y^2 - 3z^2}{r^5} = 0. \tag{3.32}$$

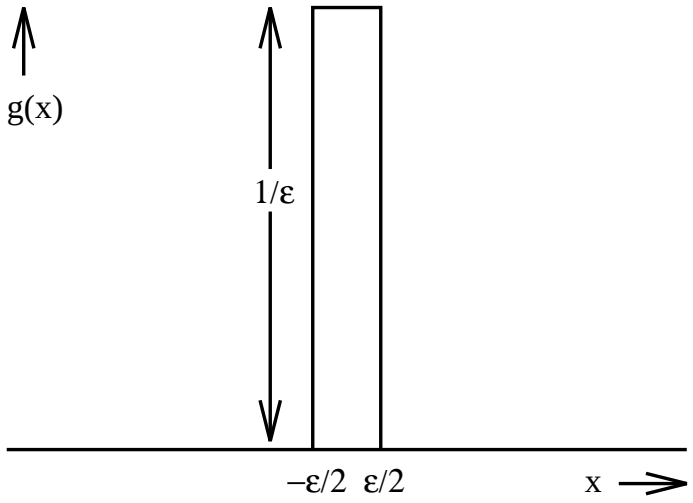
This is a puzzling result! We have from Eqs. (3.27) and (3.28) that

$$\int_V \nabla \cdot \boldsymbol{E}\ d^3\boldsymbol{r} = \frac{q}{\epsilon_0}, \tag{3.33}$$

and yet we have just proved that $\nabla \cdot \boldsymbol{E} = 0$. This paradox can be resolved after a close examination of Eq. (3.32). At the origin ($r = 0$) we find that $\nabla \cdot \boldsymbol{E} = 0/0$, which means that $\nabla \cdot \boldsymbol{E}$ can take any value at this point. Thus, Eqs. (3.32) and (3.33) can be reconciled if $\nabla \cdot \boldsymbol{E}$ is some sort of "spike" function; *i.e.*, it is zero everywhere except arbitrarily close to the origin, where it becomes very large. This must occur in such a manner that the volume integral over the "spike" is finite.

Let us examine how we might construct a one-dimensional "spike" function. Consider the "box-car" function

$$g(x,\epsilon) = \begin{cases} 1/\epsilon & \text{for } |x| < \epsilon/2 \\ 0 & \text{otherwise.} \end{cases} \tag{3.34}$$



It is clear that that

$$\int_{-\infty}^{\infty} g(x,\epsilon)\,dx = 1. \tag{3.35}$$

Now consider the function

$$\delta(x) = \lim_{\epsilon\rightarrow 0} g(x,\epsilon). \tag{3.36}$$

This is zero everywhere except arbitrarily close to $x = 0$. According to Eq. (3.35), it also possess a finite integral;

$$\int_{-\infty}^{\infty} \delta(x)\,dx = 1. \tag{3.37}$$

Thus, $\delta(x)$ has all of the required properties of a "spike" function. The one-dimensional "spike" function $\delta(x)$ is called the "Dirac delta-function" after the Cambridge physicist Paul Dirac who invented it in 1927 while investigating quantum mechanics. The delta-function is an example of what mathematicians call a "generalized function"; it is not well-defined at $x = 0$, but its integral is nevertheless well-defined. Consider the integral

$$\int_{-\infty}^{\infty} f(x)\,\delta(x)\,dx, \tag{3.38}$$

where $f(x)$ is a function which is well-behaved in the vicinity of $x = 0$. Since the delta-function is zero everywhere apart from very close to $x = 0$, it is clear that

$$\int_{-\infty}^{\infty} f(x)\,\delta(x)\,dx = f(0)\int_{-\infty}^{\infty} \delta(x)\,dx = f(0), \tag{3.39}$$

where use has been made of Eq. (3.37). The above equation, which is valid for any well-behaved function $f(x)$, is effectively the definition of a delta-function. A simple change of variables allows us to define $\delta(x - x_0)$, which is a "spike" function centred on $x = x_0$. Equation (3.39) gives

$$\int_{-\infty}^{\infty} f(x)\,\delta(x - x_0)\,dx = f(x_0). \tag{3.40}$$

We actually want a three-dimensional "spike" function; *i.e.*, a function which is zero everywhere apart from close to the origin, and whose volume integral is unity. If we denote this function by $\delta(\boldsymbol{r})$ then it is easily seen that the three-dimensional delta-function is the product of three one-dimensional delta-functions:

$$\delta(\boldsymbol{r}) = \delta(x)\delta(y)\delta(z). \tag{3.41}$$

This function is clearly zero everywhere except the origin. But is its volume integral unity? Let us integrate over a cube of dimensions $2a$ which is centred on the origin and aligned along the Cartesian axes. This volume integral is obviously separable, so that

$$\int \delta(\boldsymbol{r})\,d^3\boldsymbol{r} = \int_{-a}^{a} \delta(x)\,dx \int_{-a}^{a} \delta(y)\,dy \int_{-a}^{a} \delta(z)\,dz. \tag{3.42}$$

The integral can be turned into an integral over all space by taking the limit $a \rightarrow \infty$. However, we know that for one-dimensional delta-functions $\int_{-\infty}^{\infty} \delta(x)\,dx = 1$, so it follows from the above equation that

$$\int \delta(\boldsymbol{r})\,d^3\boldsymbol{r} = 1, \tag{3.43}$$

which is the desired result. A simple generalization of previous arguments yields

$$\int f(\boldsymbol{r})\,\delta(\boldsymbol{r})\,d^3\boldsymbol{r} = f(\boldsymbol{0}), \tag{3.44}$$

where $f(\boldsymbol{r})$ is any well-behaved scalar field. Finally, we can change variables and write

$$\delta(\boldsymbol{r} - \boldsymbol{r}') = \delta(x - x')\delta(y - y')\delta(z - z'), \tag{3.45}$$

which is a three-dimensional "spike" function centred on $\boldsymbol{r} = \boldsymbol{r}'$. It is easily demonstrated that

$$\int f(\boldsymbol{r})\,\delta(\boldsymbol{r}-\boldsymbol{r}')\,d^3\boldsymbol{r} = f(\boldsymbol{r}'). \tag{3.46}$$

Up to now we have only considered volume integrals taken over all space. However, it should be obvious that the above result also holds for integrals over any finite volume V which contains the point $\boldsymbol{r} = \boldsymbol{r}'$. Likewise, the integral is zero if V does not contain $\boldsymbol{r} = \boldsymbol{r}'$.

Let us now return to the problem in hand. The electric field generated by a charge q located at the origin has $\nabla \cdot \boldsymbol{E} = 0$ everywhere apart from the origin, and also satisfies

$$\int_V \nabla \cdot \boldsymbol{E}\; d^3\boldsymbol{r} = \frac{q}{\epsilon_0} \tag{3.47}$$

for a spherical volume V centered on the origin. These two facts imply that

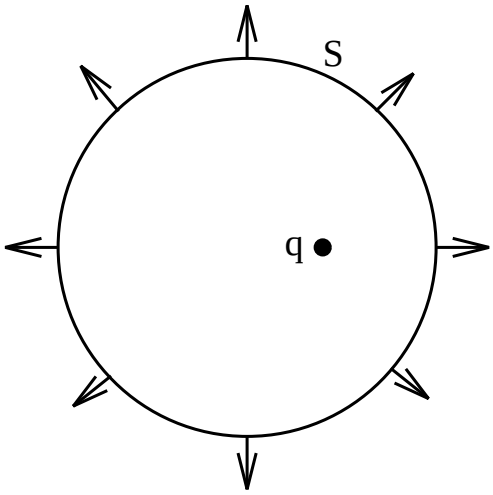
$$\nabla \cdot \boldsymbol{E} = \frac{q}{\epsilon_0}\,\delta(\boldsymbol{r}), \tag{3.48}$$

where use has been made of Eq. (3.43).

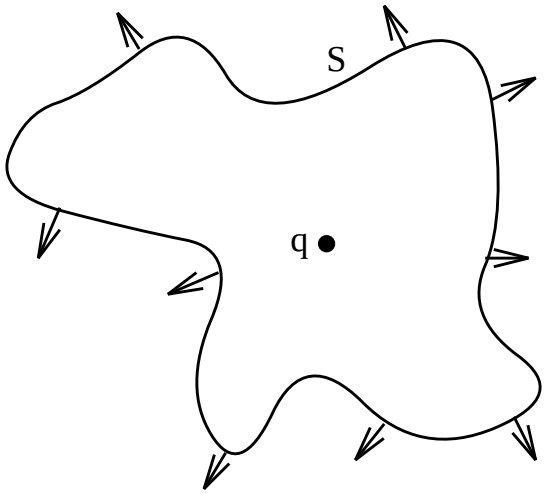
At this stage, you are probably not all that impressed with vector field theory. After all we have just spent an inordinately long time proving something using vector field theory which we previously proved in one line [see Eq. (3.27)] using conventional analysis! Let me now demonstrate the power of vector field theory. Consider, again, a charge q at the origin surrounded by a spherical surface S which is centered on the origin. Suppose that we now displace the surface S, so that it is no longer centered on the origin. What is the flux of the electric field out of S? This is no longer a simple problem for conventional analysis because the normal to the surface is not parallel to the local electric field. However, using vector field theory this problem is no more difficult than the previous one. We have

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{E}\; d^3\boldsymbol{r} \tag{3.49}$$

from Gauss' theorem, plus Eq. (3.48). From these, it is clear that the flux of $\boldsymbol{E}$ out of S is q/ϵ_0 for a spherical surface displaced from the origin. However, the flux becomes zero when the displacement is sufficiently large that the origin is



no longer enclosed by the sphere. It is possible to prove this from conventional analysis, but it is not easy! Suppose that the surface S is not spherical but is instead highly distorted. What now is the flux of $\boldsymbol{E}$ out of S? This is a virtually impossible problem in conventional analysis, but it is easy using vector field theory. Gauss' theorem and Eq. (3.48) tell us that the flux is q/ϵ_0 provided that the surface contains the origin, and that the flux is zero otherwise. This result is independent of the shape of S.



Consider N charges q_i located at $\boldsymbol{r}_i$. A simple generalization of Eq. (3.48)

gives

$$\nabla \cdot \boldsymbol{E} = \sum_{i=1}^{N} \frac{q_i}{\epsilon_0}\, \delta(\boldsymbol{r} - \boldsymbol{r}_i). \tag{3.50}$$

Thus, Gauss' theorem (3.49) implies that

$$\int_S \boldsymbol{E} \cdot d\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{E}\; d^3\boldsymbol{r} = \frac{Q}{\epsilon_0}, \tag{3.51}$$

where Q is the total charge enclosed by the surface S. This result is called Gauss' law and does not depend on the shape of the surface.

Suppose, finally, that instead of having a set of discrete charges we have a continuous charge distribution described by a charge density $\rho(\boldsymbol{r})$. The charge contained in a small rectangular volume of dimensions dx, dy, and dz, located at position $\boldsymbol{r}$ is $Q = \rho(\boldsymbol{r})\, dx\, dy\, dz$. However, if we integrate $\nabla \cdot \boldsymbol{E}$ over this volume element we obtain

$$\nabla \cdot \boldsymbol{E}\; dx\, dy\, dz = \frac{Q}{\epsilon_0} = \frac{\rho\, dx\, dy\, dz}{\epsilon_0}, \tag{3.52}$$

where use has been made of Eq. (3.51). Here, the volume element is assumed to be sufficiently small that $\nabla \cdot \boldsymbol{E}$ does not vary significantly across it. Thus, we obtain

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}. \tag{3.53}$$

This is the first of four field equations, called Maxwell's equations, which together form a complete description of electromagnetism. Of course, our derivation of Eq. (3.53) is only valid for electric fields generated by stationary charge distributions. In principle, additional terms might be required to describe fields generated by moving charge distributions. However, it turns out that this is not the case and that Eq. (3.53) is universally valid.

Equation (3.53) is a differential equation describing the electric field generated by a set of charges. We already know the solution to this equation when the charges are stationary; it is given by Eq. (3.12):

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \rho(\boldsymbol{r}')\, \frac{\boldsymbol{r} - \boldsymbol{r}'}{|\boldsymbol{r} - \boldsymbol{r}'|^3}\, d^3\boldsymbol{r}'. \tag{3.54}$$

Equations (3.53) and (3.54) can be reconciled provided

$$\nabla\cdot\left(\frac{\boldsymbol{r}-\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\right)=-\nabla^2\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)=4\pi\,\delta(\boldsymbol{r}-\boldsymbol{r}'),\tag{3.55}$$

where use has been made of Eq. (3.15). It follows that

$$\begin{aligned}\nabla\cdot\boldsymbol{E}(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int\rho(\boldsymbol{r}')\,\nabla\cdot\left(\frac{\boldsymbol{r}-\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\right)d^3\boldsymbol{r}'\\ &= \int\frac{\rho(\boldsymbol{r}')}{\epsilon_0}\,\delta(\boldsymbol{r}-\boldsymbol{r}')\,d^3\boldsymbol{r}'=\frac{\rho(\boldsymbol{r})}{\epsilon_0},\end{aligned}\tag{3.56}$$

which is the desired result. The most general form of Gauss' law, Eq. (3.51), is obtained by integrating Eq. (3.53) over a volume V surrounded by a surface S and making use of Gauss' theorem:

$$\oint_S\boldsymbol{E}\cdot d\boldsymbol{S}=\frac{1}{\epsilon_0}\int_V\rho(\boldsymbol{r})\,d^3\boldsymbol{r}.\tag{3.57}$$

3.4 Poisson's equation

We have seen that the electric field generated by a set of stationary charges can be written as the gradient of a scalar potential, so that

$$\boldsymbol{E}=-\nabla\phi.\tag{3.58}$$

This equation can be combined with the field equation (3.53) to give a partial differential equation for the scalar potential:

$$\nabla^2\phi=-\frac{\rho}{\epsilon_0}.\tag{3.59}$$

This is an example of a very famous type of partial differential equation known as "Poisson's equation."

In its most general form Poisson's equation is written

$$\nabla^2 u=v,\tag{3.60}$$

where $u(\boldsymbol{r})$ is some scalar potential which is to be determined and $v(\boldsymbol{r})$ is a known "source function." The most common boundary condition applied to this equation is that the potential u is zero at infinity. The solutions to Poisson's equation are completely superposable. Thus, if u_1 is the potential generated by the source function v_1, and u_2 is the potential generated by the source function v_2, so that

$$\nabla^2 u_1 = v_1, \qquad \nabla^2 u_2 = v_2, \tag{3.61}$$

then the potential generated by $v_1 + v_2$ is $u_1 + u_2$, since

$$\nabla^2 (u_1 + u_2) = \nabla^2 u_1 + \nabla^2 u_2 = v_1 + v_2. \tag{3.62}$$

Poisson's equation has this property because it is *linear* in both the potential and the source term.

The fact that the solutions to Poisson's equation are superposable suggests a general method for solving this equation. Suppose that we could construct all of the solutions generated by point sources. Of course, these solutions must satisfy the appropriate boundary conditions. Any general source function can be built up out of a set of suitably weighted point sources, so the general solution of Poisson's equation must be expressible as a weighted sum over the point source solutions. Thus, once we know all of the point source solutions we can construct any other solution. In mathematical terminology we require the solution to

$$\nabla^2 G(\boldsymbol{r}, \boldsymbol{r}') = \delta(\boldsymbol{r} - \boldsymbol{r}') \tag{3.63}$$

which goes to zero as $|\boldsymbol{r}| \rightarrow \infty$. The function $G(\boldsymbol{r}, \boldsymbol{r}')$ is the solution generated by a point source located at position $\boldsymbol{r}'$. In mathematical terminology this function is known as a "Green's function." The solution generated by a general source function $v(\boldsymbol{r})$ is simply the appropriately weighted sum of all of the Green's function solutions:

$$u(\boldsymbol{r}) = \int G(\boldsymbol{r}, \boldsymbol{r}') v(\boldsymbol{r}')\, d^3\boldsymbol{r}'. \tag{3.64}$$

We can easily demonstrate that this is the correct solution:

$$\nabla^2 u(\boldsymbol{r}) = \int \left[\nabla^2 G(\boldsymbol{r}, \boldsymbol{r}')\right]\, v(\boldsymbol{r}')\, d^3\boldsymbol{r}' = \int \delta(\boldsymbol{r} - \boldsymbol{r}')\, v(\boldsymbol{r}')\, d^3\boldsymbol{r}' = v(\boldsymbol{r}). \tag{3.65}$$

Let us return to Eq. (3.59):

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}. \tag{3.66}$$

The Green's function for this equation satisfies Eq. (3.63) with $|G| \rightarrow \infty$ as $|r| \rightarrow 0$. It follows from Eq. (3.55) that

$$G(\boldsymbol{r}, \boldsymbol{r}') = -\frac{1}{4\pi} \frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{3.67}$$

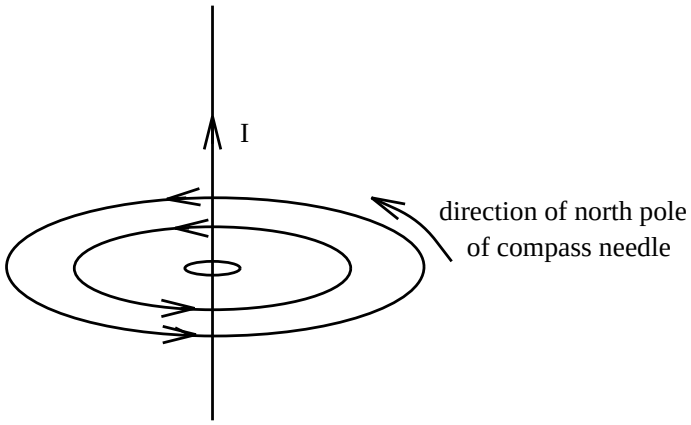
Note from Eq. (3.20) that the Green's function has the same form as the potential generated by a point charge. This is hardly surprising given the definition of a Green's function. It follows from Eq. (3.64) and (3.67) that the general solution to Poisson's equation (3.66) is written

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}'. \tag{3.68}$$

In fact, we have already obtained this solution by another method [see Eq. (3.17)].

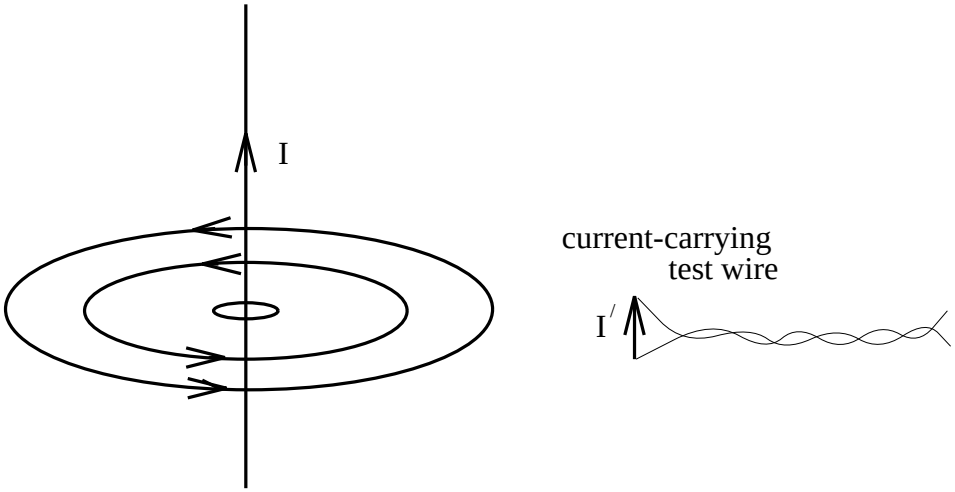
3.5 Ampère's experiments

In 1820 the Danish physicist Hans Christian Ørsted was giving a lecture demonstration of various electrical and magnetic effects. Suddenly, much to his surprise, he noticed that the needle of a compass he was holding was deflected when he moved it close to a current carrying wire. Up until then magnetism has been thought of as solely a property of some rather unusual rocks called loadstones. Word of this discovery spread quickly along the scientific grapevine, and the French physicist Andre Marie Ampère immediately decided to investigate further. Ampère's apparatus consisted (essentially) of a long straight wire carrying an electric current current I. Ampère quickly discovered that the needle of a small compass maps out a series of concentric circular loops in the plane perpendicular to a current carrying wire. The direction of circulation around these magnetic loops is conventionally taken to be the direction in which the *north* pole of the compass needle points. Using this convention, the circulation of the loops is given



by a right-hand rule: if the thumb of the right-hand points along the direction of the current then the fingers of the right-hand circulate in the same sense as the magnetic loops.

Ampère's next series of experiments involved bringing a short test wire, carrying a current I', close to the original wire and investigating the force exerted on the test wire. This experiment is not quite as clear cut as Coulomb's experiment



because, unlike electric charges, electric currents cannot exist as point entities; they have to flow in complete circuits. We must imagine that the circuit which connects with the central wire is sufficiently far away that it has no appreciable influence on the outcome of the experiment. The circuit which connects with the test wire is more problematic. Fortunately, if the feed wires are twisted around each other, as indicated in the diagram, then they effectively cancel one another

out and also do not influence the outcome of the experiment.

Ampère discovered that the force exerted on the test wire is directly proportional to its length. He also made the following observations. If the current in the test wire (*i.e.*, the test current) flows parallel to the current in the central wire then the two wires attract one another. If the current in the test wire is reversed then the two wires repel one another. If the test current points radially towards the central wire (and the current in the central wire flows upwards) then the test wire is subject to a downwards force. If the test current is reversed then the force is upwards. If the test current is rotated in a single plane, so that it starts parallel to the central current and ends up pointing radially towards it, then the force on the test wire is of constant magnitude and is always at right angles to the test current. If the test current is parallel to a magnetic loop then there is no force exerted on the test wire. If the test current is rotated in a single plane, so that it starts parallel to the central current and ends up pointing along a magnetic loop, then the magnitude of the force on the test wire attenuates like $\cos\theta$ (where θ is the angle the current is turned through; $\theta = 0$ corresponds to the case where the test current is parallel to the central current), and its direction is again always at right angles to the test current. Finally, Ampère was able to establish that the attractive force between two parallel current carrying wires is proportional to the product of the two currents, and falls off like one over the perpendicular distance between the wires.

This rather complicated force law can be summed up succinctly in vector notation provided that we define a vector field $\boldsymbol{B}$, called the magnetic field, whose direction is always parallel to the loops mapped out by a small compass. The dependence of the force per unit length, $\boldsymbol{F}$, acting on a test wire with the different possible orientations of the test current is described by

$$\boldsymbol{F} = \boldsymbol{I}' \wedge \boldsymbol{B}, \tag{3.69}$$

where $\boldsymbol{I}'$ is a vector whose direction and magnitude are the same as those of the test current. Incidentally, the SI unit of electric current is the ampere (A), which is the same as a coulomb per second. The SI unit of magnetic field strength is the tesla (T), which is the same as a newton per ampere per meter. The variation of the force per unit length acting on a test wire with the strength of the central current and the perpendicular distance r to the central wire is summed

up by saying that the magnetic field strength is proportional to I and inversely proportional to r. Thus, defining cylindrical polar coordinates aligned along the axis of the central current, we have

$$B_\theta = \frac{\mu_0\, I}{2\pi r}, \tag{3.70}$$

with $B_r = B_z = 0$. The constant of proportionality μ_0 is called the "permeability of free space" and takes the value

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}. \tag{3.71}$$

The concept of a magnetic field allows the calculation of the force on a test wire to be conveniently split into two parts. In the first part, we calculate the magnetic field generated by the current flowing in the central wire. This field circulates in the plane normal to the wire; its magnitude is proportional to the central current and inversely proportional to the perpendicular distance from the wire. In the second part, we use Eq. (3.69) to calculate the force per unit length acting on a short current carrying wire located in the magnetic field generated by the central current. This force is perpendicular to both the magnetic field and the direction of the test current. Note that, at this stage, we have no reason to suppose that the magnetic field has any real existence. It is introduced merely to facilitate the calculation of the force exerted on the test wire by the central wire.

3.6 The Lorentz force

The flow of an electric current down a conducting wire is ultimately due to the motion of electrically charged particles (in most cases, electrons) through the conducting medium. It seems reasonable, therefore, that the force exerted on the wire when it is placed in a magnetic field is really the resultant of the forces exerted on these moving charges. Let us suppose that this is the case.

Let A be the (uniform) cross-sectional area of the wire, and let n be the number density of mobile charges in the conductor. Suppose that the mobile charges each have charge q and velocity $\boldsymbol{v}$. We must assume that the conductor also contains stationary charges, of charge $-q$ and number density n, say, so that

the net charge density in the wire is zero. In most conductors the mobile charges are electrons and the stationary charges are atomic nuclei. The magnitude of the electric current flowing through the wire is simply the number of coulombs per second which flow past a given point. In one second a mobile charge moves a distance v, so all of the charges contained in a cylinder of cross-sectional area A and length v flow past a given point. Thus, the magnitude of the current is $q\,nA\,v$. The direction of the current is the same as the direction of motion of the charges, so the vector current is $\boldsymbol{I}' = q\,nA\,\boldsymbol{v}$. According to Eq. (3.69) the force per unit length acting on the wire is

$$\boldsymbol{F} = q\,nA\,\boldsymbol{v}\wedge\boldsymbol{B}. \tag{3.72}$$

However, a unit length of the wire contains nA moving charges. So, assuming that each charge is subject to an equal force from the magnetic field (we have no reason to suppose otherwise), the force acting on an individual charge is

$$\boldsymbol{f} = q\,\boldsymbol{v}\wedge\boldsymbol{B}. \tag{3.73}$$

We can combine this with Eq. (3.9) to give the force acting on a charge q moving with velocity $\boldsymbol{v}$ in an electric field $\boldsymbol{E}$ and a magnetic field $\boldsymbol{B}$:

$$\boldsymbol{f} = q\,\boldsymbol{E} + q\,\boldsymbol{v}\wedge\boldsymbol{B}. \tag{3.74}$$

This is called the "Lorentz force law" after the Dutch physicist Hendrik Antoon Lorentz who first formulated it. The electric force on a charged particle is parallel to the local electric field. The magnetic force, however, is perpendicular to both the local magnetic field and the particle's direction of motion. No magnetic force is exerted on a stationary charged particle.

The equation of motion of a free particle of charge q and mass m moving in electric and magnetic fields is

$$m\frac{d\boldsymbol{v}}{dt} = q\,\boldsymbol{E} + q\,\boldsymbol{v}\wedge\boldsymbol{B}, \tag{3.75}$$

according to the Lorentz force law. This equation of motion was verified in a famous experiment carried out by the Cambridge physicist J.J. Thompson in 1897. Thompson was investigating "cathode rays," a then mysterious form of

radiation emitted by a heated metal element held at a large negative voltage (i.e. a cathode) with respect to another metal element (*i.e.*, an anode) in an evacuated tube. German physicists held that cathode rays were a form of electromagnetic radiation, whilst British and French physicists suspected that they were, in reality, a stream of charged particles. Thompson was able to demonstrate that the latter view was correct. In Thompson's experiment the cathode rays passed though a region of "crossed" electric and magnetic fields (still in vacuum). The fields were perpendicular to the original trajectory of the rays and were also mutually perpendicular.

Let us analyze Thompson's experiment. Suppose that the rays are originally traveling in the x-direction, and are subject to a uniform electric field E in the z-direction and a uniform magnetic field B in the $-y$-direction. Let us assume, as Thompson did, that cathode rays are a stream of particles of mass m and charge q. The equation of motion of the particles in the z-direction is

$$m\,\frac{d^2z}{dt^2} = q\,(E - vB)\,, \tag{3.76}$$

where v is the velocity of the particles in the x-direction. Thompson started off his experiment by only turning on the electric field in his apparatus and measuring the deflection d of the ray in the z-direction after it had traveled a distance l through the electric field. It is clear from the equation of motion that

$$d = \frac{q}{m}\frac{E\,t^2}{2} = \frac{q}{m}\frac{E\,l^2}{2v^2}, \tag{3.77}$$

where the "time of flight" t is replaced by l/v. This formula is only valid if $d \ll l$, which is assumed to be the case. Next, Thompson turned on the magnetic field in his apparatus and adjusted it so that the cathode ray was no longer deflected. The lack of deflection implies that the net force on the particles in the z-direction is zero. In other words, the electric and magnetic forces balance exactly. It follows from Eq. (3.76) that with a properly adjusted magnetic field strength

$$v = \frac{E}{B}. \tag{3.78}$$

Thus, Eqs. (3.77) and (3.78) and can be combined and rearranged to give the

charge to mass ratio of the particles in terms of measured quantities:

$$\frac{q}{m} = \frac{2dE}{l^2 B^2}. \tag{3.79}$$

Using this method Thompson inferred that cathode rays were made up of negatively charged particles (the sign of the charge is obvious from the direction of the deflection in the electric field) with a charge to mass ratio of -1.7×10^{11} C/kg. A decade later in 1908 the American Robert Millikan performed his famous "oil drop" experiment and discovered that mobile electric charges are quantized in units of -1.6×10^{-19} C. Assuming that mobile electric charges and the particles which make up cathode rays are one and the same thing, Thompson's and Millikan's experiments imply that the mass of these particles is 9.4×10^{-31} kg. Of course, this is the mass of an electron (the modern value is 9.1×10^{-31} kg), and -1.6×10^{-19} C is the charge of an electron. Thus, cathode rays are, in fact, streams of electrons which are emitted from a heated cathode and then accelerated because of the large voltage difference between the cathode and anode.

If a particle is subject to a force $\boldsymbol{f}$ and moves a distance $\delta \boldsymbol{r}$ in a time interval δt then the work done on the particle by the force is

$$\delta W = \boldsymbol{f} \cdot \delta \boldsymbol{r}. \tag{3.80}$$

The power input to the particle from the force field is

$$P = \lim_{\delta t \rightarrow 0} \frac{\delta W}{\delta t} = \boldsymbol{f} \cdot \boldsymbol{v}, \tag{3.81}$$

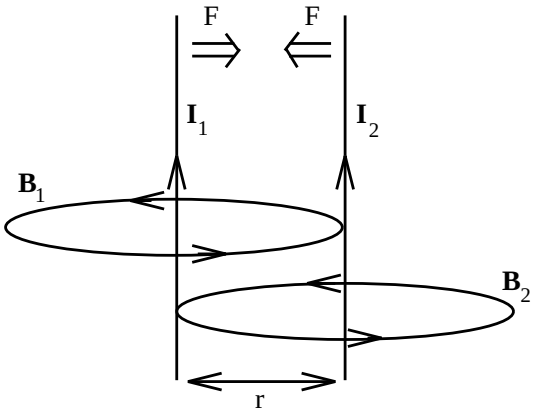
where $\boldsymbol{v}$ is the particle's velocity. It follows from the Lorentz force law, Eq. (3.74), that the power input to a particle moving in electric and magnetic fields is

$$P = q\, \boldsymbol{v} \cdot \boldsymbol{E}. \tag{3.82}$$

Note that a charged particle can gain (or lose) energy from an electric field but not from a magnetic field. This is because the magnetic force is always perpendicular to the particle's direction of motion and, therefore, does no work on the particle [see Eq. (3.80)]. Thus, in particle accelerators magnetic fields are often used to guide particle motion (*e.g.*, in a circle) but the actual acceleration is performed by electric fields.

3.7 Ampère's law

Magnetic fields, like electric fields, are completely superposable. So, if a field $\boldsymbol{B}_1$ is generated by a current I_1 flowing through some circuit, and a field $\boldsymbol{B}_2$ is generated by a current I_2 flowing through another circuit, then when the currents I_1 and I_2 flow through both circuits simultaneously the generated magnetic field is $\boldsymbol{B}_1 + \boldsymbol{B}_2$.



Consider two parallel wires separated by a perpendicular distance r and carrying electric currents I_1 and I_2, respectively. The magnetic field strength at the second wire due to the current flowing in the first wire is $B = \mu_0 I_1/2\pi r$. This field is orientated at right angles to the second wire, so the force per unit length exerted on the second wire is

$$F = \frac{\mu_0 I_1 I_2}{2\pi r}. \tag{3.83}$$

This follows from Eq. (3.69), which is valid for continuous wires as well as short test wires. The force acting on the second wire is directed radially inwards towards the first wire. The magnetic field strength at the first wire due to the current flowing in the second wire is $B = \mu_0 I_2/2\pi r$. This field is orientated at right angles to the first wire, so the force per unit length acting on the first wire is equal and opposite to that acting on the second wire, according to Eq. (3.69). Equation (3.83) is sometimes called "Ampère's law" and is clearly another example of an "action at a distance" law; *i.e.*, if the current in the first wire is suddenly changed then the force on the second wire immediately adjusts, whilst in reality there should be a short time delay, at least as long as the propagation time for a

light signal between the two wires. Clearly, Ampère's law is not strictly correct. However, as long as we restrict our investigations to steady currents it is perfectly adequate.

3.8 Magnetic monopoles?

Suppose that we have an infinite straight wire carrying an electric current I. Let the wire be aligned along the z-axis. The magnetic field generated by such a wire is written

$$\boldsymbol{B} = \frac{\mu_0 I}{2\pi}\left(\frac{-y}{r^2}, \frac{x}{r^2}, 0\right) \tag{3.84}$$

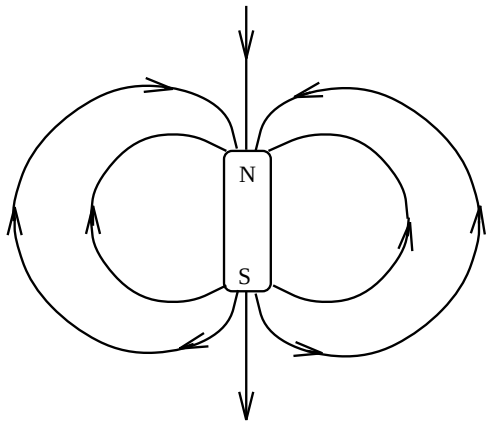
in Cartesian coordinates, where $r = \sqrt{x^2+y^2}$. The divergence of this field is

$$\nabla\cdot\boldsymbol{B} = \frac{\mu_0 I}{2\pi}\left(\frac{2yx}{r^4} - \frac{2xy}{r^4}\right) = 0, \tag{3.85}$$

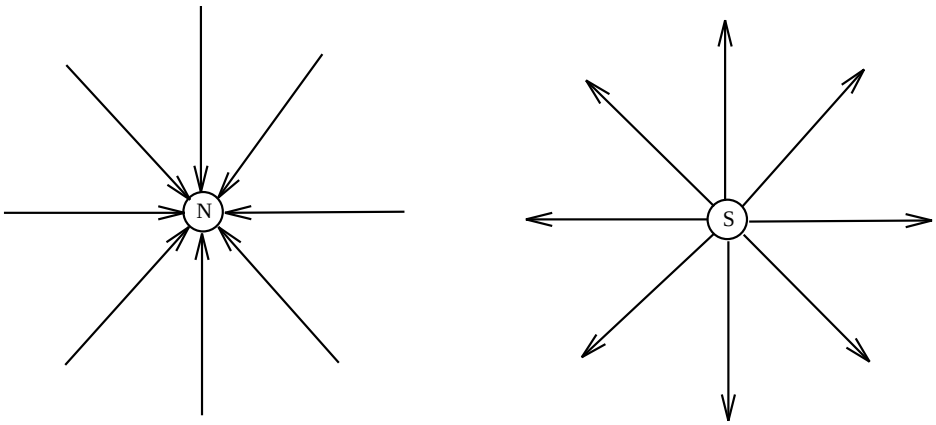
where use has been made of $\partial r/\partial x = x/r$, *etc.* We saw in Section 3.3 that the divergence of the electric field appeared, at first sight, to be zero, but, in reality, it was a delta-function because the volume integral of $\nabla\cdot\boldsymbol{E}$ was non-zero. Does the same sort of thing happen for the divergence of the magnetic field? Well, if we could find a closed surface S for which $\oint_S \boldsymbol{B}\cdot d\boldsymbol{S} \neq 0$ then according to Gauss' theorem $\int_V \nabla\cdot\boldsymbol{B}\, dV \neq 0$ where V is the volume enclosed by S. This would certainly imply that $\nabla\cdot\boldsymbol{B}$ is some sort of delta-function. So, can we find such a surface? The short answer is, no. Consider a cylindrical surface aligned with the wire. The magnetic field is everywhere tangential to the outward surface element, so this surface certainly has zero magnetic flux coming out of it. In fact, it is impossible to invent any closed surface for which $\oint_S \boldsymbol{B}\cdot d\boldsymbol{S} \neq 0$ with $\boldsymbol{B}$ given by Eq. (3.84) (if you do not believe me, try it yourselves!). This suggests that the divergence of a magnetic field generated by steady electric currents really is zero. Admittedly, we have only proved this for infinite straight currents, but, as will be demonstrated presently, it is true in general.

If $\nabla\cdot\boldsymbol{B} = 0$ then $\boldsymbol{B}$ is a *solenoidal* vector field. In other words, field lines of $\boldsymbol{B}$ never begin or end; instead, they form closed loops. This is certainly the case in Eq. (3.84) where the field lines are a set of concentric circles centred

on the z-axis. In fact, the magnetic field lines generated by any set of electric currents form closed loops, as can easily be checked by tracking the magnetic lines of force using a small compass. What about magnetic fields generated by permanent magnets (the modern equivalent of loadstones)? Do they also always form closed loops? Well, we know that a conventional bar magnet has both a north and south magnetic pole (like the Earth). If we track the magnetic field lines with a small compass they all emanate from the south pole, spread out, and eventually reconverge on the north pole. It appears likely (but we cannot prove it with a compass) that the field lines inside the magnet connect from the north to the south pole so as to form closed loops.



Can we produce an isolated north or south magnetic pole; for instance, by snapping a bar magnet in two? A compass needle would always point towards an isolated north pole, so this would act like a negative "magnetic charge." Likewise, a compass needle would always point away from an isolated south pole, so this would act like a positive "magnetic charge." It is clear from the diagram that if we take a closed surface S containing an isolated magnetic pole, which is usually termed a "magnetic monopole," then $\oint_S \boldsymbol{B}\cdot d\boldsymbol{S}\neq 0$; the flux will be positive for an isolated south pole and negative for an isolated north pole. It follows from Gauss' theorem that if $\oint_S \boldsymbol{B}\cdot d\boldsymbol{S}\neq 0$ then $\nabla\cdot\boldsymbol{B}\neq 0$. Thus, the statement that magnetic fields are solenoidal, or that $\nabla\cdot\boldsymbol{B}=0$, is equivalent to the statement that *there are no magnetic monopoles*. It is not clear, *a priori*, that this is a true statement. In fact, it is quite possible to formulate electromagnetism so as to allow for magnetic monopoles. However, as far as we know, there are no magnetic monopoles in the universe. At least, if there are any then they are all



hiding from us! We know that if we try to make a magnetic monopole by snapping a bar magnet in two then we just end up with two smaller bar magnets. If we snap one of these smaller magnets in two then we end up with two even smaller bar magnets. We can continue this process down to the atomic level without ever producing a magnetic monopole. In fact, permanent magnetism is generated by electric currents circulating on the atomic scale, so this type of magnetism is not fundamentally different to the magnetism generated by macroscopic currents.

In conclusion, *all* steady magnetic fields in the universe are generated by circulating electric currents of some description. Such fields are solenoidal; that is, they form closed loops and satisfy the field equation

$$\nabla \cdot \boldsymbol{B} = 0. \tag{3.86}$$

This, incidentally, is the second of Maxwell's equations. Essentially, it says that there are no such things as magnetic monopoles. We have only proved that $\nabla \cdot \boldsymbol{B} = 0$ for steady magnetic fields but, in fact, this is also the case for time dependent fields (see later).

3.9 Ampère's other law

Consider, again, an infinite straight wire aligned along the z-axis and carrying a current I. The field generated by such a wire is written

$$B_\theta = \frac{\mu_0 I}{2\pi r} \tag{3.87}$$

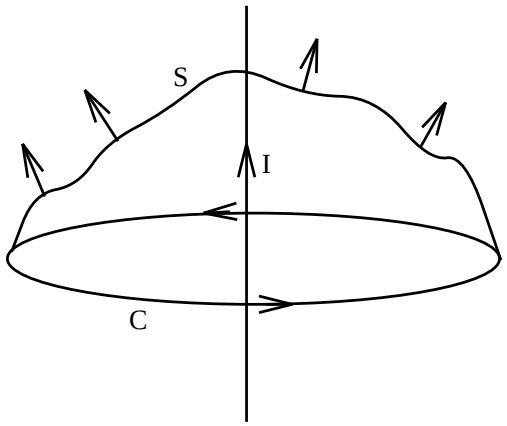
in cylindrical polar coordinates. Consider a circular loop C in the x-y plane which is centred on the wire. Suppose that the radius of this loop is r. Let us evaluate the line integral $\oint_C \boldsymbol{B}\cdot d\boldsymbol{l}$. This integral is easy to perform because the magnetic field is always parallel to the line element. We have

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \oint B_\theta\, r d\theta = \mu_0 I. \tag{3.88}$$

However, we know from Stokes' theorem that

$$\oint_S \boldsymbol{B}\cdot d\boldsymbol{l} = \int_S \nabla\wedge\boldsymbol{B}\cdot d\boldsymbol{S}, \tag{3.89}$$

where S is any surface attached to the loop C. Let us evaluate $\nabla\wedge\boldsymbol{B}$ directly.



According to Eq. (3.84):

$$\begin{aligned}
(\nabla\wedge\boldsymbol{B})_x &= \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} = 0,\\
(\nabla\wedge\boldsymbol{B})_y &= \frac{\partial B_x}{\partial z} - \frac{\partial B_z}{\partial x} = 0,\\
(\nabla\wedge\boldsymbol{B})_z &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \frac{\mu_0 I}{2\pi}\left(\frac{1}{r^2} - \frac{2x^2}{r^4} + \frac{1}{r^2} - \frac{2y^2}{r^4}\right) = 0,
\end{aligned} \tag{3.90}$$

where use has been made of $\partial r/\partial x = x/r$, *etc.* We now have a problem. Equations (3.88) and (3.89) imply that

$$\oint_S \nabla\wedge\boldsymbol{B}\cdot d\boldsymbol{S} = \mu_0 I; \tag{3.91}$$

but we have just demonstrated that $\nabla \wedge \boldsymbol{B} = \boldsymbol{0}$. This problem is very reminiscent of the difficulty we had earlier with $\nabla \cdot \boldsymbol{E}$. Recall that $\int_V \nabla \cdot \boldsymbol{E}\, dV = q/\epsilon_0$ for a volume V containing a discrete charge q, but that $\nabla \cdot \boldsymbol{E} = 0$ at a general point. We got around this problem by saying that $\nabla \cdot \boldsymbol{E}$ is a three-dimensional delta-function whose "spike" is coincident with the location of the charge. Likewise, we can get around our present difficulty by saying that $\nabla \wedge \boldsymbol{B}$ is a two-dimensional delta-function. A three-dimensional delta-function is a singular (but integrable) *point* in space, whereas a two-dimensional delta-function is a singular *line* in space. It is clear from an examination of Eqs. (3.90) that the only component of $\nabla \wedge \boldsymbol{B}$ which can be singular is the z-component, and that this can only be singular on the z-axis (*i.e.*, $r = 0$). Thus, the singularity coincides with the location of the current, and we can write

$$\nabla \wedge \boldsymbol{B} = \mu_0 I\, \delta(x)\delta(y)\, \hat{\boldsymbol{z}}. \tag{3.92}$$

The above equation certainly gives $(\nabla \wedge \boldsymbol{B})_x = (\nabla \wedge \boldsymbol{B})_y = 0$, and $(\nabla \wedge \boldsymbol{B})_z = 0$ everywhere apart from the z-axis, in accordance with Eqs. (3.90). Suppose that we integrate over a plane surface S connected to the loop C. The surface element is $d\boldsymbol{S} = dx\, dy\, \hat{\boldsymbol{z}}$, so

$$\int_S \nabla \wedge \boldsymbol{B} \cdot d\boldsymbol{S} = \mu_0 I \int\int \delta(x)\delta(y)\, dx\, dy \tag{3.93}$$

where the integration is performed over the region $\sqrt{x^2 + y^2} \leq r$. However, since the only part of S which actually contributes to the surface integral is the bit which lies infinitesimally close to the z-axis, we can integrate over all x and y without changing the result. Thus, we obtain

$$\int_S \nabla \wedge \boldsymbol{B} \cdot d\boldsymbol{S} = \mu_0 I \int_{-\infty}^{\infty} \delta(x)\, dx \int_{-\infty}^{\infty} \delta(y)\, dy = \mu_0 I, \tag{3.94}$$

which is in agreement with Eq. (3.91).

You might again be wondering why we have gone to so much trouble to prove something using vector field theory which can be demonstrated in one line via conventional analysis [see Eq. (3.88)]. The answer, of course, is that the vector field result is easily generalized whereas the conventional result is just a special

case. For instance, suppose that we distort our simple circular loop C so that it is no longer circular or even lies in one plane. What now is the line integral of $\boldsymbol{B}$ around the loop? This is no longer a simple problem for conventional analysis, because the magnetic field is not parallel to the line element of the loop. However, according to Stokes' theorem

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \int_S \nabla\wedge\boldsymbol{B}\cdot d\boldsymbol{S}, \tag{3.95}$$

with $\nabla\wedge\boldsymbol{B}$ given by Eq. (3.92). Note that the only part of S which contributes to the surface integral is an infinitesimal region centered on the z-axis. So, as long as S actually intersects the z-axis it does not matter what shape the rest the surface is, we always get the same answer for the surface integral, namely

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \int_S \nabla\wedge\boldsymbol{B}\cdot d\boldsymbol{S} = \mu_0 I. \tag{3.96}$$

Thus, provided the curve C circulates the z-axis, and therefore any surface S attached to C intersects the z-axis, the line integral $\oint_C \boldsymbol{B}\cdot d\boldsymbol{l}$ is equal to $\mu_0 I$. Of course, if C does not circulate the z-axis then an attached surface S does not intersect the z-axis and $\oint_C \boldsymbol{B}\cdot d\boldsymbol{l}$ is zero. There is one more proviso. The line integral $\oint_C \boldsymbol{B}\cdot d\boldsymbol{l}$ is $\mu_0 I$ for a loop which circulates the z-axis in a clockwise direction (looking up the z-axis). However, if the loop circulates in an anti-clockwise direction then the integral is $-\mu_0 I$. This follows because in the latter case the z-component of the surface element $d\boldsymbol{S}$ is oppositely directed to the current flow at the point where the surface intersects the wire.

Let us now consider N wires directed along the z-axis, with coordinates (x_i, y_i) in the x-y plane, each carrying a current I_i in the positive z-direction. It is fairly obvious that Eq. (3.92) generalizes to

$$\nabla\wedge\boldsymbol{B} = \mu_0 \sum_{i=1}^{N} I_i\,\delta(x-x_i)\delta(y-y_i)\,\hat{\boldsymbol{z}}. \tag{3.97}$$

If we integrate the magnetic field around some closed curve C, which can have any shape and does not necessarily lie in one plane, then Stokes' theorem and the above equation imply that

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \int_S \nabla\wedge\boldsymbol{B}\cdot d\boldsymbol{S} = \mu_0\,\mathcal{I}, \tag{3.98}$$

where $\mathcal{I}$ is the total current enclosed by the curve. Again, if the curve circulates the ith wire in a clockwise direction (looking down the direction of current flow) then the wire contributes I_i to the aggregate current $\mathcal{I}$. On the other hand, if the curve circulates in an anti-clockwise direction then the wire contributes $-I_i$. Finally, if the curve does not circulate the wire at all then the wire contributes nothing to $\mathcal{I}$.

Equation (3.97) is a field equation describing how a set of z-directed current carrying wires generate a magnetic field. These wires have zero-thickness, which implies that we are trying to squeeze a finite amount of current into an infinitesimal region. This accounts for the delta-functions on the right-hand side of the equation. Likewise, we obtained delta-functions in Section 3.3 because we were dealing with point charges. Let us now generalize to the more realistic case of diffuse currents. Suppose that the z-current flowing through a small rectangle in the x-y plane, centred on coordinates $(x,\, y)$ and of dimensions dx and dy, is $j_z(x,y)\,dx\,dy$. Here, j_z is termed the current density in the z-direction. Let us integrate $(\nabla\wedge\boldsymbol{B})_z$ over this rectangle. The rectangle is assumed to be sufficiently small that $(\nabla\wedge\boldsymbol{B})_z$ does not vary appreciably across it. According to Eq. (3.98) this integral is equal to μ_0 times the total z-current flowing through the rectangle. Thus,

$$(\nabla\wedge\boldsymbol{B})_z\,dx\,dy=\mu_0 j_z\,dx\,dy, \tag{3.99}$$

which implies that

$$(\nabla\wedge\boldsymbol{B})_z=\mu_0 j_z. \tag{3.100}$$

Of course, there is nothing special about the z-axis. Suppose we have a set of diffuse currents flowing in the x-direction. The current flowing through a small rectangle in the y-z plane, centred on coordinates $(y,\, z)$ and of dimensions dy and dz, is given by $j_x(y,z)\,dy\,dz$, where j_x is the current density in the x-direction. It is fairly obvious that we can write

$$(\nabla\wedge\boldsymbol{B})_x=\mu_0 j_x, \tag{3.101}$$

with a similar equation for diffuse currents flowing along the y-axis. We can combine these equations with Eq. (3.100) to form a single vector field equation which describes how electric currents generate magnetic fields:

$$\nabla\wedge\boldsymbol{B}=\mu_0\boldsymbol{j}, \tag{3.102}$$

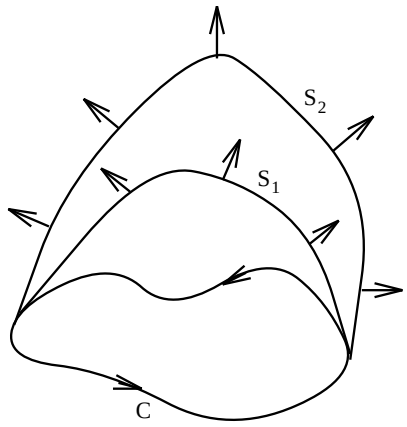
where $\boldsymbol{j} = (j_x, j_y, j_z)$ is the vector current density. This is the third Maxwell equation. The electric current flowing through a small area $d\boldsymbol{S}$ located at position $\boldsymbol{r}$ is $\boldsymbol{j}(\boldsymbol{r}) \cdot d\boldsymbol{S}$. Suppose that space is filled with particles of charge q, number density $n(\boldsymbol{r})$, and velocity $\boldsymbol{v}(\boldsymbol{r})$. The charge density is given by $\rho(\boldsymbol{r}) = qn$. The current density is given by $\boldsymbol{j}(\boldsymbol{r}) = qn\boldsymbol{v}$ and is obviously a proper vector field (velocities are proper vectors since they are ultimately derived from displacements).

If we form the line integral of $\boldsymbol{B}$ around some general closed curve C, making use of Stokes' theorem and the field equation (3.102), then we obtain

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{j} \cdot d\boldsymbol{S}. \tag{3.103}$$

In other words, the line integral of the magnetic field around any closed loop C is equal to μ_0 times the flux of the current density through C. This result is called Ampère's (other) law. If the currents flow in zero-thickness wires then Ampère's law reduces to Eq. (3.98).

The flux of the current density through C is evaluated by integrating $\boldsymbol{j} \cdot d\boldsymbol{S}$ over any surface S attached to C. Suppose that we take two different surfaces S_1 and S_2. It is clear that if Ampère's law is to make any sense then the surface integral $\int_{S_1} \boldsymbol{j} \cdot d\boldsymbol{S}$ had better equal the integral $\int_{S_2} \boldsymbol{j} \cdot d\boldsymbol{S}$. That is, when we work out the flux of the current though C using two different attached surfaces then we had better get the same answer, otherwise Eq. (3.103) is wrong. We saw in



Section 2 that if the integral of a vector field $\boldsymbol{A}$ over some surface attached to a loop depends only on the loop, and is independent of the surface which spans it,

then this implies that $\nabla \cdot \boldsymbol{A} = 0$. The flux of the current density through any loop C is is calculated by evaluating the integral $\int_S \boldsymbol{j} \cdot d\boldsymbol{S}$ for any surface S which spans the loop. According to Ampère's law, this integral depends only on C and is completely independent of S (*i.e.*, it is equal to the line integral of $\boldsymbol{B}$ around C, which depends on C but not on S). This implies that $\nabla \cdot \boldsymbol{j} = 0$. In fact, we can obtain this relation directly from the field equation (3.102). We know that the divergence of a curl is automatically zero, so taking the divergence of Eq. (3.102) we obtain

$$\nabla \cdot \boldsymbol{j} = 0. \tag{3.104}$$

We have shown that if Ampère's law is to make any sense then we need $\nabla \cdot \boldsymbol{j} = 0$. Physically this implies that the net current flowing through any closed surface S is zero. Up to now we have only considered stationary charges and steady currents. It is clear that if all charges are stationary and all currents are steady then there can be no net current flowing through a closed surface S, since this would imply a build up of charge in the volume V enclosed by S. In other words, as long as we restrict our investigation to stationary charges and steady currents then we expect $\nabla \cdot \boldsymbol{j} = 0$, and Ampère's law makes sense. However, suppose that we now relax this restriction. Suppose that some of the charges in a volume V decide to move outside V. Clearly, there will be a non-zero net flux of electric current through the bounding surface S whilst this is happening. This implies from Gauss' theorem that $\nabla \cdot \boldsymbol{j} \neq 0$. Under these circumstances Ampère's law collapses in a heap. We shall see later that we can rescue Ampère's law by adding an extra term involving a time derivative to the right-hand side of the field equation (3.102). For steady state situations (*i.e.*, $\partial/\partial t = 0$) this extra term can be neglected. Thus, the field equation $\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j}$ is, in fact, only two-thirds of Maxwell's third equation; there is a term missing on the right-hand side.

We have now derived two field equations involving magnetic fields (actually, we have only derived one and two-thirds):

$$\nabla \cdot \boldsymbol{B} = 0, \tag{3.105a}$$

$$\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j}. \tag{3.105b}$$

We obtained these equations by looking at the fields generated by infinitely long, straight, steady currents. This, of course, is a rather special class of currents.

We should now go back and repeat the process for general currents. In fact, if we did this we would find that the above field equations still hold (provided that the currents are steady). Unfortunately, this demonstration is rather messy and extremely tedious. There is a better approach. Let us *assume* that the above field equations are valid for any set of steady currents. We can then, with relatively little effort, use these equations to generate the correct formula for the magnetic field induced by a general set of steady currents, thus proving that our assumption is correct. More of this later.

3.10 Helmholtz's theorem: A mathematical digression

Let us now embark on a slight mathematical digression. Up to now we have only studied the electric and magnetic fields generated by stationary charges and steady currents. We have found that these fields are describable in terms of four field equations:

$$\begin{aligned} \nabla\cdot\boldsymbol{E} &= \frac{\rho}{\epsilon_0},\\ \nabla\wedge\boldsymbol{E} &= \boldsymbol{0} \end{aligned} \tag{3.106}$$

for electric fields, and

$$\begin{aligned} \nabla\cdot\boldsymbol{B} &= 0,\\ \nabla\wedge\boldsymbol{B} &= \mu_0\boldsymbol{j} \end{aligned} \tag{3.107}$$

for magnetic fields. There are no other field equations. This strongly suggests that if we know the divergence and the curl of a vector field then we know everything there is to know about the field. In fact, this is the case. There is a mathematical theorem which sums this up. It is called Helmholtz's theorem after the German polymath Hermann Ludwig Ferdinand von Helmholtz.

Let us start with scalar fields. Field equations are a type of differential equation; *i.e.*, they deal with the infinitesimal differences in quantities between neighbouring points. The question is, what differential equation completely specifies a scalar field? This is easy. Suppose that we have a scalar field ϕ and a field

equation which tells us the gradient of this field at all points: something like

$$\nabla\phi = \boldsymbol{A}, \tag{3.108}$$

where $\boldsymbol{A}(\boldsymbol{r})$ is a vector field. Note that we need $\nabla \wedge \boldsymbol{A} = \boldsymbol{0}$ for self consistency, since the curl of a gradient is automatically zero. The above equation completely specifies ϕ once we are given the value of the field at a single point, P say. Thus,

$$\phi(Q) = \phi(P) + \int_P^Q \nabla\phi \cdot d\boldsymbol{l} = \phi(P) + \int_P^Q \boldsymbol{A} \cdot d\boldsymbol{l}, \tag{3.109}$$

where Q is a general point. The fact that $\nabla\wedge\boldsymbol{A} = \boldsymbol{0}$ means that $\boldsymbol{A}$ is a conservative field which guarantees that the above equation gives a unique value for ϕ at a general point in space.

Suppose that we have a vector field $\boldsymbol{F}$. How many differential equations do we need to completely specify this field? Hopefully, we only need two: one giving the divergence of the field and one giving its curl. Let us test this hypothesis. Suppose that we have two field equations:

$$\nabla \cdot \boldsymbol{F} = D, \tag{3.110a}$$

$$\nabla \wedge \boldsymbol{F} = \boldsymbol{C}, \tag{3.110b}$$

where D is a scalar field and $\boldsymbol{C}$ is a vector field. For self-consistency we need

$$\nabla \cdot \boldsymbol{C} = 0, \tag{3.111}$$

since the divergence of a curl is automatically zero. The question is, do these two field equations plus some suitable boundary conditions completely specify $\boldsymbol{F}$? Suppose that we write

$$\boldsymbol{F} = -\nabla U + \nabla \wedge \boldsymbol{W}. \tag{3.112}$$

In other words, we are saying that a general field $\boldsymbol{F}$ is the sum of a conservative field, ∇U, and a solenoidal field, $\nabla \wedge \boldsymbol{W}$. This sounds plausible, but it remains to be proved. Let us start by taking the divergence of the above equation and making use of Eq. (3.110a). We get

$$\nabla^2 U = -D. \tag{3.113}$$

Note that the vector field $\boldsymbol{W}$ does not figure in this equation because the divergence of a curl is automatically zero. Let us now take the curl of Eq. (3.112):

$$\nabla \wedge \boldsymbol{F} = \nabla \wedge \nabla \wedge \boldsymbol{W} = \nabla(\nabla \cdot \boldsymbol{W}) - \nabla^2 \boldsymbol{W} = -\nabla^2 \boldsymbol{W}. \tag{3.114}$$

Here, we assume that the divergence of $\boldsymbol{W}$ is zero. This is another thing which remains to be proved. Note that the scalar field U does not figure in this equation because the curl of a divergence is automatically zero. Using Eq. (3.110b) we get

$$\begin{aligned} \nabla^2 W_x &= -C_x, \\ \nabla^2 W_y &= -C_y, \\ \nabla^2 W_z &= -C_z, \end{aligned} \tag{3.115}$$

So, we have transformed our problem into four differential equations, Eq. (3.113) and Eqs. (3.115), which we need to solve. Let us look at these equations. We immediately notice that they all have exactly the same form. In fact, they are all versions of Poisson's equation. We can now make use of a principle made famous by Richard P. Feynman: "the same equations have the same solutions." Recall that earlier on we came across the following equation:

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}, \tag{3.116}$$

where ϕ is the electrostatic potential and ρ is the charge density. We proved that the solution to this equation, with the boundary condition that ϕ goes to zero at infinity, is

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}'. \tag{3.117}$$

Well, if the same equations have the same solutions, and Eq. (3.117) is the solution to Eq. (3.116), then we can immediately write down the solutions to Eq. (3.113) and Eqs. (3.115). We get

$$U(\boldsymbol{r}) = \frac{1}{4\pi} \int \frac{D(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}', \tag{3.118}$$

and

$$W_x(\boldsymbol{r}) \;=\; \frac{1}{4\pi} \int \frac{C_x(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}',$$

$$W_y(\boldsymbol{r}) = \frac{1}{4\pi}\int \frac{C_y(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}', \tag{3.119}$$

$$W_z(\boldsymbol{r}) = \frac{1}{4\pi}\int \frac{C_z(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'.$$

The last three equations can be combined to form a single vector equation:

$$\boldsymbol{W}(\boldsymbol{r}) = \frac{1}{4\pi}\int \frac{\boldsymbol{C}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.120}$$

We assumed earlier that $\nabla\cdot\boldsymbol{W}=0$. Let us check to see if this is true. Note that

$$\frac{\partial}{\partial x}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) = -\frac{x-x'}{|\boldsymbol{r}-\boldsymbol{r}'|^3} = \frac{x'-x}{|\boldsymbol{r}-\boldsymbol{r}'|^3} = -\frac{\partial}{\partial x'}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right), \tag{3.121}$$

which implies that

$$\nabla\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) = -\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right), \tag{3.122}$$

where ∇' is the operator $(\partial/\partial x', \partial/\partial y', \partial/\partial z')$. Taking the divergence of Eq. (3.120) and making use of the above relation, we obtain

$$\nabla\cdot\boldsymbol{W} = \frac{1}{4\pi}\int \boldsymbol{C}(\boldsymbol{r}')\cdot\nabla\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)\,d^3\boldsymbol{r}' = -\frac{1}{4\pi}\int \boldsymbol{C}(\boldsymbol{r}')\cdot\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)\,d^3\boldsymbol{r}'. \tag{3.123}$$

Now

$$\int_{-\infty}^{\infty} g\,\frac{\partial f}{\partial x}\,dx = [gf]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f\,\frac{\partial g}{\partial x}\,dx. \tag{3.124}$$

However, if $gf\rightarrow 0$ as $x\rightarrow\pm\infty$ then we can neglect the first term on the right-hand side of the above equation and write

$$\int_{-\infty}^{\infty} g\,\frac{\partial f}{\partial x}\,dx = -\int_{-\infty}^{\infty} f\,\frac{\partial g}{\partial x}\,dx. \tag{3.125}$$

A simple generalization of this result yields

$$\int \boldsymbol{g}\cdot\nabla f\,d^3\boldsymbol{r} = -\int f\,\nabla\cdot\boldsymbol{g}\,d^3\boldsymbol{r}, \tag{3.126}$$

provided that $g_x f \rightarrow 0$ as $|\boldsymbol{r}| \rightarrow \infty$, *etc.* Thus, we can deduce that

$$\nabla \cdot \boldsymbol{W} = \frac{1}{4\pi} \int \frac{\nabla' \cdot \boldsymbol{C}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}' \tag{3.127}$$

from Eq. (3.123), provided $|\boldsymbol{C}(\boldsymbol{r})|$ is bounded as $|\boldsymbol{r}| \rightarrow \infty$. However, we have already shown that $\nabla \cdot \boldsymbol{C} = 0$ from self-consistency arguments, so the above equation implies that $\nabla \cdot \boldsymbol{W} = 0$, which is the desired result.

We have constructed a vector field $\boldsymbol{F}$ which satisfies Eqs. (3.110) and behaves sensibly at infinity; i.e., $|\boldsymbol{F}| \rightarrow 0$ as $|\boldsymbol{r}| \rightarrow \infty$. But, is our solution the only possible solution of Eqs. (3.110) with sensible boundary conditions at infinity? Another way of posing this question is to ask whether there are any solutions of

$$\nabla^2 U = 0, \quad \nabla^2 W_i = 0, \tag{3.128}$$

where i denotes x, y, or z, which are bounded at infinity. If there are then we are in trouble, because we can take our solution and add to it an arbitrary amount of a vector field with zero divergence and zero curl and thereby obtain another solution which also satisfies physical boundary conditions. This would imply that our solution is not unique. In other words, it is not possible to unambiguously reconstruct a vector field given its divergence, its curl, and physical boundary conditions. Fortunately, the equation

$$\nabla^2 \phi = 0, \tag{3.129}$$

which is called Laplace's equation, has a very nice property: its solutions are *unique*. That is, if we can find a solution to Laplace's equation which satisfies the boundary conditions then we are guaranteed that this is the only solution. We shall prove this later on in the course. Well, let us invent some solutions to Eqs. (3.128) which are bounded at infinity. How about

$$U = W_i = 0? \tag{3.130}$$

These solutions certainly satisfy Laplace's equation and are well-behaved at infinity. Because the solutions to Laplace's equations are unique, we know that Eqs. (3.130) are the only solutions to Eqs. (3.128). This means that there is no vector field which satisfies physical boundary equations at infinity and has

zero divergence and zero curl. In other words, our solution to Eqs. (3.110) is the *only* solution. Thus, we have unambiguously reconstructed the vector field $\boldsymbol{F}$ given its divergence, its curl, and sensible boundary conditions at infinity. This is Helmholtz's theorem.

We have just proved a number of very useful, and also very important, points. First, according to Eq. (3.112), a general vector field can be written as the sum of a conservative field and a solenoidal field. Thus, we ought to be able to write electric and magnetic fields in this form. Second, a general vector field which is zero at infinity is completely specified once its divergence and its curl are given. Thus, we can guess that the laws of electromagnetism can be written as four field equations,

$$\begin{aligned} \nabla\cdot\boldsymbol{E} &= \textit{something},\\ \nabla\wedge\boldsymbol{E} &= \boldsymbol{something},\\ \nabla\cdot\boldsymbol{B} &= \textit{something},\\ \nabla\wedge\boldsymbol{B} &= \boldsymbol{something}, \end{aligned} \tag{3.131}$$

without knowing the first thing about electromagnetism (other than the fact that it deals with two vector fields). Of course, Eq. (3.106) and (3.107) are of exactly this form. We also know that there are only four field equations, since the above equations are sufficient to completely reconstruct both $\boldsymbol{E}$ and $\boldsymbol{B}$. Furthermore, we know that we can solve the field equations without even knowing what the right-hand sides look like. After all, we solved Eqs. (3.110) for completely general right-hand sides. (Actually, the right-hand sides have to go to zero at infinity otherwise integrals like Eq. (3.118) blow up.) We also know that any solutions we find are unique. In other words, there is only one possible steady electric and magnetic field which can be generated by a given set of stationary charges and steady currents. The third thing which we proved was that if the right-hand sides of the above field equations are all zero then the only physical solution is $\boldsymbol{E}=\boldsymbol{B}=\boldsymbol{0}$. This implies that steady electric and magnetic fields cannot generate themselves, instead they have to be generated by stationary charges and steady currents. So, if we come across a steady electric field we know that if we trace the field lines back we shall eventually find a charge. Likewise, a steady magnetic field implies that there is a steady current flowing somewhere. All of these results

follow from vector field theory, *i.e.*, from the general properties of fields in three dimensional space, prior to any investigation of electromagnetism.

3.11 The magnetic vector potential

Electric fields generated by stationary charges obey

$$\nabla \wedge \boldsymbol{E} = \boldsymbol{0}. \tag{3.132}$$

This immediately allows us to write

$$\boldsymbol{E} = -\nabla \phi, \tag{3.133}$$

since the curl of a gradient is automatically zero. In fact, whenever we come across an irrotational vector field in physics we can always write it as the gradient of some scalar field. This is clearly a useful thing to do since it enables us to replace a vector field by a much simpler scalar field. The quantity ϕ in the above equation is known as the electric scalar potential.

Magnetic fields generated by steady currents (and unsteady currents, for that matter) satisfy

$$\nabla \cdot \boldsymbol{B} = 0. \tag{3.134}$$

This immediately allows us to write

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A}, \tag{3.135}$$

since the divergence of a curl is automatically zero. In fact, whenever we come across a solenoidal vector field in physics we can always write it as the curl of some other vector field. This is not an obviously useful thing to do, however, since it only allows us to replace one vector field by another. Nevertheless, Eq. (3.135) is probably the single most useful equation we shall come across in this lecture course. The quantity $\boldsymbol{A}$ is known as the magnetic vector potential.

We know from Helmholtz's theorem that a vector field is fully specified by its divergence and its curl. The curl of the vector potential gives us the magnetic field via Eq. (3.135). However, the divergence of $\boldsymbol{A}$ has no physical significance.

In fact, we are completely free to choose $\nabla \cdot \boldsymbol{A}$ to be whatever we like. Note that, according to Eq. (3.135), the magnetic field is invariant under the transformation

$$\boldsymbol{A} \rightarrow \boldsymbol{A} - \nabla \psi. \tag{3.136}$$

In other words, the vector potential is undetermined to the gradient of a scalar field. This is just another way of saying that we are free to choose $\nabla \cdot \boldsymbol{A}$. Recall that the electric scalar potential is undetermined to an arbitrary additive constant, since the transformation

$$\phi \rightarrow \phi + c \tag{3.137}$$

leaves the electric field invariant in Eq. (3.133). The transformations (3.136) and (3.137) are examples of what mathematicians call "gauge transformations." The choice of a particular function ψ or a particular constant c is referred to as a choice of the gauge. We are free to fix the gauge to be whatever we like. The most sensible choice is the one which makes our equations as simple as possible. The usual gauge for the scalar potential ϕ is such that $\phi \rightarrow 0$ at infinity. The usual gauge for $\boldsymbol{A}$ is such that

$$\nabla \cdot \boldsymbol{A} = 0. \tag{3.138}$$

This particular choice is known as the "Coulomb gauge."

It is obvious that we can always add a constant to ϕ so as to make it zero at infinity. But it is not at all obvious that we can always perform a gauge transformation such as to make $\nabla \cdot \boldsymbol{A}$ zero. Suppose that we have found some vector field $\boldsymbol{A}$ whose curl gives the magnetic field but whose divergence in non-zero. Let

$$\nabla \cdot \boldsymbol{A} = v(\boldsymbol{r}). \tag{3.139}$$

The question is, can we find a scalar field ψ such that after we perform the gauge transformation (3.136) we are left with $\nabla \cdot \boldsymbol{A} = 0$. Taking the divergence of Eq. (3.136) it is clear that we need to find a function ψ which satisfies

$$\nabla^2 \psi = v. \tag{3.140}$$

But this is just Poisson's equation (again!). We know that we can always find a unique solution of this equation (see Section 3.10). This proves that, in practice, we can always set the divergence of $\boldsymbol{A}$ equal to zero.

Let us consider again an infinite straight wire directed along the z-axis and carrying a current I. The magnetic field generated by such a wire is written

$$\boldsymbol{B} = \frac{\mu_0 I}{2\pi}\left(\frac{-y}{r^2}, \frac{x}{r^2}, 0\right). \tag{3.141}$$

We wish to find a vector potential $\boldsymbol{A}$ whose curl is equal to the above magnetic field and whose divergence is zero. It is not difficult to see that

$$\boldsymbol{A} = -\frac{\mu_0 I}{4\pi}\left(0, 0, \ln(x^2+y^2)\right) \tag{3.142}$$

fits the bill. Note that the vector potential is parallel to the direction of the current. This would seem to suggest that there is a more direct relationship between the vector potential and the current than there is between the magnetic field and the current. The potential is not very well behaved on the z-axis, but this is just because we are dealing with an infinitely thin current.

Let us take the curl of Eq. (3.135). We find that

$$\nabla \wedge \boldsymbol{B} = \nabla \wedge \nabla \wedge \boldsymbol{A} = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = -\nabla^2 \boldsymbol{A}, \tag{3.143}$$

where use has been made of the Coulomb gauge condition (3.138). We can combine the above relation with the field equation (3.102) to give

$$\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{j}. \tag{3.144}$$

Writing this in component form, we obtain

$$\begin{aligned} \nabla^2 A_x &= -\mu_0 j_x, \\ \nabla^2 A_y &= -\mu_0 j_y, \\ \nabla^2 A_z &= -\mu_0 j_z. \end{aligned} \tag{3.145}$$

But, this is just Poisson's equation three times over. We can immediately write the unique solutions to the above equations:

$$A_x(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{j_x(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\, d^3\boldsymbol{r}',$$

$$A_y(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{j_y(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}', \tag{3.146}$$

$$A_z(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{j_z(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'.$$

These solutions can be recombined to form a single vector solution

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.147}$$

Of course, we have seen a equation like this before:

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.148}$$

Equations (3.147) and (3.148) are the unique solutions (given the arbitrary choice of gauge) to the field equations (3.106) and (3.107); they specify the magnetic vector and electric scalar potentials generated by a set of stationary charges, of charge density $\rho(\boldsymbol{r})$, and a set of steady currents, of current density $\boldsymbol{j}(\boldsymbol{r})$. Incidentally, we can prove that Eq. (3.147) satisfies the gauge condition $\nabla\cdot\boldsymbol{A}=0$ by repeating the analysis of Eqs. (3.121)–(3.127) (with $\boldsymbol{W}\rightarrow\boldsymbol{A}$ and $\boldsymbol{C}\rightarrow\mu_0\boldsymbol{j}$) and using the fact that $\nabla\cdot\boldsymbol{j}=0$ for steady currents.

3.12 The Biot-Savart law

According to Eq. (3.133) we can obtain an expression for the electric field generated by stationary charges by taking minus the gradient of Eq. (3.148). This yields

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \rho(\boldsymbol{r}')\,\frac{\boldsymbol{r}-\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,d^3\boldsymbol{r}', \tag{3.149}$$

which we recognize as Coulomb's law written for a continuous charge distribution. According to Eq. (3.135) we can obtain an equivalent expression for the magnetic field generated by steady currents by taking the curl of Eq. (3.147). This gives

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')\wedge(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,d^3\boldsymbol{r}', \tag{3.150}$$

where use has been made of the vector identity $\nabla \wedge (\phi \boldsymbol{A}) = \phi \nabla \wedge \boldsymbol{A} + \nabla \phi \wedge \boldsymbol{A}$. Equation (3.150) is known as the "Biot-Savart law" after the French physicists Jean Baptiste Biot and Felix Savart; it completely specifies the magnetic field generated by a steady (but otherwise quite general) distributed current.

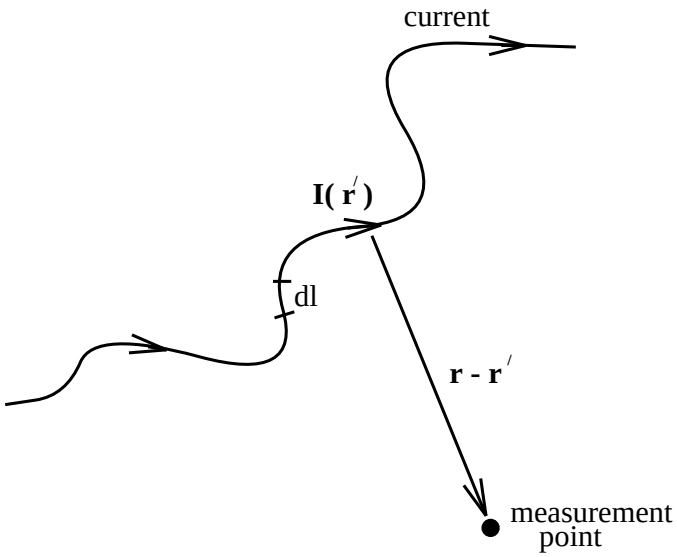
Let us reduce our distributed current to an idealized zero thickness wire. We can do this by writing

$$\boldsymbol{j}(\boldsymbol{r})\, d^3\boldsymbol{r} = \boldsymbol{I}(\boldsymbol{r})\, dl, \tag{3.151}$$

where $\boldsymbol{I}$ is the vector current (*i.e.*, its direction and magnitude specify the direction and magnitude of the current) and dl is an element of length along the wire. Equations (3.150) and (3.151) can be combined to give

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{I}(\boldsymbol{r}') \wedge (\boldsymbol{r} - \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|^3}\, dl, \tag{3.152}$$

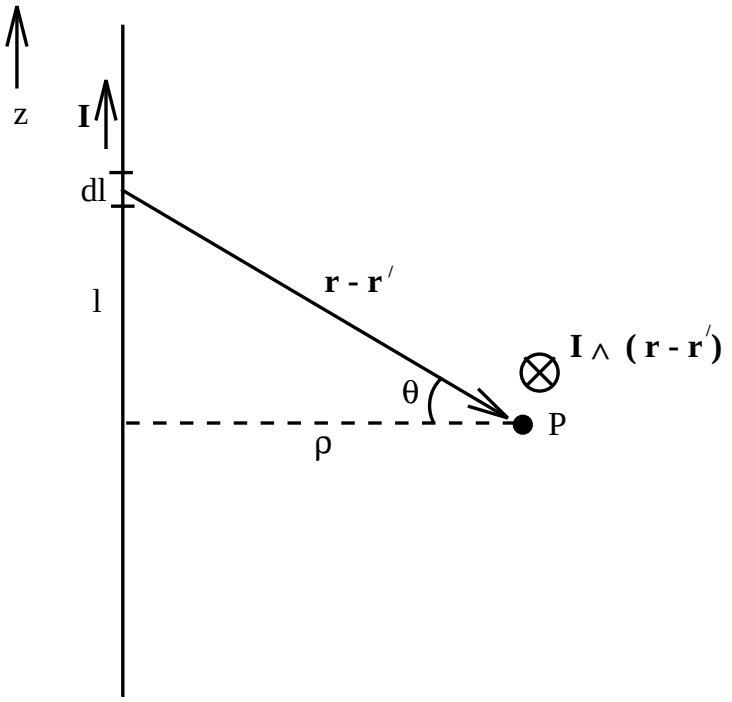
which is the form in which the Biot-Savart law is most usually written. This



law is to magnetostatics (*i.e.*, the study of magnetic fields generated by steady currents) what Coulomb's law is to electrostatics (*i.e.*, the study of electric fields generated by stationary charges). Furthermore, it can be experimentally verified given a set of currents, a compass, a test wire, and a great deal of skill and patience. This justifies our earlier assumption that the field equations (3.105) are valid for general current distributions (recall that we derived them by studying

the fields generated by infinite, straight wires). Note that both Coulomb's law and the Biot-Savart law are "gauge independent"; *i.e.*, they do not depend on the particular choice of gauge.

Consider (for the last time!) an infinite, straight wire directed along the z-axis and carrying a current I. Let us reconstruct the magnetic field generated by



the wire at point P using the Biot-Savart law. Suppose that the perpendicular distance to the wire is ρ. It is easily seen that

$$\begin{aligned} \boldsymbol{I}\wedge(\boldsymbol{r}-\boldsymbol{r}') &= I\rho\,\hat{\boldsymbol{\theta}},\\ l &= \rho\,\tan\theta,\\ dl &= \frac{\rho}{\cos^2\theta}\,d\theta,\\ |\boldsymbol{r}-\boldsymbol{r}'| &= \frac{\rho}{\cos\theta}. \end{aligned} \tag{3.153}$$

Thus, according to Eq. (3.152) we have

$$B_\theta = \frac{\mu_0}{4\pi}\int_{-\pi/2}^{\pi/2}\frac{I\rho}{\rho^3(\cos\theta)^{-3}}\,\frac{\rho}{\cos^2\theta}\,d\theta$$

$$= \frac{\mu_0 I}{4\pi\rho}\int_{-\pi/2}^{\pi/2}\cos\theta\,d\theta = \frac{\mu_0 I}{4\pi\rho}\left[\sin\theta\right]_{-\pi/2}^{\pi/2}, \tag{3.154}$$

which gives the familiar result

$$B_\theta = \frac{\mu_0 I}{2\pi\rho}. \tag{3.155}$$

So, we have come full circle in our investigation of magnetic fields. Note that the simple result (3.155) can only be obtained from the Biot-Savart law after some non-trivial algebra. Examination of more complicated current distributions using this law invariably leads to lengthy, involved, and extremely unpleasant calculations.

3.13 Electrostatics and magnetostatics

We have now completed our theoretical investigation of electrostatics and magnetostatics. Our next task is to incorporate time variation into our analysis. However, before we start this let us briefly review our progress so far. We have found that the electric fields generated by stationary charges and the magnetic fields generated by steady currents are describable in terms of four field equations:

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{3.156a}$$

$$\nabla\wedge\boldsymbol{E} = \boldsymbol{0}, \tag{3.156b}$$

$$\nabla\cdot\boldsymbol{B} = 0, \tag{3.156c}$$

$$\nabla\wedge\boldsymbol{B} = \mu_0\boldsymbol{j}. \tag{3.156d}$$

The boundary conditions are that the fields are zero at infinity, assuming that the generating charges and currents are localized to some region in space. According to Helmholtz's theorem the above field equations, plus the boundary conditions, are sufficient to *uniquely* specify the electric and magnetic fields. The physical significance of this is that divergence and curl are the only *rotationally invariant* differential properties of a general vector field; *i.e.*, the only quantities which do not change when the axes are rotated. Since physics does not depend on the

orientation of the axes (which is, after all, quite arbitrary) divergence and curl are the *only* quantities which can appear in field equations which claim to describe physical phenomena.

The field equations can be integrated to give:

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int_V \rho\, dV, \tag{3.157a}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{3.157b}$$

$$\oint_S \boldsymbol{B}\cdot d\boldsymbol{S} = 0, \tag{3.157c}$$

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \mu_0\int_{S'} \boldsymbol{j}\cdot d\boldsymbol{S}. \tag{3.157d}$$

Here, S is a closed surface enclosing a volume V. Also, C is a closed loop, and S' is some surface attached to this loop. The field equations (3.156) can be deduced from Eqs. (3.157) using Gauss' theorem and Stokes' theorem. Equation (3.157a) is called Gauss' law and says that the flux of the electric field out of a closed surface is proportional to the enclosed electric charge. Equation (3.157c) has no particular name and says that there are no such things as magnetic monopoles. Equation (3.157d) is called Ampère's law and says that the line integral of the magnetic field around any closed loop is proportional to the flux of the current through the loop. Equations (3.157b) and (3.157d) are incomplete; each acquires an extra term on the right-hand side in time dependent situations.

The field equation (3.156b) is automatically satisfied if we write

$$\boldsymbol{E} = -\nabla\phi. \tag{3.158}$$

Likewise, the field equation (3.156c) is automatically satisfied if we write

$$\boldsymbol{B} = \nabla\wedge\boldsymbol{A}. \tag{3.159}$$

Here, ϕ is the electric scalar potential and $\boldsymbol{A}$ is the magnetic vector potential. The electric field is clearly unchanged if we add a constant to the scalar potential:

$$\boldsymbol{E}\rightarrow\boldsymbol{E} \quad \text{as} \quad \phi\rightarrow\phi + c. \tag{3.160}$$

The magnetic field is similarly unchanged if we add the gradient of a scalar field to the vector potential:

$$\boldsymbol{B} \rightarrow \boldsymbol{B} \quad \text{as} \quad \boldsymbol{A} \rightarrow \boldsymbol{A} + \nabla\psi. \tag{3.161}$$

The above transformations, which leave the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields invariant, are called gauge transformations. We are free to choose c and ψ to be whatever we like; *i.e.*, we are free to choose the gauge. The most sensible gauge is the one which make our equations as simple and symmetric as possible. This corresponds to the choice

$$\phi(\boldsymbol{r}) \rightarrow 0 \quad \text{as} \quad |\boldsymbol{r}| \rightarrow \infty, \tag{3.162}$$

and

$$\nabla \cdot \boldsymbol{A} = 0. \tag{3.163}$$

The latter convention is known as the Coulomb gauge.

Taking the divergence of Eq. (3.158) and the curl of Eq. (3.159), and making use of the Coulomb gauge, we find that the four field equations (3.156) can be reduced to Poisson's equation written four times over:

$$\nabla^2\phi = -\frac{\rho}{\epsilon_0}, \tag{3.164a}$$

$$\nabla^2\boldsymbol{A} = -\mu_0\boldsymbol{j}. \tag{3.164b}$$

Poisson's equation is just about the simplest *rotationally invariant* partial differential equation it is possible to write. Note that ∇^2 is clearly rotationally invariant since it is the divergence of a gradient, and both divergence and gradient are rotationally invariant. We can always construct the solution to Poisson's equation, given the boundary conditions. Furthermore, we have a *uniqueness theorem* which tells us that our solution is the only possible solution. Physically, this means that there is only one electric and magnetic field which is consistent with a given set of stationary charges and steady currents. This sounds like an obvious, almost trivial, statement. But there are many areas of physics (for instance, fluid mechanics and plasma physics) where we also believe, for physical reasons, that for a given set of boundary conditions the solution should be unique. The problem is that in most cases when we reduce the problem to a partial differential

equation we end up with something far nastier than Poisson's equation. In general, we cannot solve this equation. In fact, we usually cannot even prove that it possess a solution for general boundary conditions, let alone that the solution is unique. So, we are very fortunate indeed that in electrostatics and magnetostatics the problem boils down to solving a nice partial differential equation. When you hear people say things like "electromagnetism is the best understood theory in physics" what they are really saying is that the partial differential equations which crop up in this theory are soluble and have nice properties.

Poisson's equation

$$\nabla^2 u = v \tag{3.165}$$

is *linear*, which means that its solutions are superposable. We can exploit this fact to construct a general solution to this equation. Suppose that we can find the solution to

$$\nabla^2 G(\boldsymbol{r},\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}') \tag{3.166}$$

which satisfies the boundary conditions. This is the solution driven by a unit amplitude point source located at position vector $\boldsymbol{r}'$. Since any general source can be built up out of a weighted sum of point sources it follows that a general solution to Poisson's equation can be built up out of a weighted superposition of point source solutions. Mathematically, we can write

$$u(\boldsymbol{r}) = \int G(\boldsymbol{r},\boldsymbol{r}')\, v(\boldsymbol{r}')\, d^3\boldsymbol{r}'. \tag{3.167}$$

The function G is called the Green's function. The Green's function for Poisson's equation is

$$G(\boldsymbol{r},\boldsymbol{r}') = -\frac{1}{4\pi}\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.168}$$

Note that this Green's function is proportional to the scalar potential of a point charge located at $\boldsymbol{r}'$; this is hardly surprising given the definition of a Green's function and Eq. (3.164a).

According to Eqs. (3.164), (3.165), (3.167), and (3.168), the scalar and vector potentials generated by a set of stationary charges and steady currents take the

form

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}', \tag{3.169a}$$

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{3.169b}$$

Making use of Eqs. (3.158) and (3.159) we obtain the fundamental force laws for electric and magnetic fields. Coulomb's law

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \rho(\boldsymbol{r}')\,\frac{\boldsymbol{r}-\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,d^3\boldsymbol{r}', \tag{3.170}$$

and the Biot-Savart law

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')\wedge(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,d^3\boldsymbol{r}'. \tag{3.171}$$

Of course, both of these laws are examples of action at a distance laws and, therefore, violate relativity. However, this is not a problem as long as we restrict ourselves to fields generated by time independent charge and current distributions.

The question, now, is how badly is this scheme we have just worked out going to be disrupted when we take time variation into account. The answer, somewhat surprisingly, is by very little indeed. So, in Eqs. (3.156)–(3.171) we can already discern the basic outline of classical electromagnetism. Let us continue our investigation.

3.14 Faraday's law

The history of mankind's development of physics is really the history of the synthesis of ideas. Physicists keep finding that apparently disparate phenomena can be understood as different aspects of some more fundamental phenomenon. This process has continued until today all physical phenomena can be described in terms of three fundamental forces: gravity, the electroweak force, and the

strong force. One of the main goals of modern physics is to find some way of combining these three forces so that all of physics can be described in terms of a single unified force. This, essentially, is the purpose of super-symmetry theories.

The first great synthesis of ideas in physics took place in 1666 when Issac Newton realised that the force which causes apples to fall downwards is the same as the force which maintains the planets in elliptical orbits around the Sun. The second great synthesis, which we are about to study in more detail, took place in 1830 when Michael Faraday discovered that electricity and magnetism are two aspects of the same thing, usually referred to as "electromagnetism." The third great synthesis, which we shall discuss presently, took place in 1873 when James Clerk Maxwell demonstrated that light and electromagnetism are intimately related. The last (but, hopefully, not the final) great synthesis took place in 1967 when Steve Weinberg and Abdus Salam showed that the electromagnetic force and the weak nuclear force (*i.e.*, the one which is responsible for β decays) can be combined to give the electroweak force. Unfortunately, Weinberg's work lies beyond the scope of this lecture course.

Let us now consider Faraday's experiments, having put them in their proper historical context. Prior to 1830 the only known way to make an electric current flow through a conducting wire was to connect the ends of the wire to the positive and negative terminals of a battery. We measure a battery's ability to push current down a wire in terms of its "voltage," by which we mean the voltage difference between its positive and negative terminals. What does voltage correspond to in physics? Well, volts are the units used to measure electric scalar potential, so when we talk about a 6V battery what we are really saying is that the difference in electric scalar potential between its positive and negative terminals is six volts. This insight allows us to write

$$V = \phi(\oplus) - \phi(\ominus) = -\int_{\oplus}^{\ominus} \nabla\phi \cdot d\boldsymbol{l} = \int_{\oplus}^{\ominus} \boldsymbol{E} \cdot d\boldsymbol{l}, \tag{3.172}$$

where V is the battery voltage, $\oplus$ denotes the positive terminal, $\ominus$ the negative terminal, and $d\boldsymbol{l}$ is an element of length along the wire. Of course, the above equation is a direct consequence of $\boldsymbol{E} = -\nabla\phi$. Clearly, a voltage difference between two ends of a wire attached to a battery implies the presence of an electric field which pushes charges through the wire. This field is directed from

the positive terminal of the battery to the negative terminal and is, therefore, such as to force electrons to flow through the wire from the negative to the positive terminal. As expected, this means that a net positive current flows from the positive to the negative terminal. The fact that $\boldsymbol{E}$ is a conservative field ensures that the voltage difference V is independent of the path of the wire. In other words, two different wires attached to the same battery develop identical voltage differences. This is just as well. The quantity V is usually called the *electromotive force*, or e.m.f. for short. "Electromotive force" is a bit of a misnomer. The e.m.f. is certainly what causes current to flow through a wire, so it is electromotive (*i.e.*, it causes electrons to move), but it is not a force. In fact, it is a difference in electric scalar potential.

Let us now consider a closed loop of wire (with no battery). The electromotive force around such a loop is

$$V = \oint \boldsymbol{E} \cdot d\boldsymbol{l} = 0. \tag{3.173}$$

This is a direct consequence of the field equation $\nabla \wedge \boldsymbol{E} = \boldsymbol{0}$. So, since $\boldsymbol{E}$ is a conservative field then the electromotive force around a closed loop of wire is automatically zero and no current flows around the wire. This all seems to make sense. However, Michael Faraday is about to throw a spanner in our works! He discovered in 1830 that a changing magnetic field can cause a current to flow around a closed loop of wire (in the absence of a battery). Well, if current flows through a wire then there must be an electromotive force. So,

$$V = \oint \boldsymbol{E} \cdot d\boldsymbol{l} \neq 0, \tag{3.174}$$

which immediately implies that $\boldsymbol{E}$ is not a conservative field, and that $\nabla \wedge \boldsymbol{E} \neq \boldsymbol{0}$. Clearly, we are going to have to modify some of our ideas regarding electric fields!

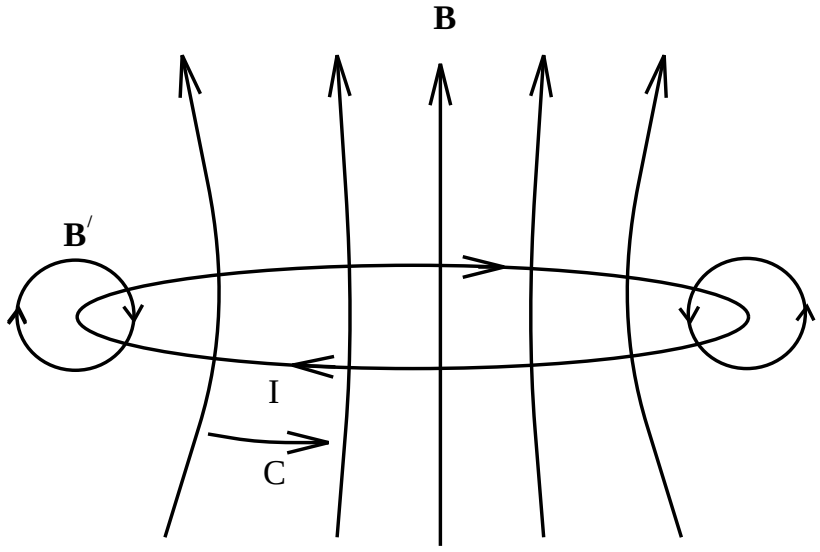
Faraday continued his experiments and found that another way of generating an electromotive force around a loop of wire is to keep the magnetic field constant and move the loop. Eventually, Faraday was able to formulate a law which accounted for all of his experiments. The e.m.f. generated around a loop of wire in a magnetic field is proportional to the rate of change of the flux of the magnetic field through the loop. So, if the loop is denoted C and S is some surface attached

to the loop then Faraday's experiments can be summed up by writing

$$V = \oint_C \boldsymbol{E} \cdot d\boldsymbol{l} = A \frac{\partial}{\partial t} \int_S \boldsymbol{B} \cdot d\boldsymbol{S}, \tag{3.175}$$

where A is a constant of proportionality. Thus, the changing flux of the magnetic field through the loop creates an electric field directed around the loop. This process is know as "magnetic induction."

S.I. units have been carefully chosen so as to make $|A| = 1$ in the above equation. The only thing we now have to decide is whether $A = +1$ or $A = -1$. In other words, which way around the loop does the induced e.m.f. want to drive the current? We possess a general principle which allows us to decide questions like this. It is called Le Chatelier's principle. According to Le Chatelier's principle every change generates a reaction which tries to minimize the change. Essentially, this means that the universe is stable to small perturbations. When this principle is applied to the special case of magnetic induction it is usually called Lenz's law. According to Lenz's law, the current induced around a closed loop is always such that the magnetic field it produces tries to counteract the change in magnetic flux which generates the electromotive force. From the diagram, it is clear that if the



magnetic field $\boldsymbol{B}$ is increasing and the current I circulates clockwise (as seen from above) then it generates a field $\boldsymbol{B'}$ which opposes the increase in magnetic flux through the loop, in accordance with Lenz's law. The direction of the current is

opposite to the sense of the current loop C (assuming that the flux of $\boldsymbol{B}$ through the loop is positive), so this implies that $A = -1$ in Eq. (3.175). Thus, Faraday's law takes the form

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\frac{\partial}{\partial t}\int_S \boldsymbol{B}\cdot d\boldsymbol{S}. \tag{3.176}$$

Experimentally, Faraday's law is found to correctly predict the e.m.f. (*i.e.*, $\oint \boldsymbol{E}\cdot d\boldsymbol{l}$) generated in any wire loop, irrespective of the position or shape of the loop. It is reasonable to assume that the same e.m.f. would be generated in the absence of the wire (of course, no current would flow in this case). Thus, Eq. (3.176) is valid for any closed loop C. If Faraday's law is to make any sense it must also be true for any surface S attached to the loop C. Clearly, if the flux of the magnetic field through the loop depends on the surface upon which it is evaluated then Faraday's law is going to predict different e.m.f.s for different surfaces. Since there is no preferred surface for a general non-coplanar loop, this would not make very much sense. The condition for the flux of the magnetic field, $\int_S \boldsymbol{B}\cdot d\boldsymbol{S}$, to depend only on the loop C to which the surface S is attached, and not on the nature of the surface itself, is

$$\oint_{S'} \boldsymbol{B}\cdot d\boldsymbol{S}' = 0, \tag{3.177}$$

for any closed surface S'.

Faraday's law, Eq. (3.176), can be converted into a field equation using Stokes' theorem. We obtain

$$\nabla\wedge\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t}. \tag{3.178}$$

This is the final Maxwell equation. It describes how a changing magnetic field can generate, or induce, an electric field. Gauss' theorem applied to Eq. (3.177) yields the familiar field equation

$$\nabla\cdot\boldsymbol{B} = 0. \tag{3.179}$$

This ensures that the magnetic flux through a loop is a well defined quantity.

The divergence of Eq. (3.178) yields

$$\frac{\partial\,\nabla\cdot\boldsymbol{B}}{\partial t} = 0. \tag{3.180}$$

Thus, the field equation (3.178) actually demands that the divergence of the magnetic field be constant in time for self-consistency (this means that the flux of the magnetic field through a loop need not be a well defined quantity as long as its time derivative is well defined). However, a constant non-solenoidal magnetic field can only be generated by magnetic monopoles, and magnetic monopoles do not exist (as far as we are aware). Hence, $\nabla \cdot \boldsymbol{B} = 0$. The absence of magnetic monopoles is an observational fact, it cannot be predicted by any theory. If magnetic monopoles were discovered tomorrow this would not cause physicists any problems. We know how to generalize Maxwell's equations to include both magnetic monopoles and currents of magnetic monopoles. In this generalized formalism Maxwell's equations are completely symmetric with respect to electric and magnetic fields, and $\nabla \cdot \boldsymbol{B} \neq 0$. However, an extra term (involving the current of magnetic monopoles) must be added to the right-hand side of Eq. (3.178) in order to make it self-consistent.

3.15 Electric scalar potential?

We now have a problem. We can only write the electric field in terms of a scalar potential (*i.e.*, $\boldsymbol{E} = -\nabla\phi$) provided that $\nabla \wedge \boldsymbol{E} = \boldsymbol{0}$. However, we have just found that in the presence of a changing magnetic field the curl of the electric field is non-zero. In other words, $\boldsymbol{E}$ is not, in general, a conservative field. Does this mean that we have to abandon the concept of electric scalar potential? Fortunately, no. It is still possible to define a scalar potential which is physically meaningful.

Let us start from the equation

$$\nabla \cdot \boldsymbol{B} = 0, \tag{3.181}$$

which is valid for both time varying and non time varying magnetic fields. Since the magnetic field is solenoidal we can write it as the curl of a vector potential:

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A}. \tag{3.182}$$

So, there is no problem with the vector potential in the presence of time varying fields. Let us substitute Eq. (3.182) into the field equation (3.178). We obtain

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial\, \nabla \wedge \boldsymbol{A}}{\partial t}, \tag{3.183}$$

which can be written

$$\nabla \wedge \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} \right) = 0. \tag{3.184}$$

We know that a curl free vector field can always be expressed as the gradient of a scalar potential, so let us write

$$\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} = -\nabla \phi, \tag{3.185}$$

or

$$\boldsymbol{E} = -\nabla \phi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{3.186}$$

This is a very nice equation! It tells us that the scalar potential ϕ only describes the conservative electric field generated by electric charges. The electric field induced by time varying magnetic fields is non-conservative, and is described by the magnetic vector potential.

3.16 Gauge transformations

Electric and magnetic fields can be written in terms of scalar and vector potentials, as follows:

$$\begin{aligned} \boldsymbol{E} &= -\nabla \phi - \frac{\partial \boldsymbol{A}}{\partial t}, \\ \boldsymbol{B} &= \nabla \wedge \boldsymbol{A}. \end{aligned} \tag{3.187}$$

However, this prescription is not unique. There are many different potentials which generate the same fields. We have come across this problem before. It is called gauge invariance. The most general transformation which leaves the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields unchanged in Eqs. (3.187) is

$$\begin{aligned} \phi &\rightarrow \phi + \frac{\partial \psi}{\partial t}, \\ \boldsymbol{A} &\rightarrow \boldsymbol{A} - \nabla \psi. \end{aligned} \tag{3.188}$$

This is clearly a generalization of the gauge transformation which we found earlier for static fields:

$$\begin{aligned} \phi &\rightarrow \phi + c, \\ \boldsymbol{A} &\rightarrow \boldsymbol{A} - \nabla\psi, \end{aligned} \tag{3.189}$$

where c is a constant. In fact, if $\psi(\boldsymbol{r},t) \rightarrow \psi(\boldsymbol{r}) + c\,t$ then Eqs. (3.188) reduce to Eqs. (3.189).

We are free to choose the gauge so as to make our equations as simple as possible. As before, the most sensible gauge for the scalar potential is to make it go to zero at infinity:

$$\phi(\boldsymbol{r}) \rightarrow 0 \qquad \text{as} \quad |\boldsymbol{r}| \rightarrow \infty. \tag{3.190}$$

For steady fields we found that the optimum gauge for the vector potential was the so called Coulomb gauge:

$$\nabla\cdot\boldsymbol{A} = 0. \tag{3.191}$$

We can still use this gauge for non-steady fields. The argument which we gave earlier (see Section 3.11), that it is always possible to transform away the divergence of a vector potential, remains valid. One of the nice features of the Coulomb gauge is that when we write the electric field,

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial\boldsymbol{A}}{\partial t}, \tag{3.192}$$

we find that the part which is generated by charges (*i.e.*, the first term on the right-hand side) is conservative and the part induced by magnetic fields (*i.e.*, the second term on the right-hand side) is purely solenoidal. Earlier on, we proved mathematically that a general vector field can be written as the sum of a conservative field and a solenoidal field (see Section 3.10). Now we are finding that when we split up the electric field in this manner the two fields have different physical origins: the conservative part of the field emanates from electric charges whereas the solenoidal part is induced by magnetic fields.

Equation (3.192) can be combined with the field equation

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{3.193}$$

(which remains valid for non-steady fields) to give

$$-\nabla^2\phi - \frac{\partial\, \nabla\cdot \boldsymbol{A}}{\partial t} = \frac{\rho}{\epsilon_0}. \tag{3.194}$$

With the Coulomb gauge condition, $\nabla\cdot\boldsymbol{A} = 0$, the above expression reduces to

$$\nabla^2\phi = -\frac{\rho}{\epsilon_0}, \tag{3.195}$$

which is just Poisson's equation. Thus, we can immediately write down an expression for the scalar potential generated by non-steady fields. It is exactly the same as our previous expression for the scalar potential generated by steady fields, namely

$$\phi(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}\, d^3\boldsymbol{r}'. \tag{3.196}$$

However, this apparently simple result is *extremely* deceptive. Equation (3.196) is a typical action at a distance law. If the charge density changes suddenly at $\boldsymbol{r}'$ then the potential at $\boldsymbol{r}$ responds *immediately*. However, we shall see later that the full time dependent Maxwell's equations only allow information to propagate at the speed of light (*i.e.*, they do not violate relativity). How can these two statements be reconciled? The crucial point is that the scalar potential cannot be measured directly, it can only be inferred from the electric field. In the time dependent case there are two parts to the electric field; that part which comes from the scalar potential, and that part which comes from the vector potential [see Eq. (3.192)]. So, if the scalar potential responds immediately to some distance rearrangement of charge density it does not necessarily follow that the electric field also has an immediate response. What actually happens is that the change in the part of the electric field which comes from the scalar potential is balanced by an equal and opposite change in the part which comes from the vector potential, so that the overall electric field remains unchanged. This state of affairs persists at least until sufficient time has elapsed for a light ray to travel from the distant charges to the region in question. Thus, relativity is not violated since it is the electric field, and not the scalar potential, which carries physically accessible information.

It is clear that the apparent action at a distance nature of Eq. (3.196) is highly misleading. This suggests, very strongly, that the Coulomb gauge is not

the optimum gauge in the time dependent case. A more sensible choice is the so called "Lorentz gauge":

$$\nabla \cdot \boldsymbol{A} = -\epsilon_0 \mu_0 \, \frac{\partial \phi}{\partial t}. \tag{3.197}$$

It can be shown, by analogy with earlier arguments (see Section 3.11), that it is always possible to make a gauge transformation, at a given instance in time, such that the above equation is satisfied. Substituting the Lorentz gauge condition into Eq. (3.194), we obtain

$$\epsilon_0 \mu_0 \, \frac{\partial^2 \phi}{\partial t^2} - \nabla^2 \phi = \frac{\rho}{\epsilon_0}. \tag{3.198}$$

It turns out that this is a three dimensional wave equation in which information propagates at the speed of light. But, more of this later. Note that the magnetically induced part of the electric field (*i.e.*, $-\partial \boldsymbol{A}/\partial t$) is not purely solenoidal in the Lorentz gauge. This is a slight disadvantage of the Lorentz gauge with respect to the Coulomb gauge. However, this disadvantage is more than offset by other advantages which will become apparent presently. Incidentally, the fact that the part of the electric field which we ascribe to magnetic induction changes when we change the gauge suggests that the separation of the field into magnetically induced and charge induced components is not unique in the general time varying case (*i.e.*, it is a convention).

3.17 The displacement current

Michael Faraday revolutionized physics in 1830 by showing that electricity and magnetism were interrelated phenomena. He achieved this breakthrough by careful experimentation. Between 1864 and 1873 James Clerk Maxwell achieved a similar breakthrough by pure thought. Of course, this was only possible because he was able to take the experimental results of Faraday, Ampère, *etc.*, as his starting point. Prior to 1864 the laws of electromagnetism were written in integral form. Thus, Gauss's law was (in S.I. units) *the flux of the electric field through a closed surface equals the total enclosed charge divided by* ϵ_0. The no magnetic monopole law was *the flux of the magnetic field through any closed surface is zero.* Faraday's law was *the electromotive force generated around a closed loop equals*

minus the rate of change of the magnetic flux through the loop. Finally, Ampère's law was *the line integral of the magnetic field around a closed loop equals the total current flowing through the loop times* μ_0. Maxwell's first great achievement was to realize that these laws could be expressed as a set of partial differential equations. Of course, he wrote his equations out in component form because modern vector notation did not come into vogue until about the time of the First World War. In modern notation, Maxwell first wrote

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{3.199a}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{3.199b}$$

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{3.199c}$$

$$\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j}. \tag{3.199d}$$

Maxwell's second great achievement was to realize that these equations are wrong.

We can see that there is something slightly unusual about Eqs. (3.199). They are very unfair to electric fields! After all, time varying magnetic fields can induce electric fields, but electric fields apparently cannot affect magnetic fields in any way. However, there is a far more serious problem associated with the above equations, which we alluded to earlier on. Consider the integral form of the last Maxwell equation (*i.e.*, Ampère's law)

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{l} = \mu_0 \int_S \boldsymbol{j} \cdot d\boldsymbol{S}. \tag{3.200}$$

This says that the line integral of the magnetic field around a closed loop C is equal to μ_0 times the flux of the current density through the loop. The problem is that the flux of the current density through a loop is not, in general, a well defined quantity. In order for the flux to be well defined the integral of $\boldsymbol{j} \cdot d\boldsymbol{S}$ over some surface S attached to a loop C must depend on C but not on the details of S. This is only the case if

$$\nabla \cdot \boldsymbol{j} = 0. \tag{3.201}$$

Unfortunately, the above condition is only satisfied for non time varying fields.

Why do we say that, in general, $\nabla \cdot \boldsymbol{j} \neq 0$? Well, consider the flux of $\boldsymbol{j}$ over some closed surface S enclosing a volume V. This is clearly equivalent to the rate at which charge flows through S. However, if charge is a conserved quantity (and we certainly believe that it is) then the rate at which charge flows through S must equal the rate of decrease of the charge contained in volume V. Thus,

$$\oint_S \boldsymbol{j} \cdot d\boldsymbol{S} = -\frac{\partial}{\partial t}\int_V \rho \, dV. \tag{3.202}$$

Making use of Gauss' theorem, this yields

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t}. \tag{3.203}$$

Thus, $\nabla \cdot \boldsymbol{j} = 0$ is only true in a steady state (*i.e.*, when $\partial/\partial t \equiv 0$).

The problem with Ampère's law is well illustrated by the following very famous example. Consider a long straight wire interrupted by a parallel plate capacitor. Suppose that C is some loop which circles the wire. In the non time dependent situation the capacitor acts like a break in the wire, so no current flows, and no magnetic field is generated. There is clearly no problem with Ampère's law in this case. In the time dependent situation a transient current flows in the wire as the capacitor charges up, or charges down, so a transient magnetic field is generated. Thus, the line integral of the magnetic field around C is (transiently) non-zero. According to Ampère's law, the flux of the current through any surface attached to C should also be (transiently) non-zero. Let us consider two such surfaces. The first surface, S_1, intersects the wire. This surface causes us no problem since the flux of $\boldsymbol{j}$ though the surface is clearly non-zero (because it intersects a current carrying wire). The second surface, S_2, passes between the plates of the capacitor and, therefore, does not intersect the wire at all. Clearly, the flux of the current through this surface is zero. The current fluxes through surfaces S_1 and S_2 are obviously different. However, both surfaces are attached to the same loop C, so the fluxes should be the same according to Ampère's law. It would appear that Ampère's law is about to disintegrate! However, we notice that although the surface S_2 does not intersect any electric current it does pass through a region of strong changing electric field as it threads between the plates of the charging (or discharging) capacitor. Perhaps, if we add a term involving $\partial \boldsymbol{E}/\partial t$ to the

right-hand side of Eq. (3.199d) we can somehow fix up Ampère's law? This is, essentially, how Maxwell reasoned more than one hundred years ago.

Let us try out this scheme. Suppose that we write

$$\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j} + \lambda \frac{\partial \boldsymbol{E}}{\partial t} \tag{3.204}$$

instead of Eq. (3.199d). Here, λ is some constant. Does this resolve our problem? We want the flux of the right-hand side of the above equation through some loop C to be well defined; *i.e.*, it should only depend on C and not the particular surface S (which spans C) upon which it is evaluated. This is another way of saying that we want the divergence of the right-hand side to be zero. In fact, we can see that this is necessary for self consistency since the divergence of the left-hand side is automatically zero. So, taking the divergence of Eq. (3.204) we obtain

$$0 = \mu_0 \nabla \cdot \boldsymbol{j} + \lambda \, \frac{\partial \, \nabla \cdot \boldsymbol{E}}{\partial t}. \tag{3.205}$$

But, we know that

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{3.206}$$

so combining the previous two equations we arrive at

$$\mu_0 \nabla \cdot \boldsymbol{j} + \frac{\lambda}{\epsilon_0} \frac{\partial \rho}{\partial t} = 0. \tag{3.207}$$

Now, our charge conservation law (3.203) can be written

$$\nabla \cdot \boldsymbol{j} + \frac{\partial \rho}{\partial t} = 0. \tag{3.208}$$

The previous two equations are in agreement provided $\lambda = \epsilon_0 \mu_0$. So, if we modify the final Maxwell equation such that it reads

$$\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j} + \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t} \tag{3.209}$$

then we find that the divergence of the right-hand side is zero as a consequence of charge conservation. The extra term is called the "displacement current" (this

name was invented by Maxwell). In summary, we have shown that although the flux of the real current through a loop is *not* well defined, if we form the sum of the real current and the displacement current then the flux of this new quantity through a loop *is* well defined.

Of course, the displacement current is not a current at all. It is, in fact, associated with the generation of magnetic fields by time varying electric fields. Maxwell came up with this rather curious name because many of his ideas regarding electric and magnetic fields were completely wrong. For instance, Maxwell believed in the æther, and he thought that electric and magnetic fields were some sort of stresses in this medium. He also thought that the displacement current was associated with displacements of the æther (hence, the name). The reason that these misconceptions did not invalidate his equations is quite simple. Maxwell based his equations on the results of experiments, and he added in his extra term so as to make these equations mathematically self consistent. Both of these steps are valid irrespective of the existence or non-existence of the æther.

"But, hang on a minute," you might say, "you can't go around adding terms to laws of physics just because you feel like it! The field equations (3.199) are derived directly from the results of famous nineteenth century experiments. If there is a new term involving the time derivative of the electric field which needs to be added into these equations, how come there is no corresponding nineteenth century experiment which demonstrates this? We have Ampère's law which shows that changing magnetic fields generate electric fields. Why is there no "Joe Blogg's" law that says that changing electric fields generate magnetic fields?" This is a perfectly reasonable question. The answer is that the new term describes an effect which is far too small to have been observed in nineteenth century experiments. Let us demonstrate this.

First, we shall show that it is comparatively easy to detect the induction of an electric field by a changing magnetic field in a desktop laboratory experiment. The Earth's magnetic field is about 1 gauss (that is, 10^{-4} tesla). Magnetic fields generated by electromagnets (which will fit on a laboratory desktop) are typically about one hundred times bigger that this. Let us, therefore, consider a hypothetical experiment in which a 100 gauss magnetic field is switched on suddenly. Suppose that the field ramps up in one tenth of a second. What

electromotive force is generated in a 10 centimeter square loop of wire located in this field? Ampère's law is written

$$V = -\frac{\partial}{\partial t}\oint \boldsymbol{B}\cdot d\boldsymbol{S} \sim \frac{B\,A}{t}, \tag{3.210}$$

where $B = 0.01$ tesla is the field strength, $A = 0.01$ m^2 is the area of the loop, and $t = 0.1$ seconds is the ramp time. It follows that $V \sim 1$ millivolt. Well, one millivolt is easily detectable. In fact, most hand-held laboratory voltmeters are calibrated in millivolts. It is clear that we would have no difficulty whatsoever detecting the magnetic induction of electric fields in a nineteenth century style laboratory experiment.

Let us now consider the electric induction of magnetic fields. Suppose that our electric field is generated by a parallel plate capacitor of spacing one centimeter which is charged up to 100 volts. This gives a field of 10^4 volts per meter. Suppose, further, that the capacitor is discharged in one tenth of a second. The law of electric induction is obtained by integrating Eq. (3.209) and neglecting the first term on the right-hand side. Thus,

$$\oint \boldsymbol{B}\cdot d\boldsymbol{l} = \epsilon_0\mu_0\frac{\partial}{\partial t}\int \boldsymbol{E}\cdot d\boldsymbol{S}. \tag{3.211}$$

Let us consider a loop 10 centimeters square. What is the magnetic field generated around this loop (we could try to measure this with a Hall probe). Very approximately we find that

$$lB \sim \epsilon_0\mu_0\,\frac{El^2}{t}, \tag{3.212}$$

where $l = 0.1$ meters is the dimensions of the loop, B is the magnetic field strength, $E = 10^4$ volts per meter is the electric field, and $t = 0.1$ seconds is the decay time of the field. We find that $B \sim 10^{-9}$ gauss. Modern technology is unable to detect such a small magnetic field, so we cannot really blame Faraday for not noticing electric induction in 1830.

"So," you might say, "why did you bother mentioning this displacement current thing in the first place if it is undetectable?" Again, a perfectly fair question. The answer is that the displacement current *is* detectable in some experiments.

Suppose that we take an FM radio signal, amplify it so that its peak voltage is one hundred volts, and then apply it to the parallel plate capacitor in the previous hypothetical experiment. What size of magnetic field would this generate? Well, a typical FM signal oscillates at 10^9 Hz, so t in the previous example changes from 0.1 seconds to 10^{-9} seconds. Thus, the induced magnetic field is about 10^{-1} gauss. This is certainly detectable by modern technology. So, it would seem that if the electric field is oscillating fast then electric induction of magnetic fields is an observable effect. In fact, there is a virtually infallible rule for deciding whether or not the displacement current can be neglected in Eq. (3.209). If *electromagnetic radiation* is important then the displacement current must be included. On the other hand, if electromagnetic radiation is unimportant then the displacement current can be safely neglected. Clearly, Maxwell's inclusion of the displacement current in Eq. (3.209) was a vital step in his later realization that his equations allowed propagating wave-like solutions. These solutions are, of course, electromagnetic waves. But, more of this later.

We are now in a position to write out Maxwell's equations in all their glory! We get

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{3.213a}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{3.213b}$$

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{3.213c}$$

$$\nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j} + \epsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t}. \tag{3.213d}$$

These four partial differential equations constitute a *complete* description of the behaviour of electric and magnetic fields. The first equation describes how electric fields are induced by charges. The second equation says that there is no such thing as a magnetic charge. The third equation describes the induction of electric fields by changing magnetic fields, and the fourth equation describes the generation of magnetic fields by electric currents and the induction of magnetic fields by changing electric fields. Note that with the inclusion of the displacement current these equations treat electric and magnetic fields on an equal footing; *i.e.*, electric fields can induce magnetic fields, and *vice versa*. Equations (3.213) sum up the

experimental results of Coulomb, Ampère, and Faraday very succinctly; they are called Maxwell's equations because James Clerk Maxwell was the first to write them down (in component form). Maxwell also fixed them up so that they made mathematical sense.

3.18 The potential formulation of Maxwell's equations

We have seen that Eqs. (3.213b) and (3.213c) are automatically satisfied if we write the electric and magnetic fields in terms of potentials:

$$\begin{aligned} \boldsymbol{E} &= -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \\ \boldsymbol{B} &= \nabla \wedge \boldsymbol{A}. \end{aligned} \tag{3.214}$$

This prescription is not unique, but we can make it unique by adopting the following conventions:

$$\phi(\boldsymbol{r}) \rightarrow 0 \qquad \text{as} \quad |\boldsymbol{r}| \rightarrow \infty, \tag{3.215a}$$

$$\nabla \cdot \boldsymbol{A} = -\epsilon_0 \mu_0 \frac{\partial \phi}{\partial t}. \tag{3.215b}$$

The above equations can be combined with Eq. (3.213a) to give

$$\epsilon_0 \mu_0 \frac{\partial^2 \phi}{\partial t^2} - \nabla^2 \phi = \frac{\rho}{\epsilon_0}. \tag{3.216}$$

Let us now consider Eq. (3.213d). Substitution of Eqs. (3.214) into this formula yields

$$\nabla \wedge \nabla \wedge \boldsymbol{A} \equiv \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu_0 \boldsymbol{j} - \epsilon_0 \mu_0 \frac{\partial \nabla \phi}{\partial t} - \epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{A}}{\partial t^2}, \tag{3.217}$$

or

$$\epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \nabla^2 \boldsymbol{A} = \mu_0 \boldsymbol{j} - \nabla \left(\nabla \cdot \boldsymbol{A} + \epsilon_0 \mu_0 \frac{\partial \phi}{\partial t} \right). \tag{3.218}$$

We can now see quite clearly where the Lorentz gauge condition (3.215b) comes from. The above equation is, in general, very complicated since it involves both the vector and scalar potentials. But, if we adopt the Lorentz gauge then the last term on the right-hand side becomes zero and the equation simplifies considerably so that it only involves the vector potential. Thus, we find that Maxwell's equations reduce to the following:

$$\begin{aligned}\epsilon_0\mu_0\,\frac{\partial^2\phi}{\partial t^2}-\nabla^2\phi &= \frac{\rho}{\epsilon_0},\\ \epsilon_0\mu_0\,\frac{\partial^2\boldsymbol{A}}{\partial t^2}-\nabla^2\boldsymbol{A} &= \mu_0\boldsymbol{j}.\end{aligned}\tag{3.219}$$

This is the same equation written four times over. In steady state (*i.e.*, $\partial/\partial t = 0$) it reduces to Poisson's equation, which we know how to solve. With the $\partial/\partial t$ terms included it becomes a slightly more complicated equation (in fact, a driven three dimensional wave equation).

3.19 Electromagnetic waves

This is an appropriate point at which to demonstrate that Maxwell's equations possess propagating wave-like solutions. Let us start from Maxwell's equations in free space (*i.e.*, with no charges and no currents):

$$\nabla\cdot\boldsymbol{E} = 0,\tag{3.220a}$$

$$\nabla\cdot\boldsymbol{B} = 0,\tag{3.220b}$$

$$\nabla\wedge\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t},\tag{3.220c}$$

$$\nabla\wedge\boldsymbol{B} = \epsilon_0\mu_0\,\frac{\partial\boldsymbol{E}}{\partial t}.\tag{3.220d}$$

Note that these equations exhibit a nice symmetry between the electric and magnetic fields.

There is an easy way to show that the above equations possess wave-like solutions, and a hard way. The easy way is to assume that the solutions are going

to be wave-like beforehand. Specifically, let us search for plane wave solutions of the form:

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= \boldsymbol{E}_0 \cos(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t), \\ \boldsymbol{B}(\boldsymbol{r},t) &= \boldsymbol{B}_0 \cos(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t+\phi). \end{aligned} \tag{3.221}$$

Here, $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ are constant vectors, $\boldsymbol{k}$ is called the wave-vector, and ω is the angular frequency. The frequency in hertz is related to the angular frequency via $\omega = 2\pi f$. The frequency is conventionally defined to be positive. The quantity ϕ is a phase difference between the electric and magnetic fields. It is more convenient to write

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_0\, \mathrm{e}^{\,\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \\ \boldsymbol{B} &= \boldsymbol{B}_0\, \mathrm{e}^{\,\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \end{aligned} \tag{3.222}$$

where by convention the physical solution is the *real part* of the above equations. The phase difference ϕ is absorbed into the constant vector $\boldsymbol{B}_0$ by allowing it to become complex. Thus, $\boldsymbol{B}_0 \rightarrow \boldsymbol{B}_0\, \mathrm{e}^{\,\mathrm{i}\,\phi}$. In general, the vector $\boldsymbol{E}_0$ is also complex.

A wave maximum of the electric field satisfies

$$\boldsymbol{k}\cdot\boldsymbol{r} = \omega t + n\,2\pi + \phi, \tag{3.223}$$

where n is an integer and ϕ is some phase angle. The solution to this equation is a set of equally spaced parallel planes (one plane for each possible value of n) whose normals lie in the direction of the wave vector $\boldsymbol{k}$ and which propagate in this direction with velocity

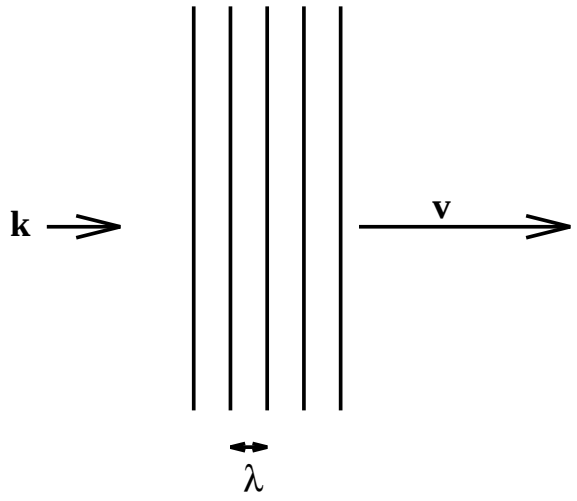
$$v = \frac{\omega}{k}. \tag{3.224}$$

The spacing between adjacent planes (*i.e.*, the wavelength) is given by

$$\lambda = \frac{2\pi}{k}. \tag{3.225}$$

Consider a general plane wave vector field

$$\boldsymbol{A} = \boldsymbol{A}_0\, \mathrm{e}^{\,\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}. \tag{3.226}$$



What is the divergence of $\boldsymbol{A}$? This is easy to evaluate. We have

$$\begin{aligned}\nabla\cdot\boldsymbol{A} &= \frac{\partial A_x}{\partial x}+\frac{\partial A_y}{\partial y}+\frac{\partial A_z}{\partial z} = (A_{0x}\,\mathrm{i}k_x + A_{0y}\,\mathrm{i}k_y + A_{0z}\,\mathrm{i}k_z)\,\mathrm{e}^{\,\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\\ &= \mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{A}.\end{aligned} \tag{3.227}$$

How about the curl of $\boldsymbol{A}$? This is slightly more difficult. We have

$$\begin{aligned}(\nabla\wedge\boldsymbol{A})_x &= \frac{\partial A_z}{\partial y}-\frac{\partial A_y}{\partial z} = (\mathrm{i}\,k_y A_z - \mathrm{i}\,k_z A_y)\\ &= \mathrm{i}\,(\boldsymbol{k}\wedge\boldsymbol{A})_x.\end{aligned} \tag{3.228}$$

This is easily generalized to

$$\nabla\wedge\boldsymbol{A} = \mathrm{i}\,\boldsymbol{k}\wedge\boldsymbol{A}. \tag{3.229}$$

We can see that vector field operations on a plane wave simplify to dot and cross products involving the wave-vector.

The first Maxwell equation (3.220a) reduces to

$$\mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{E}_0 = 0, \tag{3.230}$$

using the assumed electric and magnetic fields (3.222), and Eq. (3.227). Thus, the electric field is perpendicular to the direction of propagation of the wave.

Likewise, the second Maxwell equation gives

$$\mathrm{i}\,\boldsymbol{k}\cdot\boldsymbol{B}_0 = 0, \tag{3.231}$$

implying that the magnetic field is also perpendicular to the direction of propagation. Clearly, the wave-like solutions of Maxwell's equation are a type of *transverse wave*. The third Maxwell equation gives

$$\mathrm{i}\,\boldsymbol{k}\wedge\boldsymbol{E}_0 = \mathrm{i}\,\omega\boldsymbol{B}_0, \tag{3.232}$$

where use has been made of Eq. (3.229). Dotting this equation with $\boldsymbol{E}_0$ yields

$$\boldsymbol{E}_0\cdot\boldsymbol{B}_0 = \frac{\boldsymbol{E}_0\cdot\boldsymbol{k}\wedge\boldsymbol{E}_0}{\omega} = 0. \tag{3.233}$$

Thus, the electric and magnetic fields are mutually perpendicular. Dotting equation (3.232) with $\boldsymbol{B}_0$ yields

$$\boldsymbol{B}_0\cdot\boldsymbol{k}\wedge\boldsymbol{E}_0 = \omega\,{B_0}^2 > 0. \tag{3.234}$$

Thus, the vectors $\boldsymbol{E}_0$, $\boldsymbol{B}_0$, and $\boldsymbol{k}$ are mutually perpendicular and form a right-handed set. The final Maxwell equation gives

$$\mathrm{i}\,\boldsymbol{k}\wedge\boldsymbol{B}_0 = -\mathrm{i}\,\epsilon_0\mu_0\,\omega\boldsymbol{E}_0. \tag{3.235}$$

Combining this with Eq. (3.232) yields

$$\boldsymbol{k}\wedge(\boldsymbol{k}\wedge\boldsymbol{E}_0) = (\boldsymbol{k}\cdot\boldsymbol{E}_0)\,\boldsymbol{k} - k^2\,\boldsymbol{E}_0 = -\epsilon_0\mu_0\,\omega^2\,\boldsymbol{E}_0, \tag{3.236}$$

or

$$k^2 = \epsilon_0\mu_0\,\omega^2, \tag{3.237}$$

where use has been made of Eq. (3.230). However, we know from Eq. (3.224) that the wave-velocity c is related to the magnitude of the wave-vector and the wave frequency via $c = \omega/k$. Thus, we obtain

$$c = \frac{1}{\sqrt{\epsilon_0\mu_0}}. \tag{3.238}$$

We have found transverse wave solutions of the free-space Maxwell equations, propagating at some velocity c which is given by a combination of ϵ_0 and μ_0. The constants ϵ_0 and μ_0 are easily measurable. The former is related to the force acting between electric charges and the latter to the force acting between electric currents. Both of these constants were fairly well known in Maxwell's time. Maxwell, incidentally, was the first person to look for wave-like solutions of his equations and, thus, to derive Eq. (3.238). The modern values of ϵ_0 and μ_0 are

$$\begin{aligned} \epsilon_0 &= 8.8542 \times 10^{-12}\,\mathrm{C^2\,N^{-1}\,m^{-2}}, \\ \mu_0 &= 4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}. \end{aligned} \tag{3.239}$$

Let us use these values to find the propagation velocity of "electromagnetic waves." We get

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} = 2.998 \times 10^8\,\mathrm{m\,s^{-1}}. \tag{3.240}$$

Of course, we immediately recognize this as the velocity of light. Maxwell also made this connection back in the 1870's. He conjectured that light, whose nature has been previously unknown, was a form of electromagnetic radiation. This was a remarkable prediction. After all, Maxwell's equations were derived from the results of benchtop laboratory experiments involving charges, batteries, coils, and currents, which apparently had nothing whatsoever to do with light.

Maxwell was able to make another remarkable prediction. The wavelength of light was well known in the late nineteenth century from studies of diffraction through slits, *etc.* Visible light actually occupies a surprisingly narrow wavelength range. The shortest wavelength blue light which is visible has $\lambda = 0.4$ microns (one micron is 10^{-6} meters). The longest wavelength red light which is visible has $\lambda = 0.76$ microns. However, there is nothing in our analysis which suggests that this particular range of wavelengths is special. Electromagnetic waves can have any wavelength. Maxwell concluded that visible light was a small part of a vast spectrum of previously undiscovered types of electromagnetic radiation. Since Maxwell's time virtually all of the non-visible parts of the electromagnetic spectrum have been observed. Table 1 gives a brief guide to the electromagnetic spectrum. Electromagnetic waves are of particular importance because they are our only source of information regarding the universe around us. Radio waves and

Radiation Type	Wavelength Range (m)
Gamma Rays	$< 10^{-11}$
X-Rays	10^{-11}–10^{-9}
Ultraviolet	10^{-9}–10^{-7}
Visible	10^{-7}–10^{-6}
Infrared	10^{-6}–10^{-4}
Microwave	10^{-4}–10^{-1}
TV-FM	10^{-1}–10^{1}
Radio	$> 10^{1}$

Table 1: The electromagnetic spectrum

microwaves (which are comparatively hard to scatter) have provided much of our knowledge about the centre of our own galaxy. This is completely unobservable in visible light, which is strongly scattered by interstellar gas and dust lying in the galactic plane. For the same reason, the spiral arms of our galaxy can only be mapped out using radio waves. Infrared radiation is useful for detecting proto-stars which are not yet hot enough to emit visible radiation. Of course, visible radiation is still the mainstay of astronomy. Satellite based ultraviolet observations have yielded invaluable insights into the structure and distribution of distant galaxies. Finally, X-ray and γ-ray astronomy usually concentrates on exotic objects in the Galaxy such as pulsars and supernova remnants.

Equations (3.230), (3.232), and the relation $c = \omega/k$, imply that

$$B_0 = \frac{E_0}{c}. \tag{3.241}$$

Thus, the magnetic field associated with an electromagnetic wave is smaller in magnitude than the electric field by a factor c. Consider a free charge interacting with an electromagnetic wave. The force exerted on the charge is given by the Lorentz formula

$$\boldsymbol{f} = q\,(\boldsymbol{E} + \boldsymbol{v} \wedge \boldsymbol{B}). \tag{3.242}$$

The ratio of the electric and magnetic forces is

$$\frac{f_{\rm magnetic}}{f_{\rm electric}} \sim \frac{v\,B_0}{E_0} \sim \frac{v}{c}. \tag{3.243}$$

So, unless the charge is relativistic the electric force greatly exceeds the magnetic force. Clearly, in most terrestrial situations electromagnetic waves are an essentially *electric* phenomenon (as far as their interaction with matter goes). For this reason, electromagnetic waves are usually characterized by their wave-vector (which specifies the direction of propagation and the wavelength) and the plane of polarization (*i.e.*, the plane of oscillation) of the associated electric field. For a given wave-vector $\boldsymbol{k}$, the electric field can have any direction in the plane normal to $\boldsymbol{k}$. However, there are only two *independent* directions in a plane (*i.e.*, we can only define two linearly independent vectors in a plane). This implies that there are only two independent polarizations of an electromagnetic wave, once its direction of propagation is specified.

Let us now derive the velocity of light from Maxwell's equation the hard way. Suppose that we take the curl of the fourth Maxwell equation, Eq. (3.220d). We obtain

$$\nabla \wedge \nabla \wedge \boldsymbol{B} = \nabla(\nabla\cdot\boldsymbol{B}) - \nabla^2\boldsymbol{B} = -\nabla^2\boldsymbol{B} = \epsilon_0\mu_0\,\frac{\partial\,\nabla\wedge\boldsymbol{E}}{\partial t}. \tag{3.244}$$

Here, we have used the fact that $\nabla\cdot\boldsymbol{B} = 0$. The third Maxwell equation, Eq. (3.220c), yields

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{B} = \boldsymbol{0}, \tag{3.245}$$

where use has been made of Eq. (3.238). A similar equation can obtained for the electric field by taking the curl of Eq. (3.220c):

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{E} = \boldsymbol{0}, \tag{3.246}$$

We have found that electric and magnetic fields both satisfy equations of the form

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{A} = \boldsymbol{0} \tag{3.247}$$

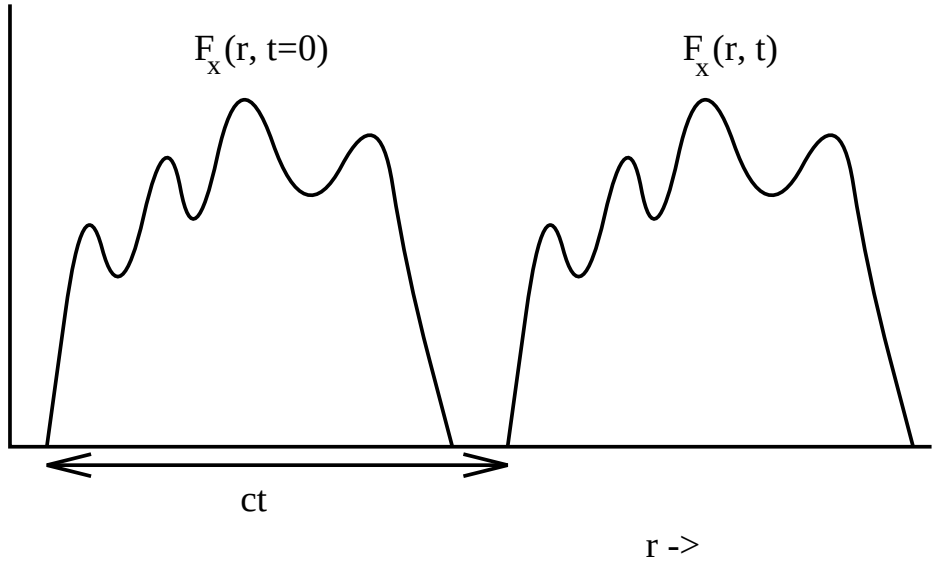
in free space. As is easily verified, the most general solution to this equation (with a positive frequency) is

$$\begin{aligned} A_x &= F_x(\boldsymbol{k}\cdot\boldsymbol{r} - kc\,t), \\ A_y &= F_y(\boldsymbol{k}\cdot\boldsymbol{r} - kc\,t), \\ A_z &= F_z(\boldsymbol{k}\cdot\boldsymbol{r} - kc\,t), \end{aligned} \tag{3.248}$$

where $F_x(\phi)$, $F_y(\phi)$, and $F_z(\phi)$ are one-dimensional scalar functions. Looking along the direction of the wave-vector, so that $\boldsymbol{r} = (\boldsymbol{k}/k)\,r$, we find that

$$\begin{aligned} A_x &= F_x(\,k\,(r - ct)\,), \\ A_y &= F_y(\,k\,(r - ct)\,), \\ A_z &= F_z(\,k\,(r - ct)\,). \end{aligned} \tag{3.249}$$

The x-component of this solution is shown schematically below; it clearly propagates in r with velocity c. If we look along a direction which is perpendicular to $\boldsymbol{k}$ then $\boldsymbol{k}\cdot\boldsymbol{r} = 0$ and there is no propagation. Thus, the components of $\boldsymbol{A}$ are arbitrarily shaped pulses which propagate, without changing shape, along the direction of $\boldsymbol{k}$ with velocity c. These pulses can be related to the sinusoidal



plane wave solutions which we found earlier by Fourier transformation. Thus, any arbitrary shaped pulse propagating in the direction of $\boldsymbol{k}$ with velocity c can be

broken down into lots of sinusoidal oscillations propagating in the same direction with the same velocity.

The operator

$$\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \tag{3.250}$$

is called the d'Alembertian. It is the four dimensional equivalent of the Laplacian. Recall that the Laplacian is invariant under rotational transformation. The d'Alembertian goes one better than this because it is both rotationally invariant and *Lorentz invariant*. The d'Alembertian is conventionally denoted $\Box^2$. Thus, electromagnetic waves in free space satisfy the wave equations

$$\begin{aligned} \Box^2 \boldsymbol{E} &= \boldsymbol{0}, \\ \Box^2 \boldsymbol{B} &= \boldsymbol{0}. \end{aligned} \tag{3.251}$$

When written in terms of the vector and scalar potentials, Maxwell's equations reduce to

$$\begin{aligned} \Box^2 \phi &= -\frac{\rho}{\epsilon_0}, \\ \Box^2 \boldsymbol{A} &= -\mu_0 \boldsymbol{j}. \end{aligned} \tag{3.252}$$

These are clearly driven wave equations. Our next task is to find the solutions to these equations.

3.20 Green's functions

Earlier on in this lecture course we had to solve Poisson's equation

$$\nabla^2 u = v, \tag{3.253}$$

where $v(\boldsymbol{r})$ is denoted the source function. The potential $u(\boldsymbol{r})$ satisfies the boundary condition

$$u(\boldsymbol{r}) \rightarrow 0 \qquad \text{as } |\boldsymbol{r}| \rightarrow \infty, \tag{3.254}$$

provided that the source function is reasonably localized. The solutions to Poisson's equation are superposable (because the equation is linear). This property

is exploited in the Green's function method of solving this equation. The Green's function $G(\boldsymbol{r},\boldsymbol{r}')$ is the potential, which satisfies the appropriate boundary conditions, generated by a unit amplitude point source located at $\boldsymbol{r}'$. Thus,

$$\nabla^2 G(\boldsymbol{r},\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{3.255}$$

Any source function $v(\boldsymbol{r})$ can be represented as a weighted sum of point sources

$$v(\boldsymbol{r}) = \int \delta(\boldsymbol{r}-\boldsymbol{r}')\, v(\boldsymbol{r}')\, d^3\boldsymbol{r}'. \tag{3.256}$$

It follows from superposability that the potential generated by the source $v(\boldsymbol{r})$ can be written as the weighted sum of point source driven potentials (*i.e.*, Green's functions)

$$u(\boldsymbol{r}) = \int G(\boldsymbol{r},\boldsymbol{r}')\, v(\boldsymbol{r}')\, d^3\boldsymbol{r}'. \tag{3.257}$$

We found earlier that the Green's function for Poisson's equation is

$$G(\boldsymbol{r},\boldsymbol{r}') = -\frac{1}{4\pi}\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.258}$$

It follows that the general solution to Eq. (3.253) is written

$$u(\boldsymbol{r}) = -\frac{1}{4\pi}\int \frac{v(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\, d^3\boldsymbol{r}'. \tag{3.259}$$

Note that the point source driven potential (3.258) is perfectly sensible. It is spherically symmetric about the source, and falls off smoothly with increasing distance from the source.

We now need to solve the wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) u = v, \tag{3.260}$$

where $v(\boldsymbol{r},t)$ is a time varying source function. The potential $u(\boldsymbol{r},t)$ satisfies the boundary conditions

$$u(\boldsymbol{r}) \rightarrow 0 \qquad \text{as } |\boldsymbol{r}| \rightarrow \infty \text{ and } |t| \rightarrow \infty. \tag{3.261}$$

The solutions to Eq. (3.260) are superposable (since the equation is linear), so a Green's function method of solution is again appropriate. The Green's function $G(\boldsymbol{r},\boldsymbol{r}';t,t')$ is the potential generated by a point *impulse* located at position $\boldsymbol{r}'$ and applied at time t'. Thus,

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)G(\boldsymbol{r},\boldsymbol{r}';t,t')=\delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t'). \tag{3.262}$$

Of course, the Green's function must satisfy the correct boundary conditions. A general source $v(\boldsymbol{r},t)$ can be built up from a weighted sum of point impulses

$$v(\boldsymbol{r},t)=\int\int \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\,v(\boldsymbol{r}',t')\,d^3\boldsymbol{r}'\,dt'. \tag{3.263}$$

It follows that the potential generated by $v(\boldsymbol{r},t)$ can be written as the weighted sum of point impulse driven potentials

$$u(\boldsymbol{r},t)=\int\int G(\boldsymbol{r},\boldsymbol{r}';t,t')\,v(\boldsymbol{r}',t')\,d^3\boldsymbol{r}'\,dt'. \tag{3.264}$$

So, how do we find the Green's function?

Consider

$$G(\boldsymbol{r},\boldsymbol{r}';t,t')=\frac{F(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{3.265}$$

where $F(\phi)$ is a general scalar function. Let us try to prove the following theorem:

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)G=-4\pi\,F(t-t')\,\delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{3.266}$$

At a general point, $\boldsymbol{r}\neq\boldsymbol{r}'$, the above expression reduces to

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)G=0. \tag{3.267}$$

So, we basically have to show that G is a valid solution of the free space wave equation. We can easily show that

$$\frac{\partial|\boldsymbol{r}-\boldsymbol{r}'|}{\partial x}=\frac{x-x'}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.268}$$

It follows by simple differentiation that

$$\frac{\partial^2 G}{\partial x^2} = \left(\frac{3(x-x')^2 - |\boldsymbol{r}-\boldsymbol{r}'|^2}{|\boldsymbol{r}-\boldsymbol{r}'|^5}\right) F + \left(\frac{3(x-x')^2 - |\boldsymbol{r}-\boldsymbol{r}'|^2}{|\boldsymbol{r}-\boldsymbol{r}'|^4}\right) \frac{F'}{c} + \frac{(x-x')^2}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\frac{F''}{c^2}, \tag{3.269}$$

where $F'(\phi) = dF(\phi)/d\phi$. We can derive analogous equations for $\partial^2 G/\partial y^2$ and $\partial^2 G/\partial z^2$. Thus,

$$\nabla^2 G = \frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} + \frac{\partial^2 G}{\partial z^2} = \frac{F''}{|\boldsymbol{r}-\boldsymbol{r}'|\,c^2} = \frac{1}{c^2}\frac{\partial^2 G}{\partial t^2}, \tag{3.270}$$

giving

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G = 0, \tag{3.271}$$

which is the desired result. Consider, now, the region around $\boldsymbol{r} = \boldsymbol{r}'$. It is clear from Eq. (3.269) that the dominant term on the left-hand side as $|\boldsymbol{r}-\boldsymbol{r}'| \rightarrow 0$ is the first one, which is essentially $F\,\partial^2(|\boldsymbol{r}-\boldsymbol{r}'|^{-1})/\partial x^2$. It is also clear that $(1/c^2)(\partial^2 G/\partial t^2)$ is negligible compared to this term. Thus, as $|\boldsymbol{r}-\boldsymbol{r}'| \rightarrow 0$ we find that

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G \rightarrow F(t-t')\,\nabla^2\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right). \tag{3.272}$$

However, according to Eqs. (3.255) and (3.258)

$$\nabla^2\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) = -4\pi\,\delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{3.273}$$

We conclude that

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G = -4\pi\,F(t-t')\,\delta(\boldsymbol{r}-\boldsymbol{r}'), \tag{3.274}$$

which is the desired result.

Let us now make the special choice

$$F(\phi) = -\frac{\delta(\phi)}{4\pi}. \tag{3.275}$$

It follows from Eq. (3.274) that

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G = \delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t'). \tag{3.276}$$

Thus,

$$G(\boldsymbol{r}, \boldsymbol{r}'; t, t') = -\frac{1}{4\pi}\frac{\delta(t - t' - |\boldsymbol{r} - \boldsymbol{r}'|/c)}{|\boldsymbol{r} - \boldsymbol{r}'|} \tag{3.277}$$

is the Green's function for the driven wave equation (3.260).

The time dependent Green's function (3.277) is the same as the steady state Green's function (3.258), apart from the delta function appearing in the former. What does this delta function do? Well, consider an observer at point $\boldsymbol{r}$. Because of the delta function our observer only measures a non-zero potential at one particular time

$$t = t' + \frac{|\boldsymbol{r} - \boldsymbol{r}'|}{c}. \tag{3.278}$$

It is clear that this is the time the impulse was applied at position $\boldsymbol{r}'$ (*i.e.*, t') *plus* the time taken for a light signal to travel between points $\boldsymbol{r}'$ and $\boldsymbol{r}$. At time $t > t'$ the locus of all points at which the potential is non-zero is

$$|\boldsymbol{r} - \boldsymbol{r}'| = c\,(t - t'). \tag{3.279}$$

In other words, it is a sphere centred on $\boldsymbol{r}'$ whose radius is the distance traveled by light in the time interval since the impulse was applied at position $\boldsymbol{r}'$. Thus, the Green's function (3.277) describes a spherical wave which emanates from position $\boldsymbol{r}'$ at time t' and propagates at the speed of light. The amplitude of the wave is inversely proportional to the distance from the source.

3.21 Retarded potentials

We are now in a position to solve Maxwell's equations. Recall that in steady state Maxwell's equations reduce to

$$\begin{aligned} \nabla^2\phi &= -\frac{\rho}{\epsilon_0}, \\ \nabla^2\boldsymbol{A} &= -\mu_0\boldsymbol{j}. \end{aligned} \tag{3.280}$$

The solutions to these equations are easily found using the Green's function for Poisson's equation (3.258):

$$\begin{aligned}\phi(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'\\ \boldsymbol{A}(\boldsymbol{r}) &= \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'.\end{aligned} \tag{3.281}$$

The time dependent Maxwell equations reduce to

$$\begin{aligned}\Box^2\phi &= -\frac{\rho}{\epsilon_0},\\ \Box^2\boldsymbol{A} &= -\mu_0\boldsymbol{j}.\end{aligned} \tag{3.282}$$

We can solve these equations using the time dependent Green's function (3.277). From Eq. (3.264) we find that

$$\phi(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon_0}\int\int \frac{\delta(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)\,\rho(\boldsymbol{r}',t')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'\,dt', \tag{3.283}$$

with a similar equation for $\boldsymbol{A}$. Using the well known property of delta functions, these equations reduce to

$$\begin{aligned}\phi(\boldsymbol{r},t) &= \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}',t-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'\\ \boldsymbol{A}(\boldsymbol{r},t) &= \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{j}(\boldsymbol{r}',t-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'.\end{aligned} \tag{3.284}$$

These are the general solutions to Maxwell's equations! Note that the time dependent solutions (3.284) are the same as the steady state solutions (3.281), apart from the weird way in which time appears in the former. According to Eqs. (3.284), if we want to work out the potentials at position $\boldsymbol{r}$ and time t we have to perform integrals of the charge density and current density over all space (just like in the steady state situation). However, when we calculate the contribution of charges and currents at position $\boldsymbol{r}'$ to these integrals we do not use the values at time t, instead we use the values at some earlier time $t-|\boldsymbol{r}-\boldsymbol{r}'|/c$. What is this earlier time? It is simply the latest time at which a light signal emitted

from position $\boldsymbol{r}'$ would be received at position $\boldsymbol{r}$ before time t. This is called the *retarded time*. Likewise, the potentials (3.284) are called *retarded potentials*. It is often useful to adopt the following notation

$$A(\boldsymbol{r}', t - |\boldsymbol{r} - \boldsymbol{r}'|/c) \equiv [A(\boldsymbol{r}', t)]\,. \tag{3.285}$$

The square brackets denote retardation (*i.e.*, using the retarded time instead of the real time). Using this notation Eqs. (3.284) become

$$\begin{aligned} \phi(\boldsymbol{r}) &= \frac{1}{4\pi\epsilon_0}\int \frac{[\rho(\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}', \\ \boldsymbol{A}(\boldsymbol{r}) &= \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{j}(\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \end{aligned} \tag{3.286}$$

The time dependence in the above equations is taken as read.

We are now in a position to understand electromagnetism at its most fundamental level. A charge distribution $\rho(\boldsymbol{r}, t)$ can thought of as built up out of a collection or series of charges which instantaneously come into existence, at some point $\boldsymbol{r}'$ and some time t', and then disappear again. Mathematically, this is written

$$\rho(\boldsymbol{r}, t) = \int\int \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\,\rho(\boldsymbol{r}', t')\,d^3\boldsymbol{r}'dt'. \tag{3.287}$$

Likewise, we can think of a current distribution $\boldsymbol{j}(\boldsymbol{r}, t)$ as built up out of a collection or series of currents which instantaneously appear and then disappear:

$$\boldsymbol{j}(\boldsymbol{r}, t) = \int\int \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t-t')\,\boldsymbol{j}(\boldsymbol{r}', t')\,d^3\boldsymbol{r}'dt'. \tag{3.288}$$

Each of these ephemeral charges and currents excites a spherical wave in the appropriate potential. Thus, the charge density at $\boldsymbol{r}'$ and t' sends out a wave in the scalar potential:

$$\phi(\boldsymbol{r}, t) = \frac{\rho(\boldsymbol{r}', t')}{4\pi\epsilon_0}\,\frac{\delta(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.289}$$

Likewise, the current density at $\boldsymbol{r}'$ and t' sends out a wave in the vector potential:

$$\boldsymbol{A}(\boldsymbol{r}, t) = \frac{\mu_0\,\boldsymbol{j}(\boldsymbol{r}', t')}{4\pi}\,\frac{\delta(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.290}$$

These waves can be thought of as little messengers which inform other charges and currents about the charges and currents present at position $\boldsymbol{r}'$ and time t'. However, the messengers travel at a finite speed; *i.e.*, the speed of light. So, by the time they reach other charges and currents their message is a little out of date. Every charge and every current in the universe emits these spherical waves. The resultant scalar and vector potential fields are given by Eqs. (3.286). Of course, we can turn these fields into electric and magnetic fields using Eqs. (3.187). We can then evaluate the force exerted on charges using the Lorentz formula. We can see that we have now escaped from the apparent action at a distance nature of Coulomb's law and the Biot-Savart law. Electromagnetic information is carried by spherical waves in the vector and scalar potentials and, therefore, travels at the velocity of light. Thus, if we change the position of a charge then a distant charge can only respond after a time delay sufficient for a spherical wave to propagate from the former to the latter charge.

Let us compare the steady-state law

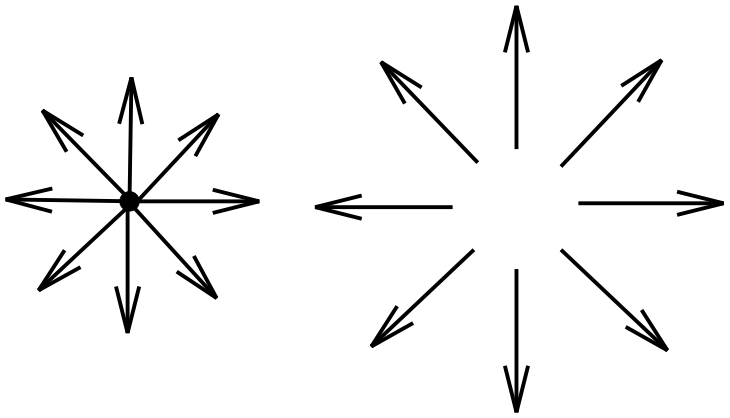
$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}' \tag{3.291}$$

with the corresponding time dependent law

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{[\rho(\boldsymbol{r}')]}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}' \tag{3.292}$$

These two formulae look very similar indeed, but there is an important difference. We can imagine (rather pictorially) that every charge in the universe is continuously performing the integral (3.291), and is also performing a similar integral to find the vector potential. After evaluating both potentials, the charge can calculate the fields and using the Lorentz force law it can then work out its equation of motion. The problem is that the information the charge receives from the rest of the universe is carried by our spherical waves, and is always slightly out of date (because the waves travel at a finite speed). As the charge considers more and more distant charges or currents its information gets more and more out of date. (Similarly, when astronomers look out to more and more distant galaxies in the universe they are also looking backwards in time. In fact, the light we receive from the most distant observable galaxies was emitted when the universe was

only about a third of its present age.) So, what does our electron do? It simply uses the most up to date information about distant charges and currents which it possesses. So, instead of incorporating the charge density $\rho(\boldsymbol{r},t)$ in its integral the electron uses the *retarded* charge density $[\rho(\boldsymbol{r},t)]$ (*i.e.*, the density evaluated at the retarded time). This is effectively what Eq. (3.292) says.



Consider a thought experiment in which a charge q appears at position $\boldsymbol{r}_0$ at time t_1, persists for a while, and then disappears at time t_2. What is the electric field generated by such a charge? Using Eq. (3.292), we find that

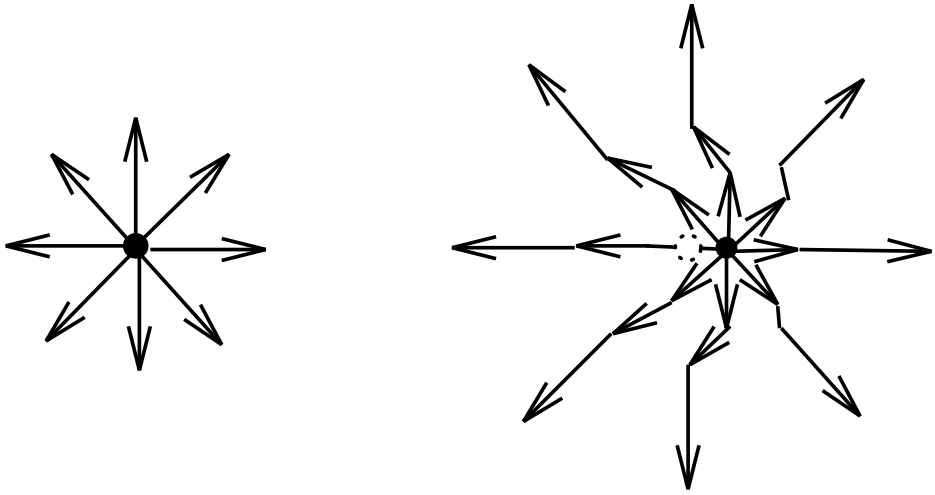
$$\begin{aligned}\phi(\boldsymbol{r}) &= \frac{q}{4\pi\epsilon_0}\frac{1}{|\boldsymbol{r}-\boldsymbol{r}_0|} \qquad \text{for } t_1 \leq t - |\boldsymbol{r}-\boldsymbol{r}_0|/c \leq t_2\\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.}\end{aligned} \tag{3.293}$$

Now, $\boldsymbol{E} = -\nabla\phi$ (since there are no currents, and therefore no vector potential is generated), so

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r}) &= \frac{q}{4\pi\epsilon_0}\frac{\boldsymbol{r}-\boldsymbol{r}_0}{|\boldsymbol{r}-\boldsymbol{r}_0|^3} \qquad \text{for } t_1 \leq t - |\boldsymbol{r}-\boldsymbol{r}_0|/c \leq t_2\\ &= \boldsymbol{0} \qquad\qquad\qquad\qquad \text{otherwise.}\end{aligned} \tag{3.294}$$

This solution is shown pictorially above. We can see that the charge effectively emits a Coulomb electric field which propagates radially away from the charge at the speed of light. Likewise, it is easy to show that a current carrying wire effectively emits an Ampèrian magnetic field at the speed of light.

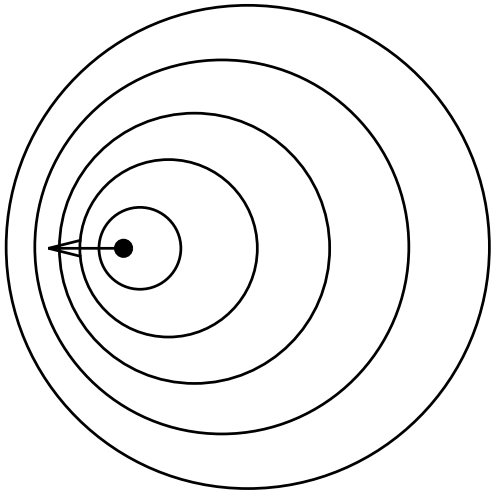
We can now appreciate the essential difference between time dependent electromagnetism and the action at a distance laws of Coulomb and Biot & Savart. In the latter theories, the field lines act rather like rigid wires attached to charges (or circulating around currents). If the charges (or currents) move then so do the field lines, leading inevitably to unphysical action at a distance type behaviour. In the time dependent theory charges act rather like water sprinklers; *i.e.*, they spray out the Coulomb field in all directions at the speed of light. Similarly, current carrying wires throw out magnetic field loops at the speed of light. If we move a charge (or current) then field lines emitted beforehand are not affected, so the field at a distant charge (or current) only responds to the change in position after a time delay sufficient for the field to propagate between the two charges (or currents) at the speed of light. In Coulomb's law and the Biot-Savart law it is not



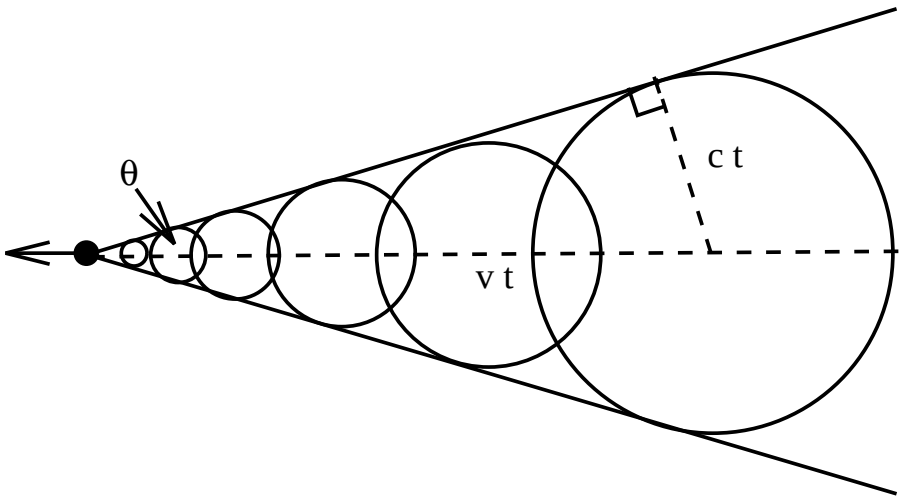
obvious that the electric and magnetic fields have any real existence. The only measurable quantities are the forces acting between charges and currents. We can describe the force acting on a given charge or current, due to the other charges and currents in the universe, in terms of the local electric and magnetic fields, but we have no way of knowing whether these fields persist when the charge or current is not present (*i.e.*, we could argue that electric and magnetic fields are just a convenient way of calculating forces but, in reality, the forces are transmitted directly between charges and currents by some form of magic). However, it is patently obvious that electric and magnetic fields have a real existence in the time dependent theory. Consider the following thought experiment. Suppose that a charge q_1 comes into existence for a period of time, emits a Coulomb field,

and then disappears. Suppose that a distant charge q_2 interacts with this field, but is sufficiently far from the first charge that by the time the field arrives the first charge has already disappeared. The force exerted on the second charge is only ascribable to the electric field; it cannot be ascribed to the first charge because this charge no longer exists by the time the force is exerted. The electric field clearly transmits energy and momentum between the two charges. Anything which possesses energy and momentum is "real" in a physical sense. Later on in this course we shall demonstrate that electric and magnetic fields conserve energy and momentum.

Let us now consider a moving charge. Such a charge is continually emitting spherical waves in the scalar potential, and the resulting wave front pattern is sketched below. Clearly, the wavefronts are more closely spaced in front of the



charge than they are behind it, suggesting that the electric field in front is larger than the field behind. In a medium, such as water or air, where waves travel at a finite speed c (say) it is possible to get a very interesting effect if the wave source travels at some velocity v which exceeds the wave speed. This is illustrated below. The locus of the outermost wave front is now a cone instead of a sphere. The wave intensity on the cone is extremely large: this is a shock wave! The half angle θ of the shock wave cone is simply $\cos^{-1}(c/v)$. In water, shock waves are produced by fast moving boats. We call these "bow waves." In air, shock waves are produced by speeding bullets and supersonic jets. In the latter case we call these "sonic booms." Is there any such thing as an electromagnetic



shock wave? At first sight, the answer to this question would appear to be, no. After all, electromagnetic waves travel at the speed of light and no wave source (*i.e.*, an electrically charged particle) can travel faster than this velocity. This is a rather disappointing conclusion. However, when an electromagnetic wave travels through matter a remarkable thing happens. The oscillating electric field of the wave induces a slight separation of the positive and negative charges in the atoms which make up the material. We call separated positive and negative charges an electric dipole. Of course, the atomic dipoles oscillate in sympathy with the field which induces them. However, an oscillating electric dipole radiates electromagnetic waves. Amazingly, when we add the original wave to these induced waves it is exactly as if the original wave propagates through the material in question at a velocity which is *slower* than the velocity of light in vacuum. Suppose, now, that we shoot a charged particle through the material faster than the slowed down velocity of electromagnetic waves. This is possible since the waves are traveling slower than the velocity of light in vacuum. In practice, the particle has to be traveling pretty close to the velocity of light in vacuum (*i.e.*, it has to be relativistic), but modern particle accelerators produce copious amount of such particles. Now, we can get an electromagnetic shock wave. We expect an intense cone of emission, just like the bow wave produced by a fast ship. In fact, this type of radiation has been observed. It is called Cherenkov radiation, and it is very useful in high energy physics. Cherenkov radiation is typically produced by surrounding a particle accelerator with perspex blocks. Relativistic charged particles emanating from the accelerator pass through the perspex traveling faster

than the local velocity of light and therefore emit Cherenkov radiation. We know the velocity of light (c_*, say) in perspex (this can be worked out from the refractive index), so if we can measure the half angle θ of the radiation cone emitted by each particle then we can evaluate the speed of the particle v via the geometric relation $\cos\theta = c_*/v$.

3.22 Advanced potentials?

We have defined the retarded time

$$t_r = t - |\boldsymbol{r} - \boldsymbol{r}'|/c \tag{3.295}$$

as the latest time at which a light signal emitted from position $\boldsymbol{r}'$ would reach position $\boldsymbol{r}$ before time t. We have also shown that a solution to Maxwell's equations can be written in terms of retarded potentials:

$$\phi(\boldsymbol{r}, t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}', t_r)}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}', \tag{3.296}$$

etc. But, is this the most general solution? Suppose that we define the *advanced* time.

$$t_a = t + |\boldsymbol{r} - \boldsymbol{r}'|/c. \tag{3.297}$$

This is the time a light signal emitted at time t from position $\boldsymbol{r}$ would reach position $\boldsymbol{r}'$. It turns out that we can also write a solution to Maxwell's equations in terms of *advanced potentials*:

$$\phi(\boldsymbol{r}, t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}', t_a)}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}', \tag{3.298}$$

etc. In fact, this is just as good a solution to Maxwell's equation as the one involving retarded potentials. To get some idea what is going on let us examine the Green's function corresponding to our retarded potential solution:

$$\phi(\boldsymbol{r}, t) = \frac{\rho(\boldsymbol{r}', t')}{4\pi\epsilon_0}\,\frac{\delta(t - t' - |\boldsymbol{r} - \boldsymbol{r}'|/c)}{|\boldsymbol{r} - \boldsymbol{r}'|}, \tag{3.299}$$

with a similar equation for the vector potential. This says that the charge density present at position $\boldsymbol{r}'$ and time t' emits a spherical wave in the scalar potential *which propagates forwards in time*. The Green's function corresponding to our advanced potential solution is

$$\phi(\boldsymbol{r},t) = \frac{\rho(\boldsymbol{r}',t')}{4\pi\epsilon_0}\frac{\delta(t-t'+|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.300}$$

This says that the charge density present at position $\boldsymbol{r}'$ and time t' emits a spherical wave in the scalar potential *which propagates backwards in time*. "But, hang on a minute," you might say, " everybody knows that electromagnetic waves can't travel backwards in time. If they did then causality would be violated." Well, *you* know that electromagnetic waves do not propagate backwards in time, *I* know that electromagnetic waves do not propagate backwards in time, but the question is do Maxwell's equations know this? Consider the wave equation for the scalar potential:

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\phi = -\frac{\rho}{\epsilon_0}. \tag{3.301}$$

This equation is manifestly symmetric in time (*i.e.*, it is invariant under the transformation $t \rightarrow -t$). Thus, backward traveling waves are just as good a solution to this equation as forward traveling waves. The equation is also symmetric in space (*i.e.*, it is invariant under the transformation $x \rightarrow -x$). So, why do we adopt the Green's function (3.299) which is symmetric in space (*i.e.*, it is invariant under $x \rightarrow -x$) but asymmetric in time (*i.e.*, it is not invariant under $t \rightarrow -t$)? Would it not be better to use the completely symmetric Green's function

$$\phi(\boldsymbol{r},t) = \frac{\rho(\boldsymbol{r}',t')}{4\pi\epsilon_0}\frac{1}{2}\left(\frac{\delta(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{\delta(t-t'+|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)? \tag{3.302}$$

In other words, a charge emits half of its waves running forwards in time (*i.e.*, retarded waves) and the other half running backwards in time (*i.e.*, advanced waves). This sounds completely crazy! However, in the 1940's Richard P. Feynman and John A. Wheeler pointed out that under certain circumstances this prescription gives the right answer. Consider a charge interacting with "the rest of the universe," where the "rest of the universe" denotes all of the distant charges in the universe and is, by implication, an awful long way away from our original

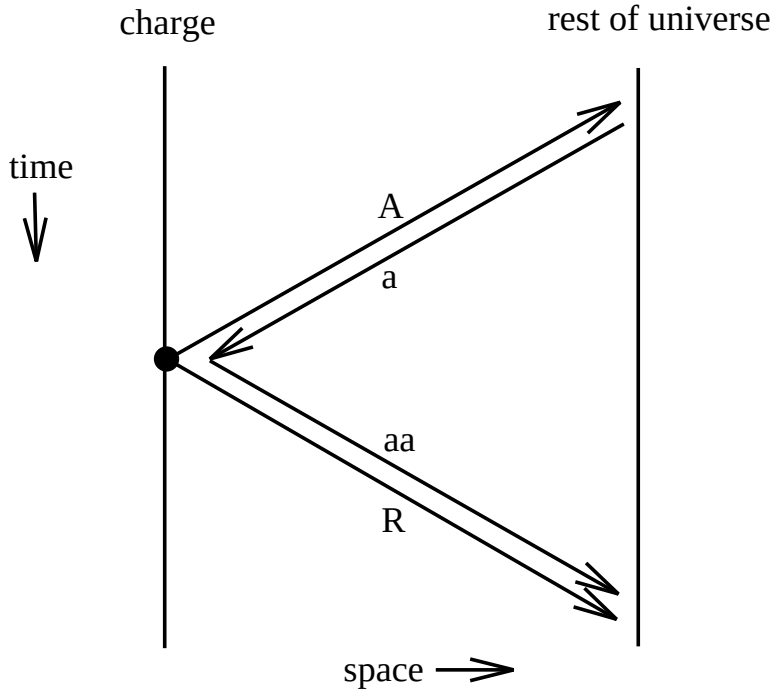
charge. Suppose that the "rest of the universe" is a perfect reflector of advanced waves and a perfect absorbed of retarded waves. The waves emitted by the charge can be written schematically as

$$F = \frac{1}{2}(\text{retarded}) + \frac{1}{2}(\text{advanced}). \tag{3.303}$$

The response of the rest of the universe is written

$$R = \frac{1}{2}(\text{retarded}) - \frac{1}{2}(\text{advanced}). \tag{3.304}$$

This is illustrated in the space-time diagram below. Here, A and R denote the



advanced and retarded waves emitted by the charge, respectively. The advanced wave travels to "the rest of the universe" and is reflected; *i.e.*, the distant charges oscillate in response to the advanced wave and emit a retarded wave a, as shown. The retarded wave a is spherical wave which converges on the original charge, passes through the charge, and then diverges again. The divergent wave is denoted aa. Note that a looks like a negative advanced wave emitted by the charge, whereas aa looks like a positive retarded wave emitted by the charge. This is

essentially what Eq. (3.304) says. The retarded waves R and aa are absorbed by "the rest of the universe."

If we add the waves emitted by the charge to the response of "the rest of the universe" we obtain

$$F' = F + R = (\text{retarded}). \tag{3.305}$$

Thus, charges *appear* to emit only retarded waves, which agrees with our everyday experience. Clearly, in this model we have side-stepped the problem of a time asymmetric Green's function by adopting time asymmetric boundary conditions to the universe; *i.e.*, the distant charges in the universe absorb retarded waves and reflect advanced waves. This is possible because the absorption takes place at the end of the universe (*i.e.*, at the "big crunch," or whatever) and the reflection takes place at the beginning of the universe (*i.e.*, at the "big bang"). It is quite plausible that the state of the universe (and, hence, its interaction with electromagnetic waves) is completely different at these two times. It should be pointed out that the Feynman-Wheeler model runs into trouble when one tries to combine electromagnetism with quantum mechanics. These difficulties have yet to be resolved, so at present the status of this model is that it is "an interesting idea" but it is still not fully accepted into the canon of physics.

3.23 Retarded fields

We know the solution to Maxwell's equations in terms of retarded potentials. Let us now construct the associated electric and magnetic fields using

$$\begin{aligned} \boldsymbol{E} &= -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \\ \boldsymbol{B} &= \nabla \wedge \boldsymbol{A}. \end{aligned} \tag{3.306}$$

It is helpful to write

$$\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}', \tag{3.307}$$

where $R = |\boldsymbol{r} - \boldsymbol{r}'|$. The retarded time becomes $t_r = t - R/c$, and a general retarded quantity is written $[F(\boldsymbol{r}, t)] \equiv F(\boldsymbol{r}, t_r)$. Thus, we can write the retarded

potential solutions of Maxwell's equations in the especially compact form:

$$\begin{aligned}\phi &= \frac{1}{4\pi\epsilon_0}\int \frac{[\rho]}{R}\,dV',\\ \boldsymbol{A} &= \frac{\mu_0}{4\pi}\int \frac{[\boldsymbol{j}]}{R}\,dV',\end{aligned} \tag{3.308}$$

where $dV' \equiv d^3\boldsymbol{r}'$.

It is easily seen that

$$\begin{aligned}\nabla\phi &= \frac{1}{4\pi\epsilon_0}\int \left([\rho]\nabla(R^{-1}) + \frac{[\partial\rho/\partial t]}{R}\,\nabla t_r\right) dV'\\ &= -\frac{1}{4\pi\epsilon_0}\int \left(\frac{[\rho]}{R^3}\,\boldsymbol{R} - \frac{[\partial\rho/\partial t]}{cR^2}\,\boldsymbol{R}\right) dV',\end{aligned} \tag{3.309}$$

where use has been made of

$$\nabla R = \frac{\boldsymbol{R}}{R},\quad \nabla(R^{-1}) = -\frac{\boldsymbol{R}}{R^3},\quad \nabla t_r = -\frac{\boldsymbol{R}}{cR}. \tag{3.310}$$

Likewise,

$$\begin{aligned}\nabla\wedge\boldsymbol{A} &= \frac{\mu_0}{4\pi}\int \left(\nabla(R^{-1})\wedge[\boldsymbol{j}] + \frac{\nabla t_r\wedge[\partial\boldsymbol{j}/\partial t]}{R}\right) dV'\\ &= -\frac{\mu_0}{4\pi}\int \left(\frac{\boldsymbol{R}\wedge[\boldsymbol{j}]}{R^3} + \frac{\boldsymbol{R}\wedge[\partial\boldsymbol{j}/\partial t]}{cR^2}\right) dV'.\end{aligned} \tag{3.311}$$

Equations (3.306), (3.309), and (3.311) can be combined to give

$$\boldsymbol{E} = \frac{1}{4\pi\epsilon_0}\int \left([\rho]\,\frac{\boldsymbol{R}}{R^3} + \left[\frac{\partial\rho}{\partial t}\right]\,\frac{\boldsymbol{R}}{cR^2} - \frac{[\partial\boldsymbol{j}/\partial t]}{c^2R}\right) dV', \tag{3.312}$$

which is the time dependent generalization of Coulomb's law, and

$$\boldsymbol{B} = \frac{\mu_0}{4\pi}\int \left(\frac{[\boldsymbol{j}]\wedge\boldsymbol{R}}{R^3} + \frac{[\partial\boldsymbol{j}/\partial t]\wedge\boldsymbol{R}}{cR^2}\right) dV', \tag{3.313}$$

which is the time dependent generalization of the Biot-Savart law.

Suppose that the typical variation time-scale of our charges and currents is t_0. Let us define $R_0 = c\,t_0$ which is the distance a light ray travels in time t_0. We can evaluate Eqs. (3.312) and (3.313) in two asymptotic limits: the "near field" region $R \ll R_0$, and the "far field" region $R \gg R_0$. In the near field region

$$\frac{|t - t_r|}{t_0} = \frac{R}{R_0} \ll 1, \tag{3.314}$$

so the difference between retarded time and standard time is relatively small. This allows us to expand retarded quantities in a Taylor series. Thus,

$$[\rho] \simeq \rho + \frac{\partial \rho}{\partial t}\,(t_r - t) + \frac{1}{2}\frac{\partial^2 \rho}{\partial t^2}\,(t_r - t) + \cdots, \tag{3.315}$$

giving

$$[\rho] \simeq \rho - \frac{\partial \rho}{\partial t}\frac{R}{c} + \frac{1}{2}\frac{\partial^2 \rho}{\partial t^2}\frac{R^2}{c^2} + \cdots. \tag{3.316}$$

Expansion of the retarded quantities in the near field region yields

$$\boldsymbol{E} \simeq \frac{1}{4\pi\epsilon_0}\int\left(\frac{\rho\,\boldsymbol{R}}{R^3} - \frac{1}{2}\frac{\partial^2\rho}{\partial t^2}\frac{\boldsymbol{R}}{c^2 R} - \frac{\partial \boldsymbol{j}/\partial t}{c^2 R} + \cdots\right) dV', \tag{3.317a}$$

$$\boldsymbol{B} \simeq \frac{\mu_0}{4\pi}\int\left(\frac{\boldsymbol{j}\wedge\boldsymbol{R}}{R^3} - \frac{1}{2}\frac{(\partial^2\boldsymbol{j}/\partial t^2)\wedge\boldsymbol{R}}{c^2 R} + \cdots\right) dV'. \tag{3.317b}$$

In Eq. (3.317a) the first term on the right-hand side corresponds to Coulomb's law, the second term is the correction due to retardation effects, and the third term corresponds to Faraday induction. In Eq. (3.317b) the first term on the right-hand side is the Biot-Savart law and the second term is the correction due to retardation effects. Note that the retardation corrections are only of order $(R/R_0)^2$. We might suppose, from looking at Eqs. (3.312) and (3.313), that the corrections should be of order R/R_0, however all of the order R/R_0 terms canceled out in the previous expansion. Suppose, then, that we have a d.c. circuit sitting on a laboratory benchtop. Let the currents in the circuit change on a typical time-scale of one tenth of a second. In this time light can travel about 3×10^7 meters, so $R_0 \sim 30,000$ kilometers. The length-scale of the experiment is

about one meter, so $R = 1$ meter. Thus, the retardation corrections are of order $(3 \times 10^7)^{-2} \sim 10^{-15}$. It is clear that we are fairly safe just using Coulomb's law, Faraday's law, and the Biot-Savart law to analyze the fields generated by this type of circuit.

In the far field region, $R \gg R_0$, Eqs. (3.312) and (3.313) are dominated by the terms which vary like R^{-1}, so

$$\boldsymbol{E} \simeq -\frac{1}{4\pi\epsilon_0}\int \frac{[\partial \boldsymbol{j}_\perp/\partial t]}{c^2 R}\, dV', \tag{3.318a}$$

$$\boldsymbol{B} \simeq \frac{\mu_0}{4\pi}\int \frac{[\partial \boldsymbol{j}_\perp/\partial t]\wedge \boldsymbol{R}}{cR^2}\, dV', \tag{3.318b}$$

where

$$\boldsymbol{j}_\perp = \boldsymbol{j} - \frac{(\boldsymbol{j}\cdot\boldsymbol{R})}{R^2}\,\boldsymbol{R}. \tag{3.318c}$$

Here, use has been made of $[\partial\rho/\partial t] = -[\nabla\cdot\boldsymbol{j}]$ and $[\nabla\cdot\boldsymbol{j}] = -[\partial\boldsymbol{j}/\partial t]\cdot\boldsymbol{R}/cR + O(1/R^2)$. Suppose that our charges and currents are localized to some region in the vicinity of $\boldsymbol{r}' = \boldsymbol{r}_*$. Let $\boldsymbol{R}_* = \boldsymbol{r} - \boldsymbol{r}_*$, with $R_* = |\boldsymbol{r} - \boldsymbol{r}_*|$. Suppose that the extent of the current and charge containing region is much less than R_*. It follows that retarded quantities can be written

$$[\rho(\boldsymbol{r}, t)] \simeq \rho(\boldsymbol{r}, t - R_*/c), \tag{3.319}$$

etc. Thus, the electric field reduces to

$$\boldsymbol{E} \simeq -\frac{1}{4\pi\epsilon_0}\frac{\left[\int \partial\boldsymbol{j}_\perp/\partial t\; dV'\right]}{c^2 R_*}, \tag{3.320}$$

whereas the magnetic field is given by

$$\boldsymbol{B} \simeq \frac{1}{4\pi\epsilon_0}\frac{\left[\int \partial\boldsymbol{j}_\perp/\partial t\; dV'\right]\wedge\boldsymbol{R}_*}{c^3 {R_*}^2}. \tag{3.321}$$

Note that

$$\frac{E}{B} = c, \tag{3.322}$$

and

$$\boldsymbol{E}\cdot\boldsymbol{B}=0. \tag{3.323}$$

This configuration of electric and magnetic fields is characteristic of an electromagnetic wave (see Section 3.19). Thus, Eqs. (3.322) and (3.323) describe an electromagnetic wave propagating *radially* away from the charge and current containing region. Note that the wave is driven by time varying electric currents. Now, charges moving with a constant velocity constitute a steady current, so a non-steady current is associated with *accelerating charges*. We conclude that accelerating electric charges emit electromagnetic waves. The wave fields, (3.320) and (3.321), fall off like the inverse of the distance from the wave source. This behaviour should be contrasted with that of Coulomb or Biot-Savart fields which fall off like the inverse square of the distance from the source. The fact that wave fields attenuate fairly gently with increasing distance from the source is what makes astronomy possible. If wave fields obeyed an inverse square law then no appreciable radiation would reach us from the rest of the universe.

In conclusion, electric and magnetic fields look simple in the near field region (they are just Coulomb fields, *etc.*) and also in the far field region (they are just electromagnetic waves). Only in the intermediate region, $R\sim R_0$, do things start getting really complicated (so we do not look in this region!).

3.24 Summary

This marks the end of our theoretical investigation of Maxwell's equations. Let us now summarize what we have learned so far. The field equations which govern electric and magnetic fields are written:

$$\nabla\cdot\boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{3.324a}$$

$$\nabla\cdot\boldsymbol{B} = 0, \tag{3.324b}$$

$$\nabla\wedge\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t}, \tag{3.324c}$$

$$\nabla\wedge\boldsymbol{B} = \mu_0\boldsymbol{j}+\frac{1}{c^2}\frac{\partial\boldsymbol{E}}{\partial t}. \tag{3.324d}$$

These equations can be integrated to give

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int_V \rho\, dV, \tag{3.325a}$$

$$\oint_S \boldsymbol{B}\cdot d\boldsymbol{S} = 0, \tag{3.325b}$$

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\frac{\partial}{\partial t}\int_S \boldsymbol{B}\cdot d\boldsymbol{S}, \tag{3.325c}$$

$$\oint_C \boldsymbol{B}\cdot d\boldsymbol{l} = \mu_0\int_S \boldsymbol{j}\cdot d\boldsymbol{S} + \frac{1}{c^2}\frac{\partial}{\partial t}\int_S \boldsymbol{E}\cdot d\boldsymbol{S}. \tag{3.325d}$$

Equations (3.324b) and (3.324c) are automatically satisfied by writing

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}, \tag{3.326a}$$

$$\boldsymbol{B} = \nabla \wedge \boldsymbol{A}. \tag{3.326b}$$

This prescription is not unique (there are many choices of ϕ and $\boldsymbol{A}$ which generate the same fields) but we can make it unique by adopting the following conventions:

$$\phi(\boldsymbol{r}) \rightarrow 0 \qquad \text{as } |\boldsymbol{r}| \rightarrow \infty, \tag{3.327}$$

and

$$\frac{1}{c^2}\frac{\partial \phi}{\partial t} + \nabla\cdot\boldsymbol{A} = 0. \tag{3.328}$$

Equations (3.324a) and (3.324d) reduce to

$$\Box^2\phi = -\frac{\rho}{\epsilon_0}, \tag{3.329a}$$

$$\Box^2\boldsymbol{A} = -\mu_0\boldsymbol{j} \tag{3.329b}$$

These are driven wave equations of the general form

$$\Box^2 u \equiv \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) u = v. \tag{3.330}$$

The Green's function for this equation which satisfies the boundary conditions and is consistent with causality is

$$G(\boldsymbol{r},\boldsymbol{r}';t,t') = -\frac{1}{4\pi}\frac{\delta(t-t'-|\boldsymbol{r}-\boldsymbol{r}'|/c)}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{3.331}$$

Thus, the solutions to Eqs. (3.329) are

$$\phi(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon_0}\int\frac{[\rho]}{R}\,dV', \tag{3.332a}$$

$$\boldsymbol{A}(\boldsymbol{r},t) = \frac{\mu_0}{4\pi}\int\frac{[\boldsymbol{j}]}{R}\,dV', \tag{3.332b}$$

where $R = |\boldsymbol{r}-\boldsymbol{r}'|$, and $dV' = d^3\boldsymbol{r}'$, with $[A] \equiv A(\boldsymbol{r}', t-R/c)$. These solutions can be combined with Eqs. (3.326) to give

$$\boldsymbol{E}(\boldsymbol{r},t) = \frac{1}{4\pi\epsilon_0}\int\left([\rho]\,\frac{\boldsymbol{R}}{R^3} + \left[\frac{\partial\rho}{\partial t}\right]\frac{\boldsymbol{R}}{cR^2} - \frac{[\partial\boldsymbol{j}/\partial t]}{c^2R}\right)dV', \tag{3.333a}$$

$$\boldsymbol{B}(\boldsymbol{r},t) = \frac{\mu_0}{4\pi}\int\left(\frac{[\boldsymbol{j}]\wedge\boldsymbol{R}}{R^3} + \frac{[\partial\boldsymbol{j}/\partial t]\wedge\boldsymbol{R}}{cR^2}\right)dV'. \tag{3.333b}$$

Equations (3.324)–(3.333) constitute the complete theory of classical electromagnetism. We can express the same information in terms of field equations [Eqs. (3.324)], integrated field equations [Eqs. (3.325)], retarded electromagnetic potentials [Eqs. (3.332)], and retarded electromagnetic fields [Eqs. (3.333)]. Let us now consider the applications of this theory.

4 Applications of Maxwell's equations

4.1 Electrostatic energy

Consider a collection of N static point charges q_i located at position vectors $\boldsymbol{r}_i$ (where i runs from 1 to N). What is the electrostatic energy stored in such a collection? Another way of asking this is, how much work would we have to do in order to assemble the charges, starting from an initial state in which they are all at rest and also very widely separated?

We know that a static electric field is conservative and can consequently be written in terms of a scalar potential:

$$\boldsymbol{E} = -\nabla\phi. \tag{4.1}$$

We also know that the electric force on a charge q is written

$$\boldsymbol{f} = q\,\boldsymbol{E}. \tag{4.2}$$

The work *we* would have to do against electrical forces in order to move the charge from point P to point Q is simply

$$W = -\int_P^Q \boldsymbol{f}\cdot d\boldsymbol{l} = -q\int_P^Q \boldsymbol{E}\cdot d\boldsymbol{l} = q\int_P^Q \nabla\phi\cdot d\boldsymbol{l} = q\left[\phi(Q)-\phi(P)\right]. \tag{4.3}$$

The negative sign in the above expression comes about because we would have to exert a force $-\boldsymbol{f}$ on the charge in order to counteract the force exerted by the electric field. Recall that the scalar potential generated by a point charge q' at position $\boldsymbol{r}'$ is

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\frac{q'}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{4.4}$$

Let us build up our collection of charges one by one. It takes no work to bring the first charge from infinity, since there is no electric field to fight against. Let us clamp this charge in position at $\boldsymbol{r}_1$. In order to bring the second charge into

position at $\boldsymbol{r}_2$ we have to do work against the electric field generated by the first charge. According to Eqs. (4.3) and Eqs. (4.4), this work is given by

$$W_2 = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|}. \tag{4.5}$$

Let us now bring the third charge into position. Since electric fields and scalar potentials are superposable the work done whilst moving the third charge from infinity to $\boldsymbol{r}_3$ is simply the sum of the work done against the electric fields generated by charges 1 and 2 taken in isolation:

$$W_3 = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1 q_3}{|\boldsymbol{r}_1 - \boldsymbol{r}_3|} + \frac{q_2 q_3}{|\boldsymbol{r}_2 - \boldsymbol{r}_3|}\right). \tag{4.6}$$

Thus, the total work done in assembling the three charges is given by

$$W = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1 q_2}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + \frac{q_1 q_3}{|\boldsymbol{r}_1 - \boldsymbol{r}_3|} + \frac{q_2 q_3}{|\boldsymbol{r}_2 - \boldsymbol{r}_3|}\right). \tag{4.7}$$

This result can easily be generalized to N charges:

$$W = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{q_i q_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}. \tag{4.8}$$

The restriction that j must be greater than i makes the above summation rather messy. If we were to sum without restriction (other than $j \neq i$) then each pair of charges would be counted twice. It is convenient to do just this and then to divide the result by two. Thus,

$$W = \frac{1}{2}\frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{q_i q_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}. \tag{4.9}$$

This is the *potential energy* (*i.e.*, the difference between the total energy and the kinetic energy) of a collection of charges. We can think of this as the work needed to bring static charges from infinity and assemble them in the required formation. Alternatively, this is the kinetic energy which would be released if the collection

were dissolved and the charges returned to infinity. But where is this potential energy stored? Let us investigate further.

Equation (4.9) can be written

$$W = \frac{1}{2}\sum_{i=1}^{N} q_i\,\phi_i, \tag{4.10}$$

where

$$\phi_i = \frac{1}{4\pi\epsilon_0}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{q_j}{|\boldsymbol{r}_i-\boldsymbol{r}_j|} \tag{4.11}$$

is the scalar potential experienced by the i th charge due to the other charges in the distribution.

Let us now consider the potential energy of a continuous charge distribution. It is tempting to write

$$W = \frac{1}{2}\int \rho\,\phi\,d^3\boldsymbol{r}, \tag{4.12}$$

by analogy with Eqs. (4.10) and (4.11), where

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}' \tag{4.13}$$

is the familiar scalar potential generated by a continuous charge distribution. Let us try this out. We know from Poisson's equation that

$$\rho = \epsilon_0\nabla\cdot\boldsymbol{E}, \tag{4.14}$$

so Eq. (4.12) can be written

$$W = \frac{\epsilon_0}{2}\int\phi\,\nabla\cdot\boldsymbol{E}\,d^3\boldsymbol{r}. \tag{4.15}$$

Vector field theory yields the standard result

$$\nabla\cdot(\boldsymbol{E}\,\phi) = \phi\,\nabla\cdot\boldsymbol{E} + \boldsymbol{E}\cdot\nabla\phi. \tag{4.16}$$

However, $\nabla\phi = -\boldsymbol{E}$, so we obtain

$$W = \frac{\epsilon_0}{2}\left[\int \nabla\cdot(\boldsymbol{E}\,\phi)\,d^3\boldsymbol{r} + \int E^2\,d^3\boldsymbol{r}\right] \tag{4.17}$$

Application of Gauss' theorem gives

$$W = \frac{\epsilon_0}{2}\left(\oint_S \phi\,\boldsymbol{E}\cdot d\boldsymbol{S} + \int_V E^2\,dV\right), \tag{4.18}$$

where V is some volume which encloses all of the charges and S is its bounding surface. Let us assume that V is a sphere, centred on the origin, and let us take the limit in which radius r of this sphere goes to infinity. We know that, in general, the electric field at large distances from a bounded charge distribution looks like the field of a point charge and, therefore, falls off like $1/r^2$. Likewise, the potential falls off like $1/r$. However, the surface area of the sphere increases like r^2. It is clear that in the limit as $r\rightarrow\infty$ the surface integral in Eq. (4.18) falls off like $1/r$ and is consequently zero. Thus, Eq. (4.18) reduces to

$$W = \frac{\epsilon_0}{2}\int E^2\,d^3\boldsymbol{r}, \tag{4.19}$$

where the integral is over all space. This is a very nice result! It tells us that the potential energy of a continuous charge distribution is stored in the electric field. Of course, we now have to assume that an electric field possesses an *energy density*

$$U = \frac{\epsilon_0}{2}\,E^2. \tag{4.20}$$

We can easily check that Eq. (4.19) is correct. Suppose that we have a charge Q which is uniformly distributed within a sphere of radius a. Let us imagine building up this charge distribution from a succession of thin spherical layers of infinitesimal thickness. At each stage we gather a small amount of charge from infinity and spread it over the surface of the sphere in a thin layer from r to $r+dr$. We continue this process until the final radius of the sphere is a. If $q(r)$ is the charge in the sphere when it has attained radius r, the work done in bringing a charge dq to it is

$$dW = \frac{1}{4\pi\epsilon_0}\frac{q(r)\,dq}{r}. \tag{4.21}$$

This follows from Eq. (4.5) since the electric field generated by a spherical charge distribution (outside itself) is the same as that of a point charge $q(r)$ located at the origin ($r = 0$) (see later). If the constant charge density in the sphere is ρ then

$$q(r) = \frac{4}{3}\pi r^3\,\rho, \tag{4.22}$$

and

$$dq = 4\pi r^2\,\rho\,dr. \tag{4.23}$$

Thus, Eq. (4.21) becomes

$$dW = \frac{4\pi}{3\epsilon_0}\,\rho^2 r^4\,dr. \tag{4.24}$$

The total work needed to build up the sphere from nothing to radius a is plainly

$$W = \frac{4\pi}{3\epsilon_0}\,\rho^2\int_0^a r^4\,dr = \frac{4\pi}{15\epsilon_0}\,\rho^2 a^5. \tag{4.25}$$

This can also be written in terms of the total charge $Q = (4/3)\pi a^3\rho$ as

$$W = \frac{3}{5}\frac{Q^2}{4\pi\epsilon_0 a}. \tag{4.26}$$

Now that we have evaluated the potential energy of a spherical charge distribution by the direct method, let us work it out using Eq. (4.19). We assume that the electric field is radial and spherically symmetric, so $\boldsymbol{E} = E_r(r)\,\hat{\boldsymbol{r}}$. Application of Gauss' law

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int_V \rho\,dV, \tag{4.27}$$

where V is a sphere of radius r, yields

$$E_r(r) = \frac{Q}{4\pi\epsilon_0}\frac{r}{a^3} \tag{4.28}$$

for $r < a$, and

$$E_r(r) = \frac{Q}{4\pi\epsilon_0\,r^2} \tag{4.29}$$

for $r \geq a$. Note that the electric field generated outside the charge distribution is the same as that of a point charge Q located at the origin, $r = 0$. Equations (4.19), (4.28), and (4.29) yield

$$W = \frac{Q^2}{8\pi\epsilon_0}\left(\frac{1}{a^6}\int_0^a r^4\,dr + \int_a^\infty \frac{dr}{r^2}\right), \tag{4.30}$$

which reduces to

$$W = \frac{Q^2}{8\pi\epsilon_0 a}\left(\frac{1}{5}+1\right) = \frac{3}{5}\frac{Q^2}{4\pi\epsilon_0 a}. \tag{4.31}$$

Thus, Eq. (4.19) gives the correct answer.

The reason we have checked Eq. (4.19) so carefully is that on close inspection it is found to be inconsistent with Eq. (4.10), from which it was supposedly derived! For instance, the energy given by Eq. (4.19) is manifestly positive definite, whereas the energy given by Eq. (4.10) can be negative (it is certainly negative for a collection of two point charges of opposite sign). The inconsistency was introduced into our analysis when we replaced Eq. (4.11) by Eq. (4.13). In Eq. (4.11) the self-interaction of the i th charge with its own electric field is excluded whereas it is included in Eq. (4.13). Thus, the potential energies (4.10) and (4.19) are different because in the former we start from ready made point charges whereas in the latter we build up the whole charge distribution from scratch. Thus, if we were to work out the potential energy of a point charge distribution using Eq. (4.19) we would obtain the energy (4.10) *plus* the energy required to assemble the point charges. What is the energy required to assemble a point charge? In fact, it is infinite! To see this let us suppose, for the sake of argument, that our point charges are actually made of charge uniformly distributed over a small sphere of radius a. According to Eq. (4.26) the energy required to assemble the i th point charge is

$$W_i = \frac{3}{5}\frac{{q_i}^2}{4\pi\epsilon_0 a}. \tag{4.32}$$

We can think of this as the self-energy of the i th charge. Thus, we can write

$$W = \frac{\epsilon_0}{2}\int E^2\,d^3\boldsymbol{r} = \frac{1}{2}\sum_{i=1}^N q_i\,\phi_i + \sum_{i=1}^N W_i \tag{4.33}$$

which enables us to reconcile Eqs. (4.10) and (4.19). Unfortunately, if our point charges really are point charges then $a \rightarrow 0$ and the self-energy of each charge becomes infinite. Thus, the potential energies predicted by Eqs. (4.10) and (4.19) differ by an infinite amount. What does this all mean? We have to conclude that the idea of locating electrostatic potential energy in the electric field is inconsistent with the assumption of the existence of point charges. One way out of this difficulty would be to say that all elementary charges, such as electrons, are not points but instead small distributions of charge. Alternatively, we could say that our classical theory of electromagnetism breaks down on very small length-scales due to quantum effects. Unfortunately, the quantum mechanical version of electromagnetism (quantum electrodynamics or QED, for short) suffers from the same infinities in the self-energies of particles as the classical version. There is a prescription, called renormalization, for steering round these infinities and getting finite answers which agree with experiments to extraordinary accuracy. However, nobody really understands why this prescription works. The problem of the infinite self-energies of charged particles is still unresolved.

4.2 Ohm's law

We all know the simplest version of Ohm's law:

$$V = IR, \tag{4.34}$$

where V is the voltage drop across a resistor of resistance R when a current I flows through it. Let us generalize this law so that it is expressed in terms of $\boldsymbol{E}$ and $\boldsymbol{j}$ rather than V and I. Consider a length l of a conductor of uniform cross-sectional area A with a current I flowing down it. In general, we expect the electrical resistance of the conductor to be proportional to its length and inversely proportional to its area (*i.e.*, it is harder to push an electrical current down a long rather than a short wire, and it is easier to push a current down a wide rather than a narrow conducting channel.) Thus, we can write

$$R = \eta \, \frac{l}{A}. \tag{4.35}$$

The constant η is called the *resistivity* and is measured in units of ohm-meters. Ohm's law becomes

$$V = \eta \frac{l}{A} I. \tag{4.36}$$

However, $I/A = j_z$ (suppose that the conductor is aligned along the z-axis) and $V/l = E_z$, so the above equation reduces to

$$E_z = \eta j_z. \tag{4.37}$$

There is nothing special about the z-axis (in an isotropic conducting medium) so the previous formula immediately generalizes to

$$\boldsymbol{E} = \eta \, \boldsymbol{j}. \tag{4.38}$$

This is the vector form of Ohm's law.

A charge q which moves through a voltage drop V acquires an energy qV from the electric field. In a resistor this energy is dissipated as heat. This type of heating is called "ohmic heating." Suppose that N charges per unit time pass through a resistor. The current flowing is obviously $I = Nq$. The total energy gained by the charges, which appears as heat inside the resistor, is

$$P = N \, qV = IV \tag{4.39}$$

per unit time. Thus, the heating power is

$$P = IV = I^2 R = \frac{V^2}{R}. \tag{4.40}$$

Equations (4.39) and (4.40) generalize to

$$P = \boldsymbol{j} \cdot \boldsymbol{E} = \eta \, j^2, \tag{4.41}$$

where P is now the power dissipated per unit volume in a resistive medium.

4.3 Conductors

Most (but not all) electrical conductors obey Ohm's law. Such conductors are termed "ohmic." Suppose that we apply an electric field to an ohmic conductor. What is going to happen? According to Eq. (4.38) the electric field drives

currents. These redistribute the charge inside the conductor until the original electric field is canceled out. At this point, the currents stop flowing. It might be objected that the currents could keep flowing in closed loops. According to Ohm's law, this would require a non-zero e.m.f., $\oint \boldsymbol{E}\cdot d\boldsymbol{l}$, acting around each loop (unless the conductor is a *superconductor*, with $\eta = 0$). However, we know that in steady-state

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = 0 \tag{4.42}$$

around any closed loop C. This proves that a steady-state e.m.f. acting around a closed loop inside a conductor is impossible. The only other alternative is

$$\boldsymbol{j} = \boldsymbol{E} = \boldsymbol{0} \tag{4.43}$$

inside a conductor. It immediately follows from Gauss' law, $\nabla\cdot\boldsymbol{E} = \rho/\epsilon_0$, that

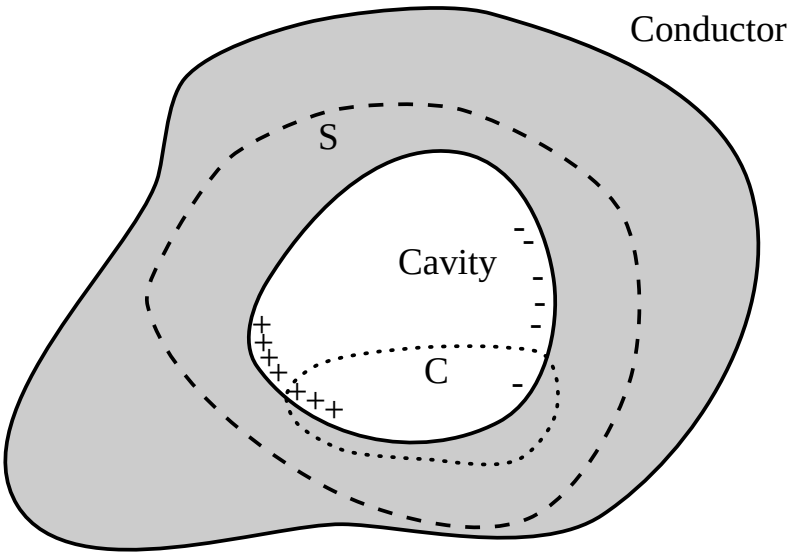
$$\rho = 0. \tag{4.44}$$

So, there are no electric charges in the interior of a conductor. But, how can a conductor cancel out an applied electric field if it contains no charges? The answer is that all of the charges reside on the surface of the conductor. In reality, the charges lie within one or two atomic layers of the surface (see any textbook on solid-state physics). The difference in scalar potential between two points P and Q is simply

$$\phi(Q) - \phi(P) = \int_P^Q \nabla\phi\cdot d\boldsymbol{l} = -\int_P^Q \boldsymbol{E}\cdot d\boldsymbol{l}. \tag{4.45}$$

However, if P and Q lie inside the same conductor then it is clear from Eq. (4.43) that the potential difference between P and Q is zero. This is true no matter where P and Q are situated inside the conductor, so we conclude that the scalar potential must be uniform inside a conductor. A corollary of this is that the surface of a conductor is an equipotential (*i.e.*, $\phi =$ constant) surface.

Not only is the electric field inside a conductor zero. It is also possible to demonstrate that the field within an empty cavity lying inside a conductor is also zero, provided that there are no charges within the cavity. Let us, first of all, integrate Gauss' law over a surface S which surrounds the cavity but lies wholly in



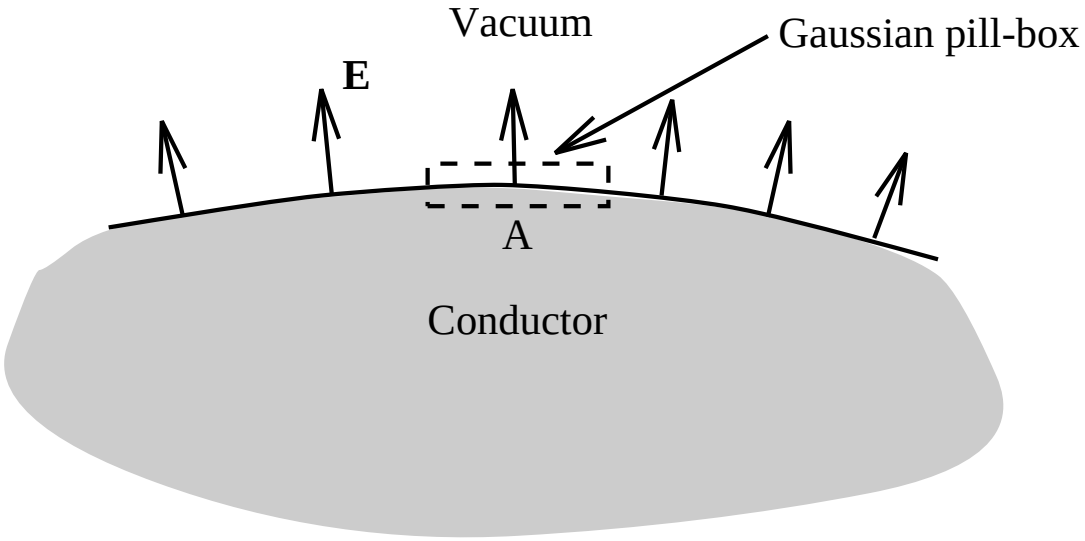
the conducting material. Since the electric field is zero in a conductor, it follows that zero net charge is enclosed by S. This does not preclude the possibility that there are equal amounts of positive and negative charges distributed on the inner surface of the conductor. However, we can easily rule out this possibility using the steady-state relation

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = 0, \tag{4.46}$$

for any closed loop C. If there are any electric field lines inside the cavity then they must run from the positive to the negative surface charges. Consider a loop C which straddles the cavity and the conductor, such as the one shown above. In the presence of field lines it is clear that the line integral of $\boldsymbol{E}$ along that portion of the loop which lies inside the cavity is non-zero. However, the line integral of $\boldsymbol{E}$ along that portion of the loop which runs through the conducting material is obviously zero (since $\boldsymbol{E} = \boldsymbol{0}$ inside a conductor). Thus, the line integral of the field around the closed loop C is non-zero. This, clearly contradicts Eq. (4.46). In fact, this equation implies that the line integral of the electric field along any path which runs through the cavity, from one point on the interior surface of the conductor to another, is zero. This can only be the case if the electric field itself is zero everywhere inside the cavity. There is one proviso to this argument. The electric field inside a cavity is only zero if the cavity contains no charges. If the cavity contains charges then our argument fails because it is possible to envisage

that the line integral of the electric field along many different paths across the cavity could be zero without the fields along these paths necessarily being zero (this argument is somewhat inexact; we shall improve it later on).

We have shown that if a cavity is completely enclosed by a conductor then no stationary distribution of charges outside can ever produce any fields inside. So, we can shield a piece of electrical equipment from stray external electric fields by placing it inside a metal can. Using similar arguments to those given above, we can also show that no static distribution of charges inside a closed conductor can ever produce a field outside it. Clearly, shielding works both ways!

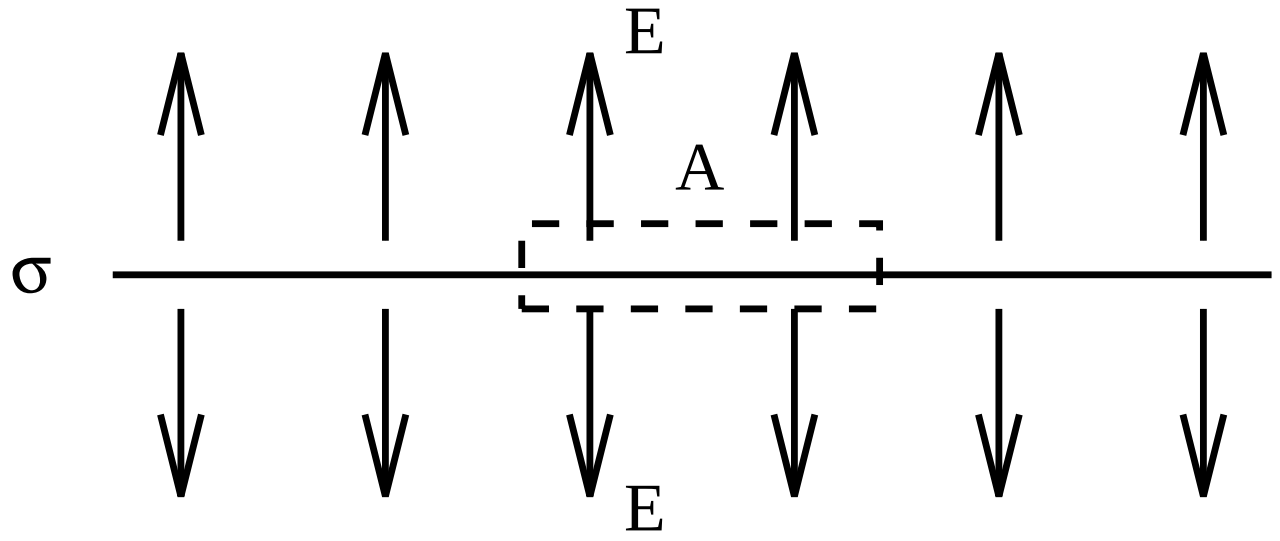


Let us consider some small region on the surface of a conductor. Suppose that the local surface charge density is σ, and that the electric field just outside the conductor is $\boldsymbol{E}$. Note that this field must be directed *normal* to the surface of the conductor. Any parallel component would be shorted out by surface currents. Another way of saying this is that the surface of a conductor is an equipotential. We know that $\nabla\phi$ is always perpendicular to equipotential surfaces, so $\boldsymbol{E} = -\nabla\phi$ must be locally perpendicular to a conducting surface. Let us use Gauss' law,

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int_V \rho\, dV, \tag{4.47}$$

where V is a so-called "Gaussian pill-box." This is a pill-box shaped volume whose two ends are aligned normal to the surface of the conductor, with the

surface running between them, and whose sides are tangential to the surface normal. It is clear that $\boldsymbol{E}$ is perpendicular to the sides of the box, so the sides make no contribution to the surface integral. The end of the box which lies inside the conductor also makes no contribution, since $\boldsymbol{E} = \boldsymbol{0}$ inside a conductor. Thus, the only non-zero contribution to the surface integral comes from the end lying in free space. This contribution is simply $E_\perp\, A$, where $E_\perp$ denotes an outward pointing (from the conductor) normal electric field, and A is the cross-sectional area of the box. The charge enclosed by the box is simply $\sigma\, A$, from the definition



of a surface charge density. Thus, Gauss' law yields

$$E_\perp = \frac{\sigma}{\epsilon_0} \tag{4.48}$$

as the relationship between the normal electric field immediately outside a conductor and the surface charge density.

Let us look at the electric field generated by a sheet charge distribution a little more carefully. Suppose that the charge per unit area is σ. By symmetry, we expect the field generated below the sheet to be the mirror image of that above the sheet (at least, locally). Thus, if we integrate Gauss' law over a pill-box of cross sectional area A, as shown above, then the two ends both contribute $E_{\rm sheet}\, A$ to the surface integral, where $E_{\rm sheet}$ is the normal electric field generated above and below the sheet. The charge enclosed by the pill-box is just $\sigma\, A$. Thus,

Gauss' law yields a symmetric electric field

$$\begin{aligned} E_{\rm sheet} &= +\frac{\sigma}{2\,\epsilon_0} && \text{above,} \\ E_{\rm sheet} &= -\frac{\sigma}{2\,\epsilon_0} && \text{below.} \end{aligned} \tag{4.49}$$

So, how do we get the asymmetric electric field of a conducting surface, which is zero immediately below the surface (*i.e.*, inside the conductor) and non-zero immediately above it? Clearly, we have to add in an external field (*i.e.*, a field which is not generated locally by the sheet charge). The requisite field is

$$E_{\rm ext} = \frac{\sigma}{2\,\epsilon_0} \tag{4.50}$$

both above and below the charge sheet. The total field is the sum of the field generated locally by the charge sheet and the external field. Thus, we obtain

$$\begin{aligned} E_{\rm total} &= +\frac{\sigma}{\epsilon_0} && \text{above,} \\ E_{\rm total} &= 0 && \text{below,} \end{aligned} \tag{4.51}$$

which is in agreement with Eq. (4.48).

The external field exerts a force on the charge sheet. The field generated locally by the sheet itself obviously cannot exert a force (the sheet cannot exert a force on itself!). The force per unit area acting on the surface of the conductor always acts outward and is given by

$$p = \sigma E_{\rm ext} = \frac{\sigma^2}{2\,\epsilon_0}. \tag{4.52}$$

Thus, there is an electrostatic pressure acting on any charged conductor. This effect can be visualized by charging up soap bubbles; the additional electrostatic pressure eventually causes them to burst. The electrostatic pressure can also be written

$$p = \frac{\epsilon_0}{2}\,E^2, \tag{4.53}$$

where E is the field strength immediately above the surface of the conductor. Note that, according to the above formula, the electrostatic pressure is equivalent to

the energy density of the electric field immediately outside the conductor. This is not a coincidence. Suppose that the conductor expands by an average distance dx due to the electrostatic pressure. The electric field is excluded from the region into which the conductor expands. The volume of this region $dV = A\,dx$, where A is the surface area of the conductor. Thus, the energy of the electric field decreases by an amount $dE = U\,dV = (\epsilon_0/2)\,E^2\,dV$, where U is the energy density of the field. This decrease in energy can be ascribed to the work which the field does on the conductor in order to make it expand. This work is $dW = p\,A\,dx$, where p is the force per unit area the field exerts on the conductor. Thus, $dE = dW$, from energy conservation, giving

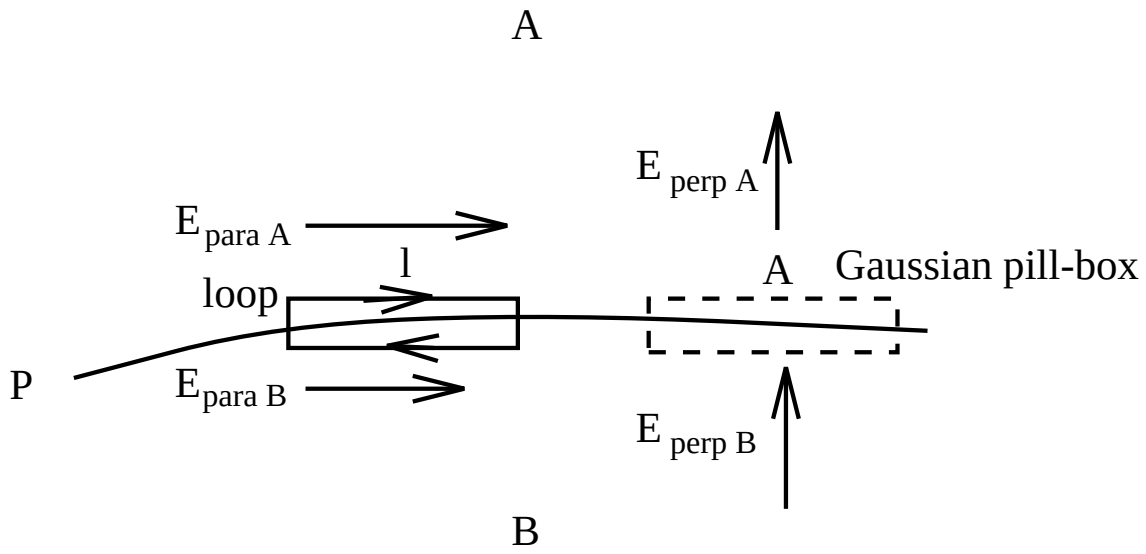
$$p = \frac{\epsilon_0}{2}\,E^2. \tag{4.54}$$

This technique for calculating a force given an expression for the energy of a system as a function of some adjustable parameter is called *the principle of virtual work*, and is very useful.

We have seen that an electric field is excluded from the inside of the conductor, but not from the outside, giving rise to a net *outward* force. We can account for this by saying that the field exerts a *negative* pressure $p = -(\epsilon_0/2)\,E^2$ on the conductor. We know that if we evacuate a metal can then the pressure difference between the inside and the outside eventually causes it to *implode*. Likewise, if we place the can in a strong electric field then the pressure difference between the inside and the outside will eventually cause it to *explode*. How big a field do we need before the electrostatic pressure difference is the same as that obtained by evacuating the can? In other words, what field exerts a negative pressure of one atmosphere (*i.e.*, 10^5 newtons per meter squared) on conductors? The answer is a field of strength $E \sim 10^8$ volts per meter. Fortunately, this is a rather large field, so there is no danger of your car exploding when you turn on the stereo!

4.4 Boundary conditions on the electric field

What are the most general boundary conditions satisfied by the electric field at the interface between two mediums; *e.g.*, the interface between a vacuum and a conductor? Consider an interface P between two mediums A and B. Let us, first



of all, integrate Gauss' law,

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = \frac{1}{\epsilon_0}\int_V \rho\, dV, \tag{4.55}$$

over a Gaussian pill-box S of cross-sectional area A whose two ends are locally parallel to the interface. The ends of the box can be made arbitrarily close together. In this limit, the flux of the electric field out of the sides of the box is obviously negligible. The only contribution to the flux comes from the two ends. In fact,

$$\oint_S \boldsymbol{E}\cdot d\boldsymbol{S} = (E_{\perp\, A} - E_{\perp\, B})\, A, \tag{4.56}$$

where $E_{\perp\, A}$ is the perpendicular (to the interface) electric field in medium A at the interface, *etc.* The charge enclosed by the pill-box is simply $\sigma\, A$, where σ is the sheet charge density on the interface. Note that any volume distribution of charge gives rise to a negligible contribution to the right-hand side of the above equation in the limit where the two ends of the pill-box are very closely spaced. Thus, Gauss' law yields

$$E_{\perp\, A} - E_{\perp\, B} = \frac{\sigma}{\epsilon_0} \tag{4.57}$$

at the interface; *i.e.*, the presence of a charge sheet on an interface causes a discontinuity in the perpendicular component of the electric field. What about

the parallel electric field? Let us integrate Faraday's law,

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = -\frac{\partial}{\partial t}\int_S \boldsymbol{B}\cdot d\boldsymbol{S}, \tag{4.58}$$

around a rectangular loop C whose long sides, length l, run parallel to the interface. The length of the short sides is assumed to be arbitrarily small. The dominant contribution to the loop integral comes from the long sides:

$$\oint_C \boldsymbol{E}\cdot d\boldsymbol{l} = (E_{\parallel\, A} - E_{\parallel\, B})\, l, \tag{4.59}$$

where $E_{\parallel\, A}$ is the parallel (to the interface) electric field in medium A at the interface, *etc.* The flux of the magnetic field through the loop is approximately $B_\perp\, A$, where $B_\perp$ is the component of the magnetic field which is normal to the loop, and A is the area of the loop. But, $A \rightarrow 0$ as the short sides of the loop are shrunk to zero so, unless the magnetic field becomes infinite (we assume that it does not), the flux also tends to zero. Thus,

$$E_{\parallel\, A} - E_{\parallel\, B} = 0; \tag{4.60}$$

i.e., there can be no discontinuity in the parallel component of the electric field across an interface.

4.5 Capacitors

We can store electrical charge on the surface of a conductor. However, electric fields will be generated immediately above the surface. The conductor can only successfully store charge if it is electrically insulated from its surroundings. Air is a very good insulator. Unfortunately, air ceases to be an insulator when the electric field strength through it exceeds some critical value which is about $E_{\rm crit} \sim 10^6$ volts per meter. This phenomenon, which is called "break-down," is associated with the formation of sparks. The most well known example of the break-down of air is during a lightning strike. Clearly, a good charge storing device is one which holds a large amount of charge but only generates small electric fields. Such a device is called a "capacitor."

Consider two thin, parallel, conducting plates of cross-sectional area A which are separated by a *small* distance d (*i.e.*, $d \ll \sqrt{A}$). Suppose that each plate carries an equal and opposite charge Q. We expect this charge to spread evenly over the plates to give an effective sheet charge density $\sigma = \pm Q/A$ on each plate. Suppose that the upper plate carries a positive charge and that the lower plate carries a negative charge. According to Eq. (4.49), the field generated by the upper plate is normal to the plate and of magnitude

$$\begin{aligned} E_{\rm upper} &= +\frac{\sigma}{2\epsilon_0} && \text{above,} \\ E_{\rm upper} &= -\frac{\sigma}{2\epsilon_0} && \text{below.} \end{aligned} \tag{4.61}$$

Likewise, the field generated by the lower plate is

$$\begin{aligned} E_{\rm lower} &= -\frac{\sigma}{2\epsilon_0} && \text{above,} \\ E_{\rm lower} &= +\frac{\sigma}{2\epsilon_0} && \text{below.} \end{aligned} \tag{4.62}$$

Note that we are neglecting any "leakage" of the field at the edges of the plates. This is reasonable if the plates are closely spaced. The total field is the sum of the two fields generated by the upper and lower plates. Thus, the net field is normal to the plates and of magnitude

$$\begin{aligned} E_\perp &= \frac{\sigma}{\epsilon_0} && \text{between,} \\ E_\perp &= 0 && \text{otherwise.} \end{aligned} \tag{4.63}$$

Since the electric field is uniform, the potential difference between the plates is simply

$$V = E_\perp\, d = \frac{\sigma\, d}{\epsilon_0}. \tag{4.64}$$

It is conventional to measure the capacity of a conductor, or set of conductors, to store charge but generate small electric fields in terms of a parameter called the "capacitance." This is usually denoted C. The capacitance of a charge storing

device is simply the ratio of the charge stored to the potential difference generated by the charge. Thus,

$$C = \frac{Q}{V}. \tag{4.65}$$

Clearly, a good charge storing device has a high capacitance. Incidentally, capacitance is measured in coulombs per volt, or farads. This is a rather unwieldy unit since good capacitors typically have capacitances which are only about one millionth of a farad. For a parallel plate capacitor it is clear that

$$C = \frac{\sigma\, A}{V} = \frac{\epsilon_0\, A}{d}. \tag{4.66}$$

Note that the capacitance only depends on geometric quantities such as the area and spacing of the plates. This is a consequence of the superposability of electric fields. If we double the charge on conductors then we double the electric fields generated around them and we, therefore, double the potential difference between the conductors. Thus, the potential difference between conductors is always directly proportional to the charge carried; the constant of proportionality (the inverse of the capacitance) can only depend on geometry.

Suppose that the charge $\pm Q$ on each plate is built up gradually by transferring small amounts of charge from one plate to another. If the instantaneous charge on the plates is $\pm q$ and an infinitesimal amount of positive charge dq is transferred from the negatively charged plate to the positively charge plate then the work done is $dW = V\, dq = q\, dq/C$, where V is the instantaneous voltage difference between the plates. Note that the voltage difference is such that it opposes any increase in the charge on either plate. The total work done in charging the capacitor is

$$W = \frac{1}{C}\int_0^Q q\, dq = \frac{Q^2}{2C} = \frac{1}{2}\, CV^2, \tag{4.67}$$

where use has been made of Eq. (4.65). The energy stored in the capacitor is the same as the work required to charge up the capacitor. Thus,

$$W = \frac{1}{2}\, CV^2. \tag{4.68}$$

This is a general result which holds for all types of capacitor.

The energy of a charged parallel plate capacitor is actually stored in the electric field between the plates. This field is of approximately constant magnitude $E_\perp = V/d$ and occupies a region of volume $A\,d$. Thus, given the energy density of an electric field $U = (\epsilon_0/2)\,E^2$, the energy stored in the electric field is

$$W = \frac{\epsilon_0}{2}\,\frac{V^2}{d^2}\,Ad = \frac{1}{2}\,CV^2, \tag{4.69}$$

where use has been made of Eq. (4.66). Note that Eqs. (4.67) and (4.69) agree. We all know that if we connect a capacitor across the terminals of a battery then a transient current flows as the capacitor charges up. The capacitor can then be placed to one side and, some time later, the stored charge can be used, for instance, to transiently light a bulb in an electrical circuit. What is interesting here is that the energy stored in the capacitor is stored as an electric field, so we can think of a capacitor as a device which either stores energy in, or extracts energy from, an electric field.

The idea, which we discussed earlier, that an electric field exerts a negative pressure $p = -(\epsilon_0/2)\,E^2$ on conductors immediately suggests that the two plates in a parallel plate capacitor *attract* one another with a mutual force

$$F = \frac{\epsilon_0}{2}\,{E_\perp}^2\,A = \frac{1}{2}\,\frac{CV^2}{d}. \tag{4.70}$$

It is not necessary to have two oppositely charged conductors in order to make a capacitor. Consider an isolated sphere of radius a which carries a charge Q. The radial electric field generated outside the sphere is given by

$$E_r(r > a) = \frac{Q}{4\pi\epsilon_0 r^2}. \tag{4.71}$$

The potential difference between the sphere and infinity, or, more realistically, some large, relatively distant reservoir of charge such as the Earth, is

$$V = \frac{Q}{4\pi\epsilon_0 a}. \tag{4.72}$$

Thus, the capacitance of the sphere is

$$C = \frac{Q}{V} = 4\pi\epsilon_0\,a. \tag{4.73}$$

The energy of a sphere when it carries a charge Q is again given by $(1/2)\,C\,V^2$. It can easily be demonstrated that this is really the energy contained in the electric field around the sphere.

Suppose that we have two spheres of radii a and b, respectively, which are connected by an electric wire. The wire allows charge to move back and forth between the spheres until they reach the same potential (with respect to infinity). Let Q be the charge on the first sphere and Q' the charge on the second sphere. Of course, the total charge $Q_0 = Q + Q'$ carried by the two spheres is a conserved quantity. It follows from Eq. (4.72) that

$$\begin{aligned} \frac{Q}{Q_0} &= \frac{a}{a+b}, \\ \frac{Q'}{Q_0} &= \frac{b}{a+b}. \end{aligned} \tag{4.74}$$

Note that if one sphere is much smaller than the other one, *e.g.*, $b \ll a$, then the large sphere grabs most of the charge:

$$\frac{Q}{Q'} = \frac{a}{b} \gg 1. \tag{4.75}$$

The ratio of the electric fields generated just above the surfaces of the two spheres follows from Eqs. (4.71) and (4.75):

$$\frac{E_b}{E_a} = \frac{a}{b}. \tag{4.76}$$

If $b \ll a$ then the field just above the smaller sphere is far bigger than that above the larger sphere. Equation (4.76) is a simple example of a far more general rule. The electric field above some point on the surface of a conductor is inversely proportional to the local radius of curvature of the surface.

It is clear that if we wish to store significant amounts of charge on a conductor then the surface of the conductor must be made as smooth as possible. Any sharp spikes on the surface will inevitably have comparatively small radii of curvature. Intense local electric fields are generated in these regions. These can easily exceed the critical field for the break down of air, leading to sparking and the eventual

loss of the charge on the conductor. Sparking can also be very destructive because the associated electric currents flow through very localized regions giving rise to intense heating.

As a final example, consider two co-axial conducting cylinders of radii a and b, where $a < b$. Suppose that the charge per unit length carried by the inner and outer cylinders is $+q$ and $-q$, respectively. We can safely assume that $\boldsymbol{E} = E_r(r)\,\hat{\boldsymbol{r}}$, by symmetry (adopting standard cylindrical polar coordinates). Let us integrate Gauss' law over a cylinder of radius r, co-axial with the conductors, and of length l. For $a < r < b$ we find that

$$2\pi r l\, E_r(r) = \frac{q l}{\epsilon_0}, \tag{4.77}$$

so

$$E_r = \frac{q}{2\pi\epsilon_0\, r} \tag{4.78}$$

for $a < r < b$. It is fairly obvious that $E_r = 0$ if r is not in the range a to b. The potential difference between the inner and outer cylinders is

$$\begin{aligned} V &= -\int_{\rm outer}^{\rm inner} \boldsymbol{E}\cdot d\boldsymbol{l} = \int_{\rm inner}^{\rm outer} \boldsymbol{E}\cdot d\boldsymbol{l} \\ &= \int_a^b E_r\, dr = \frac{q}{2\pi\epsilon_0}\int_a^b \frac{dr}{r}, \end{aligned} \tag{4.79}$$

so

$$V = \frac{q}{2\pi\epsilon_0}\ln\frac{b}{a}. \tag{4.80}$$

Thus, the capacitance per unit length of the two cylinders is

$$C = \frac{q}{V} = \frac{2\pi\epsilon_0}{\ln b/a}. \tag{4.81}$$

This is a particularly useful result which we shall need later on in this course.

4.6 Poisson's equation

We know that in steady state we can write

$$\boldsymbol{E} = -\nabla\phi, \tag{4.82}$$

with the scalar potential satisfying Poisson's equation

$$\nabla^2\phi = -\frac{\rho}{\epsilon_0}. \tag{4.83}$$

We even know the general solution to this equation:

$$\phi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,d^3\boldsymbol{r}'. \tag{4.84}$$

So, what else is there to say about Poisson's equation? Well, consider a positive (say) point charge in the vicinity of an uncharged, insulated, conducting sphere. The charge attracts negative charges to the near side of the sphere and repels positive charges to the far side. The surface charge distribution induced on the sphere is such that it is maintained at a constant electrical potential. We now have a problem. We cannot use formula (4.84) to work out the potential $\phi(\boldsymbol{r})$ around the sphere, since we do not know how the charges induced on the conducting surface are distributed. The only things which we know about the surface of the sphere are that it is an equipotential and carries zero net charge. Clearly, in the presence of conducting surfaces the solution (4.84) to Poisson's equation is completely useless. Let us now try to develop some techniques for solving Poisson's equation which allow us to solve real problems (which invariably involve conductors).

4.7 The uniqueness theorem

We have already seen the great value of the uniqueness theorem for Poisson's equation (or Laplace's equation) in our discussion of Helmholtz's theorem (see Section 3.10). Let us now examine this theorem in detail.

Consider a volume V bounded by some surface S. Suppose that we are given the charge density ρ throughout V and the value of the scalar potential ϕ_S on S.

Is this sufficient information to uniquely specify the scalar potential throughout V? Suppose, for the sake of argument, that the solution is not unique. Let there be two potentials ϕ_1 and ϕ_2 which satisfy

$$\begin{aligned} \nabla^2\phi_1 &= -\frac{\rho}{\epsilon_0}, \\ \nabla^2\phi_2 &= -\frac{\rho}{\epsilon_0} \end{aligned} \tag{4.85}$$

throughout V, and

$$\begin{aligned} \phi_1 &= \phi_S, \\ \phi_2 &= \phi_S \end{aligned} \tag{4.86}$$

on S. We can form the difference between these two potentials:

$$\phi_3 = \phi_1 - \phi_2. \tag{4.87}$$

The potential ϕ_3 clearly satisfies

$$\nabla^2\phi_3 = 0 \tag{4.88}$$

throughout V, and

$$\phi_3 = 0 \tag{4.89}$$

on S.

According to vector field theory

$$\nabla\cdot(\phi_3\,\nabla\phi_3) = (\nabla\phi_3)^2 + \phi_3\nabla^2\phi_3. \tag{4.90}$$

Thus, using Gauss' theorem

$$\int_V \left\{(\nabla\phi_3)^2 + \phi_3\nabla^2\phi_3\right\} dV = \oint_S \phi_3\nabla\phi_3\cdot d\boldsymbol{S}. \tag{4.91}$$

But, $\nabla^2\phi_3 = 0$ throughout V and $\phi_3 = 0$ on S, so the above equation reduces to

$$\int_V (\nabla\phi_3)^2\, dV = 0. \tag{4.92}$$

Note that $(\nabla\phi_3)^2$ is a *positive definite* quantity. The only way in which the volume integral of a positive definite quantity can be zero is if that quantity itself is zero throughout the volume. This is not necessarily the case for a non-positive definite quantity; we could have positive and negative contributions from various regions inside the volume which cancel one another out. Thus, since $(\nabla\phi_3)^2$ is positive definite it follows that

$$\phi_3 = \text{constant} \tag{4.93}$$

throughout V. However, we know that $\phi_3 = 0$ on S, so we get

$$\phi_3 = 0 \tag{4.94}$$

throughout V. In other words,

$$\phi_1 = \phi_2 \tag{4.95}$$

throughout V and on S. Our initial assumption that ϕ_1 and ϕ_2 are two different solutions of Laplace's equations, satisfying the same boundary conditions, turns out to be incorrect.

The fact that the solutions to Poisson's equation are unique is very useful. It means that if we find a solution to this equation—no matter how contrived the derivation—then this is the only possible solution. One immediate use of the uniqueness theorem is to prove that the electric field inside an empty cavity in a conductor is zero. Recall that our previous proof of this was rather involved, and was also not particularly rigorous (see Section 4.3). We know that the interior surface of the conductor is at some constant potential V, say. So, we have $\phi = V$ on the boundary of the cavity and $\nabla^2\phi = 0$ inside the cavity (since it contains no charges). One rather obvious solution to these equations is $\phi = V$ throughout the cavity. Since the solutions to Poisson's equation are unique this is the *only* solution. Thus,

$$\boldsymbol{E} = -\nabla\phi = -\nabla V = \boldsymbol{0} \tag{4.96}$$

inside the cavity.

Suppose that some volume V contains a number of conductors. We know that the surface of each conductor is an equipotential, but, in general, we do not know what potential each surface is at (unless we are specifically told that it is earthed,

etc.). However, if the conductors are insulated it is plausible that we might know the charge on each conductor. Suppose that there are N conductors, each carrying a charge Q_i ($i = 1$ to N), and suppose that the region V containing these conductors is filled by a known charge density ρ and bounded by some surface S which is either infinity or an enclosing conductor. Is this enough information to uniquely specify the electric field throughout V?

Well, suppose that it is not enough information, so that there are two fields $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ which satisfy

$$\begin{aligned} \nabla \cdot \boldsymbol{E}_1 &= \frac{\rho}{\epsilon_0}, \\ \nabla \cdot \boldsymbol{E}_2 &= \frac{\rho}{\epsilon_0} \end{aligned} \tag{4.97}$$

throughout V, with

$$\begin{aligned} \oint_{S_i} \boldsymbol{E}_1 \cdot d\boldsymbol{S}_i &= \frac{Q_i}{\epsilon_0}, \\ \oint_{S_i} \boldsymbol{E}_2 \cdot d\boldsymbol{S}_i &= \frac{Q_i}{\epsilon_0} \end{aligned} \tag{4.98}$$

on the surface of the ith conductor, and, finally,

$$\begin{aligned} \oint_{S} \boldsymbol{E}_1 \cdot d\boldsymbol{S}_i &= \frac{Q_{\rm total}}{\epsilon_0}, \\ \oint_{S} \boldsymbol{E}_2 \cdot d\boldsymbol{S}_i &= \frac{Q_{\rm total}}{\epsilon_0} \end{aligned} \tag{4.99}$$

over the bounding surface, where

$$Q_{\rm total} = \sum_{i=1}^{N} Q_i + \int_V \rho \, dV \tag{4.100}$$

is the total charge contained in volume V.

Let us form the difference field

$$\boldsymbol{E}_3 = \boldsymbol{E}_1 - \boldsymbol{E}_2. \tag{4.101}$$

It is clear that

$$\nabla \cdot \boldsymbol{E}_3 = 0 \tag{4.102}$$

throughout V, and

$$\oint_{S_i} \boldsymbol{E}_3 \cdot d\boldsymbol{S}_i = 0 \tag{4.103}$$

for all i, with

$$\oint_{S} \boldsymbol{E}_3 \cdot d\boldsymbol{S} = 0. \tag{4.104}$$

Now, we know that each conductor is at a constant potential, so if

$$\boldsymbol{E}_3 = -\nabla \phi_3, \tag{4.105}$$

then ϕ_3 is a constant on the surface of each conductor. Furthermore, if the outer surface S is infinity then $\phi_1 = \phi_2 = \phi_3 = 0$ on this surface. If the outer surface is an enclosing conductor then ϕ_3 is a constant on this surface. Either way, ϕ_3 is constant on S.

Consider the vector identity

$$\nabla \cdot (\phi_3 \, \boldsymbol{E}_3) = \phi_3 \, \nabla \cdot \boldsymbol{E}_3 + \boldsymbol{E}_3 \cdot \nabla \phi_3. \tag{4.106}$$

We have $\nabla \cdot \boldsymbol{E}_3 = 0$ throughout V and $\nabla \phi_3 = -\boldsymbol{E}_3$, so the above identity reduces to

$$\nabla \cdot (\phi_3 \, \boldsymbol{E}_3) = -{E_3}^2 \tag{4.107}$$

throughout V. Integrating over V and making use of Gauss' theorem yields

$$\int_V {E_3}^2 \, dV = -\sum_{i=1}^{N} \oint_{S_i} \phi_3 \, \boldsymbol{E}_3 \cdot d\boldsymbol{S}_i - \oint_S \phi_3 \, \boldsymbol{E}_3 \cdot d\boldsymbol{S}. \tag{4.108}$$

However, ϕ_3 is a constant on the surfaces S_i and S. So, making use of Eqs. (4.103) and (4.104), we obtain

$$\int_V {E_3}^2 \, dV = 0. \tag{4.109}$$

Of course, ${E_3}^2$ is a positive definite quantity, so the above relation implies that

$$\boldsymbol{E}_3 = \boldsymbol{0} \tag{4.110}$$

throughout V; *i.e.*, the fields $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are identical throughout V.

It is clear that, for a general electrostatic problem involving charges and conductors, if we are given either the potential at the surface of each conductor or the charge carried by each conductor (plus the charge density throughout the volume, *etc.*) then we can uniquely determine the electric field. There are many other uniqueness theorems which generalize this result still further; *i.e.*, we could be given the potential of some of the conductors and the charge carried by the others and the solution would still be unique.

4.8 The classical image problem

So, how do we actually solve Poisson's equation,

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = -\frac{\rho(x,y,z)}{\epsilon_0}, \tag{4.111}$$

in practice? In general, the answer is that we use a computer. However, there are a few situations, possessing a high degree of symmetry, where it is possible to find analytic solutions. Let us discuss some of these solutions.

Suppose that we have a point charge q held a distance d from an infinite, grounded, conducting plate. Let the plate lie in the x-y plane and suppose that the point charge is located at coordinates (0, 0, d). What is the scalar potential above the plane? This is not a simple question because the point charge induces surface charges on the plate, and we do not know how these are distributed.

What do we know in this problem? We know that the conducting plate is an equipotential. In fact, the potential of the plate is zero, since it is grounded. We also know that the potential at infinity is zero (this is our usual boundary condition for the scalar potential). Thus, we need to solve Poisson's equation in the region $z > 0$, for a single point charge q at position (0, 0, d), subject to the

boundary conditions

$$\phi(z=0)=0, \tag{4.112}$$

and

$$\phi \rightarrow \infty \tag{4.113}$$

as $x^2+y^2+z^2 \rightarrow \infty$. Let us forget about the real problem, for a minute, and concentrate on a slightly different one. We refer to this as the analogue problem. In the analogue problem we have a charge q located at (0, 0, d) and a charge $-q$ located at (0, 0, -d) with no conductors present. We can easily find the scalar potential for this problem, since we know where all the charges are located. We get

$$\phi_{\rm analogue}(x,y,z)=\frac{1}{4\pi\epsilon_0}\left\{\frac{q}{\sqrt{x^2+y^2+(z-d)^2}}-\frac{q}{\sqrt{x^2+y^2+(z+d)^2}}\right\}. \tag{4.114}$$

Note, however, that

$$\phi_{\rm analogue}(z=0)=0, \tag{4.115}$$

and

$$\phi_{\rm analogue} \rightarrow 0 \tag{4.116}$$

as $x^2+y^2+z^2 \rightarrow \infty$. In addition, $\phi_{\rm analogue}$ satisfies Poisson's equation for a charge at (0, 0, d), in the region $z>0$. Thus, $\phi_{\rm analogue}$ is a solution to the problem posed earlier, in the region $z>0$. Now, the uniqueness theorem tells us that there is only *one* solution to Poisson's equation which satisfies a given, well-posed set of boundary conditions. So, $\phi_{\rm analogue}$ must be the correct potential in the region $z>0$. Of course, $\phi_{\rm analogue}$ is completely wrong in the region $z<0$. We know this because the grounded plate shields the region $z<0$ from the point charge, so that $\phi=0$ in this region. Note that we are leaning pretty heavily on the uniqueness theorem here! Without this theorem, it would be hard to convince a skeptical person that $\phi=\phi_{\rm analogue}$ is the correct solution in the region $z>0$.

Now that we know the potential in the region $z>0$, we can easily work out the distribution of charges induced on the conducting plate. We already know that the relation between the electric field immediately above a conducting surface and the density of charge on the surface is

$$E_\perp=\frac{\sigma}{\epsilon_0}. \tag{4.117}$$

In this case,

$$E_\perp = E_z(z=0_+) = -\frac{\partial\phi(z=0_+)}{\partial z} = -\frac{\partial\phi_{\rm analogue}(z=0_+)}{\partial z}, \tag{4.118}$$

so

$$\sigma = -\epsilon_0 \frac{\partial\phi_{\rm analogue}(z=0_+)}{\partial z}. \tag{4.119}$$

It follows from Eq. (4.114) that

$$\frac{\partial\phi}{\partial z} = \frac{q}{4\pi\epsilon_0}\left\{\frac{-(z-d)}{[x^2+y^2+(z-d)^2]^{3/2}} + \frac{(z+d)}{[x^2+y^2+(z+d)^2]^{3/2}}\right\}, \tag{4.120}$$

so

$$\sigma(x,y) = -\frac{qd}{2\pi(x^2+y^2+d^2)^{3/2}}. \tag{4.121}$$

Clearly, the charge induced on the plate has the opposite sign to the point charge. The charge density on the plate is also symmetric about the z-axis, and is largest where the plate is closest to the point charge. The total charge induced on the plate is

$$Q = \int_{x-y\,\rm plane} \sigma\, dS, \tag{4.122}$$

which yields

$$Q = -\frac{qd}{2\pi}\int_0^\infty \frac{2\pi r\, dr}{(r^2+d^2)^{3/2}}, \tag{4.123}$$

where $r^2 = x^2+y^2$. Thus,

$$Q = -\frac{qd}{2}\int_0^\infty \frac{dk}{(k+d^2)^{3/2}} = qd\left[\frac{1}{(k+d^2)^{1/2}}\right]_0^\infty = -q. \tag{4.124}$$

So, the total charge induced on the plate is equal and opposite to the point charge which induces it.

Our point charge induces charges of the opposite sign on the conducting plate. This, presumably, gives rise to a force of attraction between the charge and the plate. What is this force? Well, since the potential, and, hence, the electric field, in the vicinity of the point charge is the same as in the analogue problem then

the force on the charge must be the same as well. In the analogue problem there are two charges $\pm q$ a net distance $2d$ apart. The force on the charge at position (0, 0, d) (*i.e.*, the real charge) is

$$\boldsymbol{F} = -\frac{1}{4\pi\epsilon_0}\frac{q^2}{(2d)^2}\,\hat{\boldsymbol{z}}. \tag{4.125}$$

What, finally, is the potential energy of the system. For the analogue problem this is just

$$W_{\rm analogue} = -\frac{1}{4\pi\epsilon_0}\frac{q^2}{2d}. \tag{4.126}$$

Note that the fields on opposite sides of the conducting plate are mirror images of one another in the analogue problem. So are the charges (apart from the change in sign). This is why the technique of replacing conducting surfaces by imaginary charges is called "the method of images." We know that the potential energy of a set of charges is equivalent to the energy stored in the electric field. Thus,

$$W = \frac{\epsilon_0}{2}\int_{\rm all\,space} E^2\,dV. \tag{4.127}$$

In the analogue problem the fields on either side of the x-y plane are mirror images of one another, so $E^2(x,y,z) = E^2(x,y,-z)$. It follows that

$$W_{\rm analogue} = 2\,\frac{\epsilon_0}{2}\int_{z>0} E^2_{\rm analogue}\,dV. \tag{4.128}$$

In the real problem

$$\begin{aligned} \boldsymbol{E}(z>0) &= \boldsymbol{E}_{\rm analogue}(z>0), \\ \boldsymbol{E}(z<0) &= \boldsymbol{0}. \end{aligned} \tag{4.129}$$

So,

$$W = \frac{\epsilon_0}{2}\int_{z>0} E^2\,dV = \frac{\epsilon_0}{2}\int_{z>0} E^2_{\rm analogue}\,dV = \frac{1}{2}W_{\rm analogue}, \tag{4.130}$$

giving

$$W = -\frac{1}{4\pi\epsilon_0}\frac{q^2}{4d}. \tag{4.131}$$

There is another method by which we can obtain the above result. Suppose that the charge is gradually moved towards the plate along the z-axis from infinity until it reaches position (0, 0, d). How much work is required to achieve this? We know that the force of attraction acting on the charge is

$$F_z = -\frac{1}{4\pi\epsilon_0}\frac{q^2}{4z^2}. \tag{4.132}$$

Thus, the work required to move this charge by dz is

$$dW = -F_z\,dz = \frac{1}{4\pi\epsilon_0}\frac{q^2}{4z^2}\,dz. \tag{4.133}$$

The total work needed to move the charge from $z = \infty$ to $z = d$ is

$$W = \frac{1}{4\pi\epsilon_0}\int_\infty^d \frac{q^2}{4z^2}\,dz = \frac{1}{4\pi\epsilon_0}\left[-\frac{q^2}{4z}\right]_\infty^d = -\frac{1}{4\pi\epsilon_0}\frac{q^2}{4d}. \tag{4.134}$$

Of course, this work is equivalent to the potential energy we evaluated earlier, and is, in turn, the same as the energy contained in the electric field.

There are many different image problems, each of which involves replacing a conductor (*e.g.*, a sphere) with an imaginary charge (or charges) which mimics the electric field in some region (but not everywhere). Unfortunately, we do not have time to discuss any more of these problems.

4.9 Complex analysis

Let us now investigate another "trick" for solving Poisson's equation (actually it only solves Laplace's equation). Unfortunately, this method can only be applied in two dimensions.

The complex variable is conventionally written

$$z = x + \mathrm{i}\,y \tag{4.135}$$

(z should not be confused with a z-coordinate; this is a strictly two dimensional problem). We can write functions $F(z)$ of the complex variable just like we would

write functions of a real variable. For instance,

$$\begin{aligned} F(z) &= z^2, \\ F(z) &= \frac{1}{z}. \end{aligned} \tag{4.136}$$

For a given function $F(z)$ we can substitute $z = x + \mathrm{i}\, y$ and write

$$F(z) = U(x, y) + \mathrm{i}\, V(x, y), \tag{4.137}$$

where U and V are two *real* two dimensional functions. Thus, if

$$F(z) = z^2, \tag{4.138}$$

then

$$F(x + \mathrm{i}\, y) = (x + \mathrm{i}\, y)^2 = (x^2 - y^2) + 2\,\mathrm{i}\, xy, \tag{4.139}$$

giving

$$\begin{aligned} U(x, y) &= x^2 - y^2, \\ V(x, y) &= 2xy. \end{aligned} \tag{4.140}$$

We can define the derivative of a complex function in just the same manner as we would define the derivative of a real function. Thus,

$$\frac{dF}{dz} = \lim_{|\delta z| \to \infty} \frac{F(z + \delta z) - F(z)}{\delta z}. \tag{4.141}$$

However, we now have a slight problem. If $F(z)$ is a "well defined" function (we shall leave it to the mathematicians to specify exactly what being well defined entails: suffice to say that most functions we can think of are well defined) then it should not matter from which direction in the complex plane we approach z when taking the limit in Eq. (4.141). There are, of course, many different directions we could approach z from, but if we look at a regular complex function, $F(z) = z^2$, say, then

$$\frac{dF}{dz} = 2z \tag{4.142}$$

is perfectly well defined and is, therefore, completely independent of the details of how the limit is taken in Eq. (4.141).

The fact that Eq. (4.141) has to give the same result, no matter which path we approach z from, means that there are some restrictions on the functions U and V in Eq. (4.137). Suppose that we approach z along the real axis, so that $\delta z = \delta x$. Then,

$$\begin{aligned}\frac{dF}{dz} &= \lim_{|\delta x|\rightarrow 0} \frac{U(x+\delta x,y)+\mathrm{i}\,V(x+\delta x,y)-U(x,y)-\mathrm{i}\,V(x,y)}{\delta x}\\ &= \frac{\partial U}{\partial x}+\mathrm{i}\,\frac{\partial V}{\partial x}.\end{aligned} \tag{4.143}$$

Suppose that we now approach z along the imaginary axis, so that $\delta z = \mathrm{i}\,\delta y$. Then,

$$\begin{aligned}\frac{dF}{dz} &= \lim_{|\delta y|\rightarrow 0} \frac{U(x,y+\delta y)+\mathrm{i}\,V(x,y+\delta y)-U(x,y)-\mathrm{i}\,V(x,y)}{\mathrm{i}\,\delta y}\\ &= -\mathrm{i}\,\frac{\partial U}{\partial y}+\frac{\partial V}{\partial y}.\end{aligned} \tag{4.144}$$

If $F(z)$ is a well defined function then its derivative must also be well defined, which implies that the above two expressions are equivalent. This requires that

$$\begin{aligned}\frac{\partial U}{\partial x} &= \frac{\partial V}{\partial y},\\ \frac{\partial V}{\partial x} &= -\frac{\partial U}{\partial y}.\end{aligned} \tag{4.145}$$

These are called the Cauchy-Riemann equations and are, in fact, sufficient to ensure that all possible ways of taking the limit (4.141) give the same answer.

So far, we have found that a general complex function $F(z)$ can be written

$$F(z) = U(x,y)+\mathrm{i}\,V(x,y), \tag{4.146}$$

where $z = x+\mathrm{i}\,y$. If $F(z)$ is well defined then U and V *automatically* satisfy the Cauchy-Riemann equations:

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y},$$

$$\frac{\partial V}{\partial x} = -\frac{\partial U}{\partial y}. \tag{4.147}$$

But, what has all of this got to do with electrostatics? Well, we can combine the two Cauchy-Riemann relations. We get

$$\frac{\partial^2 U}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial V}{\partial y} = \frac{\partial}{\partial y}\frac{\partial V}{\partial x} = -\frac{\partial}{\partial y}\frac{\partial U}{\partial y}, \tag{4.148}$$

and

$$\frac{\partial^2 V}{\partial x^2} = -\frac{\partial}{\partial x}\frac{\partial U}{\partial y} = -\frac{\partial}{\partial y}\frac{\partial U}{\partial x} = -\frac{\partial}{\partial y}\frac{\partial V}{\partial y}, \tag{4.149}$$

which reduce to

$$\begin{aligned} \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} &= 0, \\ \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} &= 0. \end{aligned} \tag{4.150}$$

Thus, both U and V *automatically* satisfy Laplace's equation in two dimensions; *i.e.*, both U and V are possible two dimensional scalar potentials in free space.

Consider the two dimensional gradients of U and V:

$$\begin{aligned} \nabla U &= \left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}\right), \\ \nabla V &= \left(\frac{\partial V}{\partial x}, \frac{\partial V}{\partial y}\right). \end{aligned} \tag{4.151}$$

Now

$$\nabla U \cdot \nabla V = \frac{\partial U}{\partial x}\frac{\partial V}{\partial x} + \frac{\partial U}{\partial y}\frac{\partial V}{\partial y}. \tag{4.152}$$

It follows from the Cauchy-Riemann equations that

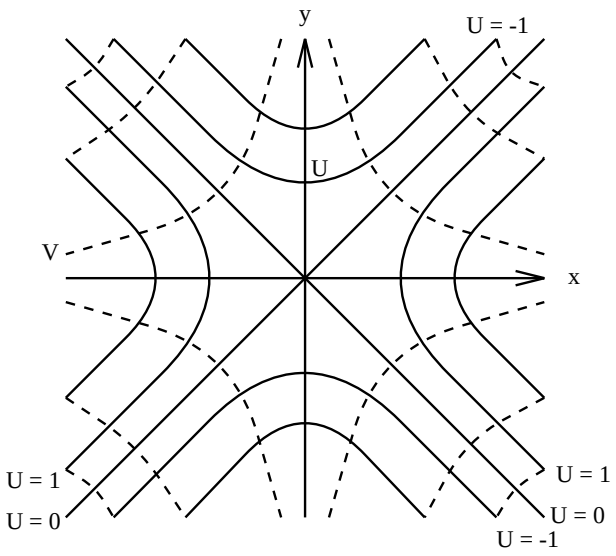
$$\nabla U \cdot \nabla V = \frac{\partial V}{\partial y}\frac{\partial V}{\partial x} - \frac{\partial V}{\partial x}\frac{\partial V}{\partial y} = 0. \tag{4.153}$$

Thus, the contours of U are everywhere perpendicular to the contours of V. It follows that if U maps out the contours of some free space scalar potential then V indicates the directions of the associated electric field lines, and *vice versa*.

For every well defined complex function $F(z)$ we can think of, we get two sets of free space potentials and the associated electric field lines. For example, consider the function $F(z) = z^2$, for which

$$\begin{aligned} U &= x^2 - y^2, \\ V &= 2xy. \end{aligned} \tag{4.154}$$

These are, in fact, the equations of two sets of orthogonal hyperboloids. So,



$U(x, y)$ (the solid lines in the figure) might represent the contours of some scalar potential and $V(x, y)$ (the dashed lines in the figure) the associated electric field lines, or *vice versa*. But, how could we actually generate a hyperboloidal potential? This is easy. Consider the contours of U at level ± 1. These could represent the surfaces of four hyperboloid conductors maintained at potentials $\pm\mathcal{V}$. The scalar potential in the region between these conductors is given by $\mathcal{V}\,U(x, y)$ and the associated electric field lines follow the contours of $V(x, y)$. Note that

$$E_x = -\frac{\partial \phi}{\partial x} = -\mathcal{V}\,\frac{\partial U}{\partial x} = -2\mathcal{V}x \tag{4.155}$$

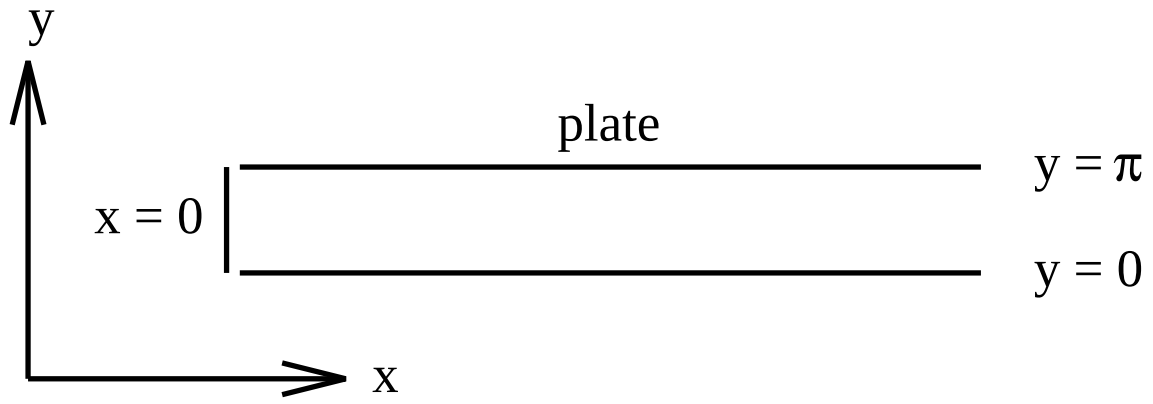
Thus, the x-component of the electric field is directly proportional to the distance from the x-axis. Likewise, for the y-component of the field. This property can be exploited to make devices (called quadrupole electrostatic lenses) which are useful for focusing particle beams.

We can think of the set of all possible well defined complex functions as a reference library of solutions to Laplace's equation in two dimensions. We have only considered a single example but there are, of course, very many complex functions which generate interesting potentials. For instance, $F(z) = z^{1/2}$ generates the potential around a semi-infinite, thin, grounded, conducting plate placed in an external field, whereas $F(z) = z^{3/2}$ yields the potential outside a grounded rectangular conducting corner under similar circumstances.

4.10 Separation of variables

The method of images and complex analysis are two rather elegant techniques for solving Poisson's equation. Unfortunately, they both have an extremely limited range of application. The final technique we shall discuss in this course, namely, the separation of variables, is somewhat messy but possess a far wider range of application. Let us examine a specific example.

Consider two semi-infinite, grounded, conducting plates lying parallel to the x-z plane, one at $y = 0$, and the other at $y = \pi$. The left end, at $x = 0$, is closed off by an infinite strip insulated from the two plates and maintained at a specified potential $\phi_0(y)$. What is the potential in the region between the plates?



We first of all assume that the potential is z-independent, since everything else in the problem is. This reduces the problem to two dimensions. Poisson's equation is written

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=0 \tag{4.156}$$

in the vacuum region between the conductors. The boundary conditions are

$$\phi(x,0) = 0, \tag{4.157a}$$

$$\phi(x,\pi) = 0 \tag{4.157b}$$

for $x>0$, since the two plates are earthed, plus

$$\phi(0,y)=\phi_0(y) \tag{4.158}$$

for $0\leq y\leq\pi$, and

$$\phi(x,y)\rightarrow 0 \tag{4.159}$$

as $x\rightarrow\infty$. The latter boundary condition is our usual one for the scalar potential at infinity.

The central assumption in the method of separation of variables is that a multi-dimensional potential can be written as the product of one-dimensional potentials, so that

$$\phi(x,y)=X(x)Y(y). \tag{4.160}$$

The above solution is obviously a very special one, and is, therefore, only likely to satisfy a very small subset of possible boundary conditions. However, it turns out that by adding together lots of different solutions of this form we can match to general boundary conditions.

Substituting (4.160) into (4.156), we obtain

$$Y\frac{d^2Y}{dx^2}+X\frac{d^2Y}{dy^2}=0. \tag{4.161}$$

Let us now separate the variables; *i.e.*, let us collect all of the x-dependent terms on one side of the equation, and all of the y-dependent terms on the other side. Thus,

$$\frac{1}{X}\frac{d^2X}{dx^2}=-\frac{1}{Y}\frac{d^2Y}{dy^2}. \tag{4.162}$$

This equation has the form

$$f(x) = g(y), \tag{4.163}$$

where f and g are general functions. The only way in which the above equation can be satisfied, for general x and y, is if both sides are equal to the same constant. Thus,

$$\frac{1}{X}\frac{d^2X}{dx^2} = k^2 = -\frac{1}{Y}\frac{d^2Y}{dy^2}. \tag{4.164}$$

The reason why we write k^2, rather than $-k^2$, will become apparent later on. Equation (4.164) separates into two ordinary differential equations:

$$\begin{aligned} \frac{d^2X}{dx^2} &= k^2\, X, \\ \frac{d^2Y}{dy^2} &= -k^2\, Y. \end{aligned} \tag{4.165}$$

We know the general solution to these equations:

$$\begin{aligned} X &= A\exp(kx) + B\exp(-kx), \\ Y &= C\sin ky + D\cos ky, \end{aligned} \tag{4.166}$$

giving

$$\phi = (\, A\exp(kx) + B\exp(-kx)\,)(C\sin ky + D\cos ky). \tag{4.167}$$

Here, A, B, C, and D are arbitrary constants. The boundary condition (4.159) is automatically satisfied if $A = 0$ and $k > 0$. Note that the choice k^2, instead of $-k^2$, in Eq. (4.164) facilitates this by making ϕ either grow or decay monotonically in the x-direction instead of oscillating. The boundary condition (4.157a) is automatically satisfied if $D = 0$. The boundary condition (4.157b) is satisfied provided that

$$\sin k\pi = 0, \tag{4.168}$$

which implies that k is a positive integer, n (say). So, our solution reduces to

$$\phi(x, y) = C\exp(-nx)\sin ny, \tag{4.169}$$

where B has been absorbed into C. Note that this solution is only able to satisfy the final boundary condition (4.158) provided $\phi_0(y)$ is proportional to $\sin ny$. Thus, at first sight, it would appear that the method of separation of variables only works for a very special subset of boundary conditions. However, this is not the case.

Now comes the clever bit! Since Poisson's equation is *linear*, any linear combination of solutions is also a solution. We can therefore form a more general solution than (4.169) by adding together lots of solutions involving different values of n. Thus,

$$\phi(x,y) = \sum_{n=1}^{\infty} C_n \exp(-nx) \sin ny, \tag{4.170}$$

where the C_n are constants. This solution automatically satisfies the boundary conditions (4.157) and (4.159). The final boundary condition (4.158) reduces to

$$\phi(0,y) = \sum_{n=1}^{\infty} C_n \sin ny = \phi_0(y). \tag{4.171}$$

The question now is what choice of the C_n fits an arbitrary function $\phi_0(y)$? To answer this question we can make use of two very useful properties of the functions $\sin ny$. Namely, that they are mutually *orthogonal* and form a *complete set*. The orthogonality property of these functions manifests itself through the relation

$$\int_0^{\pi} \sin ny \, \sin n'y \, dy = \frac{\pi}{2} \, \delta_{nn'}, \tag{4.172}$$

where the function $\delta_{nn'} = 1$ if $n = n'$ and 0 otherwise is called a Kroenecker delta. The completeness property of sine functions means that any general function $\phi_0(y)$ can always be adequately represented as a weighted sum of sine functions with various different n values. Multiplying both sides of Eq. (4.171) by $\sin n'y$ and integrating over y we obtain

$$\sum_{n=1}^{\infty} C_n \int_0^{\pi} \sin ny \, \sin n'y \, dy = \int_0^{\pi} \phi_0(y) \, \sin n'y \, dy. \tag{4.173}$$

The orthogonality relation yields

$$\frac{\pi}{2}\sum_{n=1}^{\infty} C_n\,\delta_{nn'} = \frac{\pi}{2}\,C_{n'} = \int_0^{\pi} \phi_0(y)\,\sin n'y\;dy, \tag{4.174}$$

so

$$C_n = \frac{2}{\pi}\int_0^{\pi} \phi_0(y)\,\sin ny\,dy. \tag{4.175}$$

Thus, we now have a general solution to the problem for any driving potential $\phi_0(y)$.

If the potential $\phi_0(y)$ is constant then

$$C_n = \frac{2\,\phi_0}{\pi}\int_0^{\pi} \sin ny\;dy = \frac{2\,\phi_0}{n\,\pi}\,(1-\cos n\pi), \tag{4.176}$$

giving

$$C_n = 0 \tag{4.177}$$

for even n, and

$$C_n = \frac{4\,\phi_0}{n\,\pi} \tag{4.178}$$

for odd n. Thus,

$$\phi(x,y) = \frac{4\,\phi_0}{\pi}\sum_{n=1,3,5} \frac{\exp(-nx)\sin nx}{n}. \tag{4.179}$$

This potential can be summed explicitly to give

$$\phi(x,y) = \frac{2\phi_0}{\pi}\tan^{-1}\left(\frac{\sin y}{\sinh x}\right). \tag{4.180}$$

In this form it is easy to check that Poisson's equation is obeyed and that all of the boundary conditions are satisfied.

In the above problem we write the potential as the product of one dimensional functions. Some of these functions grow and decay monotonically (*i.e.*, the exponential functions) and the others oscillate (*i.e.*, the sinusoidal functions). The

success of the method depends crucially on the orthogonality and completeness of the oscillatory functions. A set of functions $f_n(x)$ is *orthogonal* if the integral of the product of two different members of the set over some range is always zero:

$$\int_a^b f_n(x) f_m(x)\, dx = 0, \tag{4.181}$$

for $n \neq m$. A set of functions is *complete* if any other function can be expanded as a weighted sum of them. It turns out that the scheme set out above can be generalized to more complicated geometries. For instance, in spherical geometry the monotonic functions are power law functions of the radial variable and the oscillatory functions are Legendre polynomials. The latter are both mutually orthogonal and form a complete set. There are also cylindrical, ellipsoidal, hyperbolic, toroidal, *etc.* coordinates. In all cases, the associated oscillating functions are mutually orthogonal and form a complete set. This implies that the method of separation of variables is of quite general applicability.

4.11 Inductors

We have now completed our investigation of electrostatics. We should now move on to magnetostatics—*i.e.*, the study of steady magnetic fields generated by steady currents. Let us skip this topic. It contains nothing new (it merely consists of the application of Ampère's law and the Biot-Savart law) and is also exceptionally dull!

We have learned about resistance and capacitance. Let us now investigate inductance. Electrical engineers like to reduce all pieces of electrical apparatus to an *equivalent circuit* consisting only of e.m.f. sources (*e.g.*, batteries), inductors, capacitors, and resistors. Clearly, once we understand inductors we shall be ready to apply the laws of electromagnetism to real life situations.

Consider two stationary loops of wire, labeled 1 and 2. Let us run a steady current I_1 around the first loop to produce a field $\boldsymbol{B}_1$. Some of the field lines of $\boldsymbol{B}_1$ will pass through the second loop. Let Φ_2 be the flux of $\boldsymbol{B}_1$ through loop 2:

$$\Phi_2 = \int_{\text{loop 2}} \boldsymbol{B}_1 \cdot d\boldsymbol{S}_2, \tag{4.182}$$

where $d\boldsymbol{S}_2$ is a surface element of loop 2. This flux is generally quite difficult to calculate exactly (unless the two loops have a particularly simple geometry). However, we can infer from the Biot-Savart law,

$$\boldsymbol{B}_1(\boldsymbol{r}) = \frac{\mu_0 I_1}{4\pi} \oint_{\text{loop 1}} \frac{d\boldsymbol{l}_1 \wedge (\boldsymbol{r} - \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|^3}, \tag{4.183}$$

that the magnitude of $\boldsymbol{B}_1$ is proportional to the current I_1. This is ultimately a consequence of the linearity of Maxwell's equations. Here, $d\boldsymbol{l}_1$ is a line element of loop 1 located at position vector $\boldsymbol{r}'$. It follows that the flux Φ_2 must also be proportional to I_1. Thus, we can write

$$\Phi_2 = M_{21}\, I_1, \tag{4.184}$$

where M_{21} is the constant of proportionality. This constant is called the *mutual inductance* of the two loops.

Let us write the field $\boldsymbol{B}_1$ in terms of a vector potential $\boldsymbol{A}_1$, so that

$$\boldsymbol{B}_1 = \nabla \wedge \boldsymbol{A}_1. \tag{4.185}$$

It follows from Stokes' theorem that

$$\Phi_2 = \int_{\text{loop 2}} \boldsymbol{B}_1 \cdot d\boldsymbol{S}_2 = \int_{\text{loop 2}} \nabla \wedge \boldsymbol{A}_1 \cdot d\boldsymbol{S}_2 = \oint_{\text{loop 2}} \boldsymbol{A}_1 \cdot d\boldsymbol{l}_2, \tag{4.186}$$

where $d\boldsymbol{l}_2$ is a line element of loop 2. However, we know that

$$\boldsymbol{A}_1(\boldsymbol{r}) = \frac{\mu_0 I_1}{4\pi} \oint_{\text{loop 1}} \frac{d\boldsymbol{l}_1}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{4.187}$$

The above equation is just a special case of the more general law,

$$\boldsymbol{A}_1(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int_{\text{all space}} \frac{\boldsymbol{j}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}', \tag{4.188}$$

for $\boldsymbol{j}(\boldsymbol{r}') = d\boldsymbol{l}_1 I_1 / dl_1 dA$ and $d^3\boldsymbol{r}' = dl_1 dA$, where dA is the cross sectional area of loop 1. Thus,

$$\Phi_2 = \frac{\mu_0 I_1}{4\pi} \oint_{\text{loop 1}} \oint_{\text{loop 2}} \frac{d\boldsymbol{l}_1 \cdot d\boldsymbol{l}_2}{|\boldsymbol{r} - \boldsymbol{r}'|}, \tag{4.189}$$

where $\boldsymbol{r}$ is now the position vector of the line element $d\boldsymbol{l}_2$ of loop 2. The above equation implies that

$$M_{21} = \frac{\mu_0}{4\pi} \oint_{\rm loop\,1} \oint_{\rm loop\,2} \frac{d\boldsymbol{l}_1 \cdot d\boldsymbol{l}_2}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{4.190}$$

In fact, mutual inductances are rarely worked out from first principles—it is usually too difficult. However, the above formula tells us two important things. Firstly, the mutual inductance of two loops is a purely geometric quantity, having to do with the sizes, shapes, and relative orientations of the loops. Secondly, the integral is unchanged of we switch the roles of loops 1 and 2. In other words,

$$M_{21} = M_{12}. \tag{4.191}$$

In fact, we can drop the subscripts and just call these quantities M. This is a rather surprising result. It implies that no matter what the shapes and relative positions of the two loops, the flux through loop 2 when we run a current I around loop 1 is *exactly* the same as the flux through loop 1 when we send the same current around loop 2.

We have seen that a current I flowing around some loop, 1, generates a magnetic flux linking some other loop, 2. However, flux is also generated through the first loop. As before, the magnetic field, and, therefore, the flux Φ, is proportional to the current, so we can write

$$\Phi = L\, I. \tag{4.192}$$

The constant of proportionality L is called the *self inductance*. Like M it only depends on the geometry of the loop.

Inductance is measured in S.I. units called henries (H); 1 henry is 1 volt-second per ampere. The henry, like the farad, is a rather unwieldy unit since most real life inductors have a inductances of order a micro-henry.

Consider a long solenoid of length l and radius r which has N turns per unit length, and carries a current I. The longitudinal (*i.e.*, directed along the axis of the solenoid) magnetic field within the solenoid is approximately uniform, and is given by

$$B = \mu_0 N I. \tag{4.193}$$

This result is easily obtained by integrating Ampère's law over a rectangular loop whose long sides run parallel to the axis of the solenoid, one inside the solenoid and the other outside, and whose short sides run perpendicular to the axis. The magnetic flux though each turn of the loop is $B\,\pi r^2 = \mu_0 N I\,\pi r^2$. The total flux through the solenoid wire, which has Nl turns, is

$$\Phi = Nl\,\mu_0 N I\,\pi r^2. \tag{4.194}$$

Thus, the self inductance of the solenoid is

$$L = \frac{\Phi}{I} = \mu_0 N^2 \pi r^2 l. \tag{4.195}$$

Note that the self inductance only depends on geometric quantities such as the number of turns in the solenoid and the area of the coils.

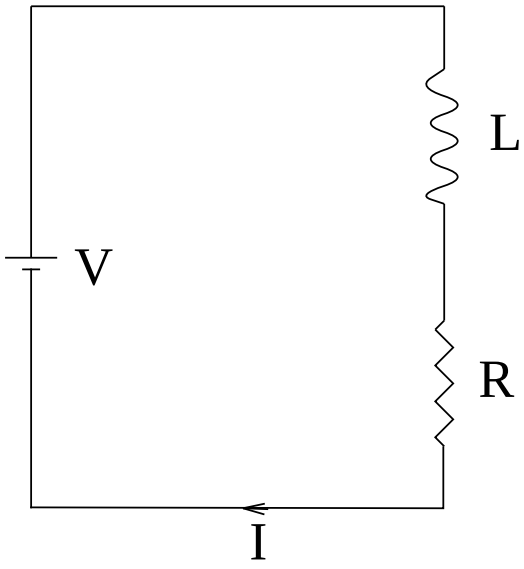
Suppose that the current I flowing through the solenoid changes. We have to assume that the change is sufficiently slow that we can neglect the displacement current and retardation effects in our calculations. This implies that the typical time-scale of the change must be much longer than the time for a light ray to traverse the circuit. If this is the case then the above formulae remain valid.

A change in the current implies a change in the magnetic flux linking the solenoid wire, since $\Phi = L\,I$. According to Faraday's law, this change generates an e.m.f. in the coils. By Lenz's law, the e.m.f. is such as to oppose the change in the current—*i.e.*, it is a back e.m.f. We can write

$$V = -\frac{d\Phi}{dt} = -L\,\frac{dI}{dt}, \tag{4.196}$$

where V is the generated e.m.f.

Suppose that our solenoid has an electrical resistance R. Let us connect the ends of the solenoid across the terminals of a battery of e.m.f. V. What is going to happen? The equivalent circuit is shown below. The inductance and resistance of the solenoid are represented by a perfect inductor L and a perfect resistor R connected in series. The voltage drop across the inductor and resistor is equal to the e.m.f. of the battery, V. The voltage drop across the resistor is simply



IR, whereas the voltage drop across the inductor (*i.e.*, minus the back e.m.f.) is $L\,dI/dt$. Here, I is the current flowing through the solenoid. It follows that

$$V = IR + L\,\frac{dI}{dt}. \tag{4.197}$$

This is a differential equation for the current I. We can rearrange it to give

$$\frac{dI}{dt} = \frac{V}{L} - \frac{R}{L}\,I. \tag{4.198}$$

The general solution is

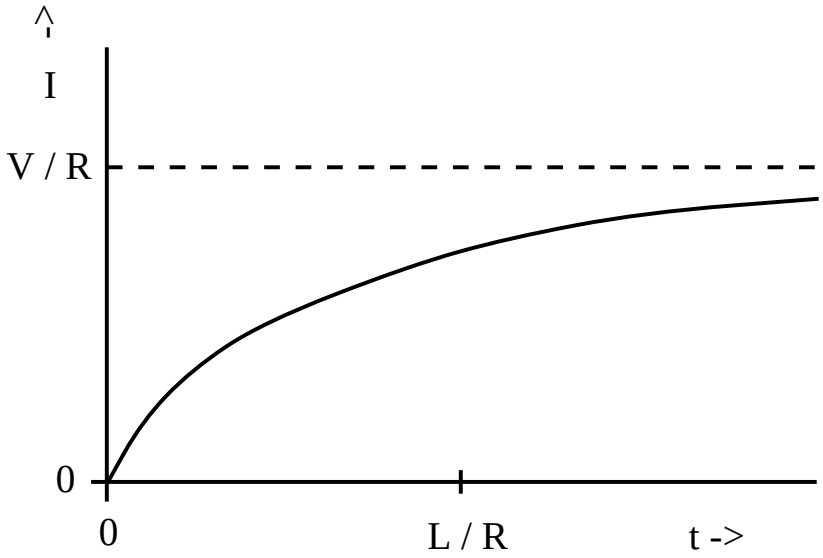
$$I(t) = \frac{V}{R} + k\exp(-R\,t/L). \tag{4.199}$$

The constant k is fixed by the boundary conditions. Suppose that the battery is connected at time $t = 0$, when $I = 0$. It follows that $k = -V/R$, so that

$$I(t) = \frac{V}{R}\,(1 - \exp(-R\,t/L)\,). \tag{4.200}$$

It can be seen from the diagram that after the battery is connected the current ramps up and attains its steady state value V/R (which comes from Ohm's law) on the characteristic time-scale

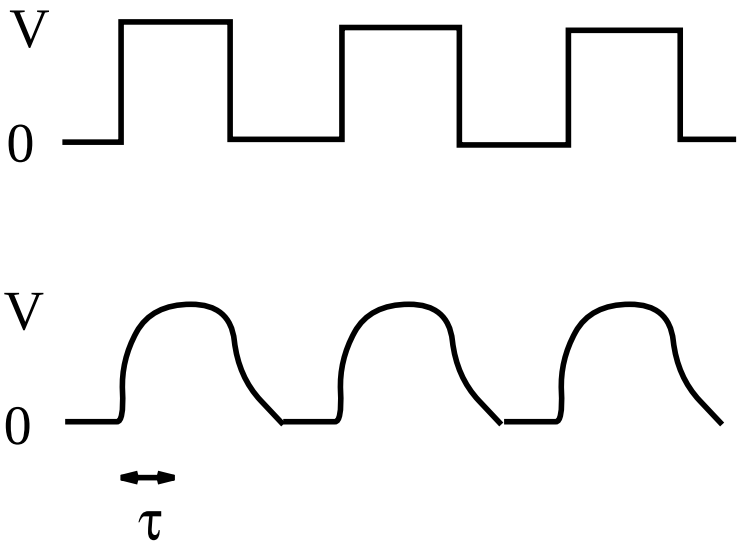
$$\tau = \frac{L}{R}. \tag{4.201}$$



This time-scale is sometimes called the "time constant" of the circuit, or, somewhat unimaginatively, the "L/R time" of the circuit.

We can now appreciate the significance of self inductance. The back e.m.f. generated in an inductor, as the current tries to change, effectively prevents the current from rising (or falling) much faster than the L/R time. This effect is sometimes advantageous, but often it is a great nuisance. All circuit elements possess some self inductance, as well as some resistance, so all have a finite L/R time. This means that when we power up a circuit the current does not jump up instantaneously to its steady state value. Instead, the rise is spread out over the L/R time of the circuit. This is a good thing. If the current were to rise instantaneously then extremely large electric fields would be generated by the sudden jump in the magnetic field, leading, inevitably, to breakdown and electric arcing. So, if there were no such thing as self inductance then every time you switched an electric circuit on or off there would be a big blue flash due to arcing between conductors. Self inductance can also be a bad thing. Suppose that we possess a fancy power supply, and we want to use it to send an electric signal down a wire (or transmission line). Of course, the wire or transmission line will possess both resistance and inductance, and will, therefore, have some characteristic L/R time. Suppose that we try to send a square wave signal down the line. Since the current in the line cannot rise or fall faster than the L/R time, the leading and trailing edges of the signal get smoothed out over an L/R time. The typical

difference between the signal fed into the wire (upper trace) and that which comes out of the other end (lower trace) is illustrated in the diagram below. Clearly, there is little point having a fancy power supply unless you also possess a low inductance wire or transmission line, so that the signal from the power supply can be transmitted to some load device without serious distortion.



Consider, now, two long thin solenoids, one wound on top of the other. The length of each solenoid is l, and the common radius is r. Suppose that the bottom coil has N_1 turns per unit length and carries a current I_1. The magnetic flux passing through each turn of the top coil is $\mu_0 N_1 I_1 \, \pi r^2$, and the total flux linking the top coil is therefore $\Phi_2 = N_2 l \, \mu_0 N_1 I_1 \, \pi r^2$, where N_2 is the number of turns per unit length in the top coil. It follows that the mutual inductance of the two coils, defined $\Phi_2 = M I_1$, is given by

$$M = \mu_0 N_1 N_2 \, \pi r^2 l. \tag{4.202}$$

Recall that the self inductance of the bottom coil is

$$L_1 = \mu_0 {N_1}^2 \, \pi r^2 l, \tag{4.203}$$

and that of the top coil is

$$L_2 = \mu_0 {N_2}^2 \, \pi r^2 l. \tag{4.204}$$

Hence, the mutual inductance can be written

$$M = \sqrt{L_1 L_2}. \tag{4.205}$$

Note that this result depends on the assumption that all of the flux produced by one coil passes through the other coil. In reality, some of the flux "leaks" out, so that the mutual inductance is somewhat less than that given in the above formula. We can write

$$M = k\sqrt{L_1 L_2}, \tag{4.206}$$

where the constant k is called the "coefficient of coupling" and lies in the range $0 \leq k \leq 1$.

Suppose that the two coils have resistances R_1 and R_2. If the bottom coil has an instantaneous current I_1 flowing through it and a total voltage drop V_1, then the voltage drop due to its resistance is $I_1 R$. The voltage drop due to the back e.m.f. generated by the self inductance of the coil is $L_1\, dI_1/dt$. There is also a back e.m.f. due to the inductive coupling to the top coil. We know that the flux through the bottom coil due to the instantaneous current I_2 flowing in the top coil is

$$\Phi_1 = M I_2. \tag{4.207}$$

Thus, by Faraday's law and Lenz's law the back e.m.f. induced in the bottom coil is

$$V = -M\,\frac{dI_2}{dt}. \tag{4.208}$$

The voltage drop across the bottom coil due to its mutual inductance with the top coil is minus this expression. Thus, the circuit equation for the bottom coil is

$$V_1 = R_1 I_1 + L_1 \frac{dI_1}{dt} + M\,\frac{dI_2}{dt}. \tag{4.209}$$

Likewise, the circuit equation for the top coil is

$$V_2 = R_2 I_2 + L_2 \frac{dI_2}{dt} + M\,\frac{dI_1}{dt}. \tag{4.210}$$

Here, V_2 is the total voltage drop across the top coil.

Suppose that we suddenly connect a battery of e.m.f. V_1 to the bottom coil at time $t = 0$. The top coil is assumed to be open circuited, or connected to a voltmeter of very high internal resistance, so that $I_2 = 0$. What is the e.m.f. generated in the top coil? Since $I_2 = 0$, the circuit equation for the bottom coil is

$$V_1 = R_1 I_1 + L_1 \frac{dI_1}{dt}, \tag{4.211}$$

where V_1 is constant, and $I_1(t = 0) = 0$. We have already seen the solution to this equation:

$$I_1 = \frac{V_1}{R_1} \left(1 - \exp(-R_1\, t/L_1)\right). \tag{4.212}$$

The circuit equation for the top coil is

$$V_2 = M \frac{dI_1}{dt}, \tag{4.213}$$

giving

$$V_2 = V_1 \frac{M}{L_1} \exp(-R_1\, t/L_1). \tag{4.214}$$

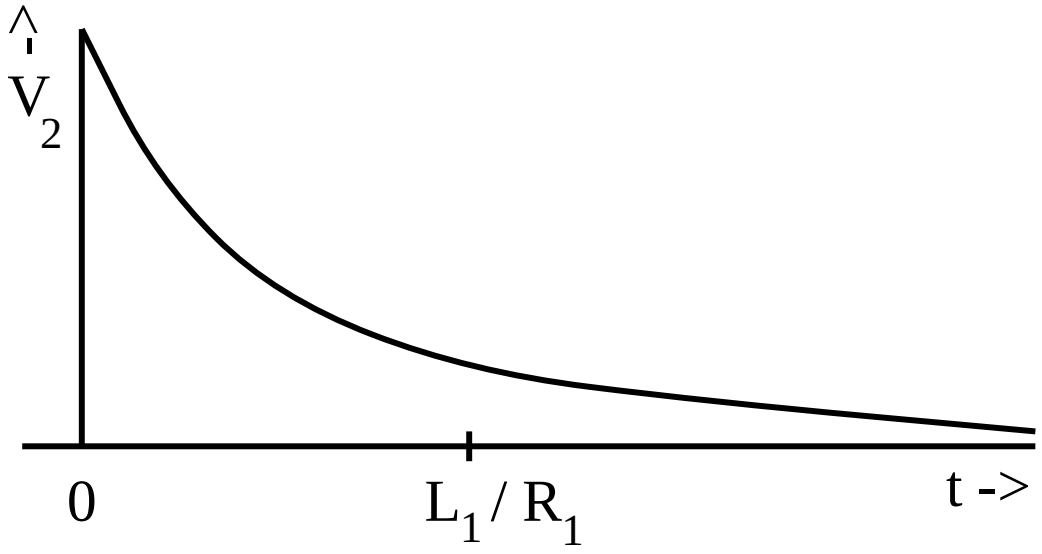
It follows from Eq. (4.206) that

$$V_2 = V_1\, k \sqrt{\frac{L_2}{L_1}} \exp(-R_1\, t/L_1). \tag{4.215}$$

Since $L_{1,\,2} \propto {N_{1,\,2}}^2$, we obtain

$$V_2 = V_1\, k\, \frac{N_2}{N_1}\, \exp(-R_1\, t/L_1). \tag{4.216}$$

Note that $V_2(t)$ is discontinuous at $t = 0$. This is not a problem since the resistance of the top circuit is infinite, so there is no discontinuity in the current (and, hence, in the magnetic field). But, what about the displacement current, which is proportional to $\partial \boldsymbol{E}/\partial t$? Surely, this is discontinuous at $t = 0$ (which is clearly unphysical)? The crucial point, here, is that we have specifically neglected the displacement current in all of our previous analysis, so it does not make much sense to start worrying about it now. If we had retained the displacement current in our calculations we would find that the voltage in the top circuit jumps up, at



$t = 0$, on a time-scale similar to the light traverse time across the circuit (*i.e.*, the jump is instantaneous to all intents and purposes, but the displacement current remains finite).

Now,

$$\frac{V_2(t=0)}{V_1} = k\,\frac{N_2}{N_1}, \tag{4.217}$$

so if $N_2 \gg N_1$ the voltage in the bottom circuit is considerably amplified in the top circuit. This effect is the basis for old-fashioned car ignition systems. A large voltage spike is induced in a secondary circuit (connected to a coil with very many turns) whenever the current in a primary circuit (connected to a coil with not so many turns) is either switched on or off. The primary circuit is connected to the car battery (whose e.m.f. is typically 12 volts). The switching is done by a set of points which are mechanically opened and closed as the engine turns. The large voltage spike induced in the secondary circuit as the points are either opened or closed causes a spark to jump across a gap in this circuit. This spark ignites a petrol/air mixture in one of the cylinders. You might think that the optimum configuration is to have only one turn in the primary circuit and lots of turns in the secondary circuit, so that the ratio N_2/N_1 is made as large as possible. However, this is not the case. Most of the magnetic field lines generated by a single turn primary coil are likely to miss the secondary coil altogether. This means that the coefficient of coupling k is small, which reduces the voltage

induced in the secondary circuit. Thus, you need a reasonable number of turns in the primary coil in order to localize the induced magnetic field so that it links effectively with the secondary coil.

4.12 Magnetic energy

Suppose that at $t = 0$ a coil of inductance L and resistance R is connected across the terminals of a battery of e.m.f. V. The circuit equation is

$$V = L\,\frac{dI}{dt} + RI. \tag{4.218}$$

The power output of the battery is VI. [Every charge q that goes around the circuit falls through a potential difference qV. In order to raise it back to the starting potential, so that it can perform another circuit, the battery must do work qV. The work done per unit time (*i.e.*, the power) is nqV, where n is the number of charges per unit time passing a given point on the circuit. But, $I = nq$, so the power output is VI.] The total work done by the battery in raising the current in the circuit from zero at time $t = 0$ to I_T at time $t = T$ is

$$W = \int_0^T VI\,dt. \tag{4.219}$$

Using the circuit equation (4.218), we obtain

$$W = L\int_0^T I\,\frac{dI}{dt}\,dt + R\int_0^T I^2\,dt, \tag{4.220}$$

giving

$$W = \frac{1}{2}LI_T^2 + R\int_0^T I^2\,dt. \tag{4.221}$$

The second term on the right-hand side represents the irreversible conversion of electrical energy into heat energy in the resistor. The first term is the amount of energy stored in the inductor at time T. This energy can be recovered after the inductor is disconnected from the battery. Suppose that the battery is

disconnected at time T. The circuit equation is now

$$0 = L\,\frac{dI}{dt} + RI, \tag{4.222}$$

giving

$$I = I_T \exp\left(-\frac{R}{L}(t-T)\right), \tag{4.223}$$

where we have made use of the boundary condition $I(T) = I_T$. Thus, the current decays away exponentially. The energy stored in the inductor is dissipated as heat in the resistor. The total heat energy appearing in the resistor after the battery is disconnected is

$$\int_T^\infty I^2 R\,dt = \frac{1}{2} L {I_T}^2, \tag{4.224}$$

where use has been made of Eq. (4.223). Thus, the heat energy appearing in the resistor is equal to the energy stored in the inductor. This energy is actually stored in the magnetic field generated around the inductor.

Consider, again, our circuit with two coils wound on top of one another. Suppose that each coil is connected to its own battery. The circuit equations are

$$\begin{aligned} V_1 &= R_1 I_1 + L\,\frac{dI_1}{dt} + M\,\frac{dI_2}{dt}, \\ V_2 &= R_2 I_2 + L\,\frac{dI_2}{dt} + M\,\frac{dI_1}{dt}, \end{aligned} \tag{4.225}$$

where V_1 is the e.m.f. of the battery in the first circuit, *etc.* The work done by the two batteries in increasing the currents in the two circuits from zero at time 0 to I_1 and I_2 at time T, respectively, is

$$\begin{aligned} W &= \int_0^T (V_1 I_1 + V_2 I_2)\,dt \\ &= \int_0^T (R_1 {I_1}^2 + R_2 {I_2}^2)\,dt + \frac{1}{2} L_1 {I_1}^2 + \frac{1}{2} L_2 {I_2}^2 \\ &\quad + M \int_0^T \left(I_1\,\frac{dI_2}{dt} + I_2\,\frac{dI_1}{dt}\right) dt. \end{aligned} \tag{4.226}$$

Thus,

$$W = \int_0^T (R_1 {I_1}^2 + R_2 {I_2}^2)\, dt + \frac{1}{2} L_1 {I_1}^2 + \frac{1}{2} L_2 {I_2}^2 + M I_1 I_2. \quad (4.227)$$

Clearly, the total magnetic energy stored in the two coils is

$$W_B = \frac{1}{2} L_1 {I_1}^2 + \frac{1}{2} L_2 {I_2}^2 + M I_1 I_2. \quad (4.228)$$

Note that the mutual inductance term increases the stored magnetic energy if I_1 and I_2 are of the same sign—*i.e.*, if the currents in the two coils flow in the same direction, so that they generate magnetic fields which reinforce one another. Conversely, the mutual inductance term decreases the stored magnetic energy if I_1 and I_2 are of the opposite sign. The total stored energy can never be negative, otherwise the coils would constitute a power source (a negative stored energy is equivalent to a positive generated energy). Thus,

$$\frac{1}{2} L_1 {I_1}^2 + \frac{1}{2} L_2 {I_2}^2 + M I_1 I_2 \geq 0, \quad (4.229)$$

which can be written

$$\frac{1}{2} \left(\sqrt{L_1} I_1 + \sqrt{L_2} I_2 \right)^2 - I_1 I_2 (\sqrt{L_1 L_2} - M) \geq 0, \quad (4.230)$$

assuming that $I_1 I_2 < 0$. It follows that

$$M \leq \sqrt{L_1 L_2}. \quad (4.231)$$

The equality sign corresponds to the situation where all of the flux generated by one coil passes through the other. If some of the flux misses then the inequality sign is appropriate. In fact, the above formula is valid for any two inductively coupled circuits.

We intimated previously that the energy stored in an inductor is actually stored in the surrounding magnetic field. Let us now obtain an explicit formula

for the energy stored in a magnetic field. Consider an ideal solenoid. The energy stored in the solenoid when a current I flows through it is

$$W = \frac{1}{2} L I^2, \tag{4.232}$$

where L is the self inductance. We know that

$$L = \mu_0 N^2 \, \pi r^2 l, \tag{4.233}$$

where N is the number of turns per unit length of the solenoid, r the radius, and l the length. The field inside the solenoid is uniform, with magnitude

$$B = \mu_0 N I, \tag{4.234}$$

and is zero outside the solenoid. Equation (4.232) can be rewritten

$$W = \frac{B^2}{2\mu_0} \, V, \tag{4.235}$$

where $V = \pi r^2 l$ is the volume of the solenoid. The above formula strongly suggests that a magnetic field possesses an energy density

$$U = \frac{B^2}{2\mu_0}. \tag{4.236}$$

Let us now examine a more general proof of the above formula. Consider a system of N circuits (labeled $i = 1$ to N), each carrying a current I_i. The magnetic flux through the ith circuit is written [*cf.*, Eq. (4.186)]

$$\Phi_i = \int \boldsymbol{B} \cdot d\boldsymbol{S}_i = \oint \boldsymbol{A} \cdot d\boldsymbol{l}_i, \tag{4.237}$$

where $\boldsymbol{B} = \nabla \wedge \boldsymbol{A}$, and $d\boldsymbol{S}_i$ and $d\boldsymbol{l}_i$ denote a surface element and a line element of this circuit, respectively. The back e.m.f. induced in the ith circuit follows from Faraday's law:

$$V_i = -\frac{d\Phi_i}{dt}. \tag{4.238}$$

The rate of work of the battery which maintains the current I_i in the ith circuit against this back e.m.f. is

$$P_i = I_i \, \frac{d\Phi_i}{dt}. \tag{4.239}$$

Thus, the total work required to raise the currents in the N circuits from zero at time 0 to $I_{0\,i}$ at time T is

$$W = \sum_{i=1}^{N} \int_0^T I_i \, \frac{d\Phi_i}{dt} \, dt. \tag{4.240}$$

The above expression for the work done is, of course, equivalent to the total energy stored in the magnetic field surrounding the various circuits. This energy is independent of the manner in which the currents are set up. Suppose, for the sake of simplicity, that the currents are ramped up linearly, so that

$$I_i = I_{0\,i} \, \frac{t}{T}. \tag{4.241}$$

The fluxes are proportional to the currents, so they must also ramp up linearly:

$$\Phi_i = \Phi_{0\,i} \, \frac{t}{T}. \tag{4.242}$$

It follows that

$$W = \sum_{i=1}^{N} \int_0^T I_{0\,i} \Phi_{0\,i} \, \frac{t}{T^2} \, dt, \tag{4.243}$$

giving

$$W = \frac{1}{2} \int_{i=1}^{N} I_{0\,i} \Phi_{0\,i}. \tag{4.244}$$

So, if instantaneous currents I_i flow in the the N circuits, which link instantaneous fluxes Φ_i, then the instantaneous stored energy is

$$W = \frac{1}{2} \int_{i=1}^{N} I_i \, \Phi_i. \tag{4.245}$$

Equations (4.237) and (4.245) imply that

$$W = \frac{1}{2}\sum_{i=1}^{N} I_i \oint \boldsymbol{A}\cdot d\boldsymbol{l}_i. \tag{4.246}$$

It is convenient, at this stage, to replace our N line currents by N current distributions of small, but finite, cross-sectional area. Equation (4.246) transforms to

$$W = \frac{1}{2}\int_V \boldsymbol{A}\cdot\boldsymbol{j}\, dV, \tag{4.247}$$

where V is a volume which contains all of the circuits. Note that for an element of the ith circuit $\boldsymbol{j} = I_i\, d\boldsymbol{l}_i/dl_i A_i$ and $dV = dl_i A_i$, where A_i is the cross-sectional area of the circuit. Now, $\mu_0\boldsymbol{j} = \nabla\wedge\boldsymbol{B}$ (we are neglecting the displacement current in this calculation), so

$$W = \frac{1}{2\mu_0}\int_V \boldsymbol{A}\cdot\nabla\wedge\boldsymbol{B}\, dV. \tag{4.248}$$

According to vector field theory,

$$\nabla\cdot(\boldsymbol{A}\wedge\boldsymbol{B}) = \boldsymbol{B}\cdot\nabla\wedge\boldsymbol{A} - \boldsymbol{A}\cdot\nabla\wedge\boldsymbol{B}, \tag{4.249}$$

which implies that

$$W = \frac{1}{2\mu_0}\int_V (-\nabla\cdot(\boldsymbol{A}\wedge\boldsymbol{B}) + \boldsymbol{B}\cdot\nabla\wedge\boldsymbol{A})\, dV. \tag{4.250}$$

Using Gauss' theorem and $\boldsymbol{B} = \nabla\wedge\boldsymbol{A}$, we obtain

$$W = -\frac{1}{2\mu_0}\oint_S \boldsymbol{A}\wedge\boldsymbol{B}\cdot d\boldsymbol{S} + \frac{1}{2\mu_0}\int_V B^2\, dV, \tag{4.251}$$

where S is the bounding surface of V. Let us take this surface to infinity. It is easily demonstrated that the magnetic field generated by a current loop falls of like r^{-3} at large distances. The vector potential falls off like r^{-2}. However, the area of surface S only increases like r^2. It follows that the surface integral is negligible in the limit $r\rightarrow\infty$. Thus, the above expression reduces to

$$W = \int_{\text{all space}} \frac{B^2}{2\mu_0}\, dV. \tag{4.252}$$

Since this expression is valid for any magnetic field whatsoever, we can conclude that the energy density of a general magnetic field is given by

$$U = \frac{B^2}{2\mu_0}. \tag{4.253}$$

4.13 Energy conservation in electromagnetism

We have seen that the energy density of an electric field is given by

$$U_E = \frac{\epsilon_0\, E^2}{2}, \tag{4.254}$$

whereas the energy density of a magnetic field satisfies

$$U_B = \frac{B^2}{2\mu_0}. \tag{4.255}$$

This suggests that the energy density of a general electromagnetic field is

$$U = \frac{\epsilon_0\, E^2}{2} + \frac{B^2}{2\mu_0}. \tag{4.256}$$

We are now in a position to demonstrate that the classical theory of electromagnetism conserves energy. We have already come across one conservation law in electromagnetism:

$$\frac{\partial \rho}{\partial t} + \nabla\cdot\boldsymbol{j} = 0. \tag{4.257}$$

This is the equation of charge conservation. Integrating over some volume V bounded by a surface S, we obtain

$$-\frac{\partial}{\partial t}\int_V \rho\, dV = \oint_S \boldsymbol{j}\cdot d\boldsymbol{S}. \tag{4.258}$$

In other words, the rate of decrease of the charge contained in volume V equals the net flux of charge across surface S. This suggests that an energy conservation law for electromagnetism should have the form

$$-\frac{\partial}{\partial t}\int_V U\, dV = \oint_S \boldsymbol{u}\cdot d\boldsymbol{S}. \tag{4.259}$$

Here, U is the energy density of the electromagnetic field and $\boldsymbol{u}$ is the flux of electromagnetic energy (*i.e.*, energy $|\boldsymbol{u}|$ per unit time, per unit cross-sectional area, passes a given point in the direction of $\boldsymbol{u}$). According to the above equation, the rate of decrease of the electromagnetic energy in volume V equals the net flux of electromagnetic energy across surface S.

Equation (4.259) is incomplete because electromagnetic fields can lose or gain energy by interacting with matter. We need to factor this into our analysis. We saw earlier (see Section 4.2) that the rate of heat dissipation per unit volume in a conductor (the so-called ohmic heating rate) is $\boldsymbol{E}\cdot\boldsymbol{j}$. This energy is extracted from electromagnetic fields, so the rate of energy loss of the fields in volume V due to interaction with matter is $\int_V \boldsymbol{E}\cdot\boldsymbol{j}\,dV$. Thus, Eq. (4.259) generalizes to

$$-\frac{\partial}{\partial t}\int_V U\,dV = \oint_S \boldsymbol{u}\cdot d\boldsymbol{S} + \int_V \boldsymbol{E}\cdot\boldsymbol{j}\,dV. \tag{4.260}$$

The above equation is equivalent to

$$\frac{\partial U}{\partial t} + \nabla\cdot\boldsymbol{u} = -\boldsymbol{E}\cdot\boldsymbol{j}. \tag{4.261}$$

Let us now see if we can derive an expression of this form from Maxwell's equations.

We start from Ampère's law (including the displacement current):

$$\nabla\wedge\boldsymbol{B} = \mu_0\boldsymbol{j} + \epsilon_0\mu_0\frac{\partial\boldsymbol{E}}{\partial t}. \tag{4.262}$$

Dotting this equation with the electric field yields

$$-\boldsymbol{E}\cdot\boldsymbol{j} = -\frac{\boldsymbol{E}\cdot\nabla\wedge\boldsymbol{B}}{\mu_0} + \epsilon_0\boldsymbol{E}\cdot\frac{\partial\boldsymbol{E}}{\partial t}. \tag{4.263}$$

This can be rewritten

$$-\boldsymbol{E}\cdot\boldsymbol{j} = -\frac{\boldsymbol{E}\cdot\nabla\wedge\boldsymbol{B}}{\mu_0} + \frac{\partial}{\partial t}\left(\frac{\epsilon_0\,E^2}{2}\right). \tag{4.264}$$

Now, from vector field theory

$$\nabla\cdot(\boldsymbol{E}\wedge\boldsymbol{B}) = \boldsymbol{B}\cdot\nabla\wedge\boldsymbol{E} - \boldsymbol{E}\cdot\nabla\wedge\boldsymbol{B}, \tag{4.265}$$

so

$$-\boldsymbol{E}\cdot\boldsymbol{j} = \nabla\cdot\left(\frac{\boldsymbol{E}\wedge\boldsymbol{B}}{\mu_0}\right) - \frac{\boldsymbol{B}\cdot\nabla\wedge\boldsymbol{E}}{\mu_0} + \frac{\partial}{\partial t}\left(\frac{\epsilon_0\,E^2}{2}\right). \tag{4.266}$$

Faraday's law yields

$$\nabla\wedge\boldsymbol{E} = -\frac{\partial\boldsymbol{B}}{\partial t}, \tag{4.267}$$

so

$$-\boldsymbol{E}\cdot\boldsymbol{j} = \nabla\cdot\left(\frac{\boldsymbol{E}\wedge\boldsymbol{B}}{\mu_0}\right) + \frac{1}{\mu_0}\boldsymbol{B}\cdot\frac{\partial\boldsymbol{B}}{\partial t} + \frac{\partial}{\partial t}\left(\frac{\epsilon_0\,E^2}{2}\right). \tag{4.268}$$

This can be rewritten

$$-\boldsymbol{E}\cdot\boldsymbol{j} = \nabla\cdot\left(\frac{\boldsymbol{E}\wedge\boldsymbol{B}}{\mu_0}\right) + \frac{\partial}{\partial t}\left(\frac{\epsilon_0\,E^2}{2} + \frac{B^2}{2\mu_0}\right). \tag{4.269}$$

Thus, we obtain the desired conservation law,

$$\frac{\partial U}{\partial t} + \nabla\cdot\boldsymbol{u} = -\boldsymbol{E}\cdot\boldsymbol{j}, \tag{4.270}$$

where

$$U = \frac{\epsilon_0\,E^2}{2} + \frac{B^2}{2\mu_0} \tag{4.271}$$

is the electromagnetic energy density, and

$$\boldsymbol{u} = \frac{\boldsymbol{E}\wedge\boldsymbol{B}}{\mu_0} \tag{4.272}$$

is the electromagnetic energy flux. The latter quantity is usually called the "Poynting flux" after its discoverer.

Let us see whether our expression for the electromagnetic energy flux makes sense. We all know that if we stand in the sun we get hot (especially in Texas!). This occurs because we absorb electromagnetic radiation emitted by the Sun. So, radiation must transport energy. The electric and magnetic fields in electromagnetic radiation are mutually perpendicular, and are also perpendicular to the direction of propagation $\hat{\boldsymbol{k}}$ (this is a unit vector). Furthermore, $B = E/c$.

Equation (3.232) can easily be transformed into the following relation between the electric and magnetic fields of an electromagnetic wave:

$$\boldsymbol{E} \wedge \boldsymbol{B} = \frac{E^2}{c}\hat{\boldsymbol{k}}. \tag{4.273}$$

Thus, the Poynting flux for electromagnetic radiation is

$$\boldsymbol{u} = \frac{E^2}{\mu_0 c}\,\hat{\boldsymbol{k}} = \epsilon_0 c E^2\,\hat{\boldsymbol{k}}. \tag{4.274}$$

This expression tells us that electromagnetic waves transport energy along their direction of propagation, which seems to make sense.

The energy density of electromagnetic radiation is

$$U = \frac{\epsilon_0\,E^2}{2} + \frac{B^2}{2\mu_0} = \frac{\epsilon_0\,E^2}{2} + \frac{E^2}{2\mu_0 c^2} = \epsilon_0 E^2, \tag{4.275}$$

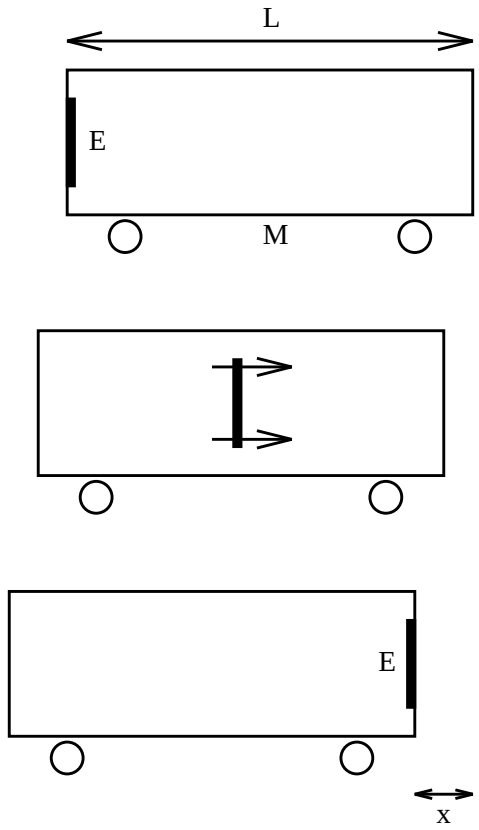
using $B = E/c$. Note that the electric and magnetic fields have equal energy densities. Since electromagnetic waves travel at the speed of light, we would expect the energy flux through one square meter in one second to equal the energy contained in a volume of length c and unit cross-sectional area; *i.e.*, c times the energy density. Thus,

$$|\boldsymbol{u}| = cU = \epsilon_0 c E^2, \tag{4.276}$$

which is in accordance with Eq. (4.274).

4.14 Electromagnetic momentum

We have seen that electromagnetic waves carry energy. It turns out that they also carry momentum. Consider the following argument, due to Einstein. Suppose that we have a railroad car of mass M and length L which is free to move in one dimension. Suppose that electromagnetic radiation of total energy E is emitted from one end of the car, propagates along the length of the car, and is then absorbed at the other end. The effective mass of this radiation is $m = E/c^2$



(from Einstein's famous relation $E = mc^2$). At first sight, the process described above appears to cause the centre of mass of the system to spontaneously shift. This violates the law of momentum conservation (assuming the railway car is subject to no external forces). The only way in which the centre of mass of the system can remain stationary is if the railway car *moves* in the opposite direction to the direction of propagation of the radiation. In fact, if the car moves by a distance x then the centre of mass of the system is the same before and after the radiation pulse provided that

$$Mx = mL = \frac{E}{c^2}\, L. \tag{4.277}$$

It is assumed that $m \ll M$ in this derivation.

But, what actually causes the car to move? If the radiation possesses momentum p then the car will recoil with the same momentum as the radiation is emitted. When the radiation hits the other end of the car then it acquires mo-

mentum p in the opposite direction, which stops the motion. The time of flight of the radiation is L/c. So, the distance traveled by a mass M with momentum p in this time is

$$x = vt = \frac{p}{M}\frac{L}{c}, \tag{4.278}$$

giving

$$p = Mx\,\frac{c}{L} = \frac{E}{c}. \tag{4.279}$$

Thus, the momentum carried by electromagnetic radiation equals its energy divided by the speed of light. The same result can be obtained from the well known relativistic formula

$$E^2 = p^2c^2 + m^2c^4 \tag{4.280}$$

relating the energy E, momentum p, and mass m of a particle. According to quantum theory, electromagnetic radiation is made up of massless particles called photons. Thus,

$$p = \frac{E}{c} \tag{4.281}$$

for individual photons, so the same must be true of electromagnetic radiation as a whole. If follows from Eq. (4.281) that the momentum density g of electromagnetic radiation equals its energy density over c, so

$$g = \frac{U}{c} = \frac{|\boldsymbol{u}|}{c^2} = \frac{\epsilon_0 E^2}{c}. \tag{4.282}$$

It is reasonable to suppose that the momentum points along the direction of the energy flow (this is obviously the case for photons), so the vector momentum density (which gives the direction as well as the magnitude, of the momentum per unit volume) of electromagnetic radiation is

$$\boldsymbol{g} = \frac{\boldsymbol{u}}{c^2}. \tag{4.283}$$

Thus, the momentum density equals the energy flux over c^2.

Of course, the electric field associated with an electromagnetic wave oscillates rapidly, which implies that the previous expressions for the energy density, energy flux, and momentum density of electromagnetic radiation are also rapidly

oscillating. It is convenient to average over many periods of the oscillation (this average is denoted $\langle\,\rangle$). Thus,

$$\begin{aligned}\langle U\rangle &= \frac{\epsilon_0 {E_0}^2}{2},\\ \langle \boldsymbol{u}\rangle &= \frac{c\epsilon_0 {E_0}^2}{2} = c\,\langle U\rangle \hat{\boldsymbol{k}},\\ \langle \boldsymbol{g}\rangle &= \frac{\epsilon_0 {E_0}^2}{2c}\,\hat{\boldsymbol{k}} = \frac{\langle U\rangle}{c}\,\hat{\boldsymbol{k}},\end{aligned} \tag{4.284}$$

where the factor $1/2$ comes from averaging $\cos^2\omega t$. Here, E_0 is the peak amplitude of the electric field associated with the wave.

Since electromagnetic radiation possesses momentum then it must exert a force on bodies which absorb (or emit) radiation. Suppose that a body is placed in a beam of perfectly collimated radiation, which it absorbs completely. The amount of momentum absorbed per unit time, per unit cross-sectional area, is simply the amount of momentum contained in a volume of length c and unit cross-sectional area; *i.e.*, c times the momentum density g. An absorbed momentum per unit time, per unit area, is equivalent to a pressure. In other words, the radiation exerts a pressure cg on the body. Thus, the "radiation pressure" is given by

$$p = \frac{\epsilon_0\, E^2}{2} = \langle U\rangle. \tag{4.285}$$

So, the pressure exerted by collimated electromagnetic radiation is equal to its average energy density.

Consider a cavity filled with electromagnetic radiation. What is the radiation pressure exerted on the walls? In this situation the radiation propagates in all directions with equal probability. Consider radiation propagating at an angle θ to the local normal to the wall. The amount of such radiation hitting the wall per unit time, per unit area, is proportional to $\cos\theta$. Moreover, the component of momentum normal to the wall which the radiation carries is also proportional to $\cos\theta$. Thus, the pressure exerted on the wall is the same as in Eq. (4.285), except that it is weighted by the average of $\cos^2\theta$ over all solid angles in order to take into account the fact that obliquely propagating radiation exerts a pressure

which is $\cos^2\theta$ times that of normal radiation. The average of $\cos^2\theta$ over all solid angles is 1/3, so for isotropic radiation

$$p = \frac{\langle U\rangle}{3}. \tag{4.286}$$

Clearly, the pressure exerted by isotropic radiation is one third of its average energy density.

The power incident on the surface of the Earth due to radiation emitted by the Sun is about 1300 W/m^2. So, what is the radiation pressure? Since,

$$\langle|\boldsymbol{u}|\rangle = c\,\langle U\rangle = 1300\,\mathrm{Wm}^{-2}, \tag{4.287}$$

then

$$p = \langle U\rangle \simeq 4\times 10^{-6}\,\mathrm{Nm}^{-2}. \tag{4.288}$$

Here, the radiation is assumed to be perfectly collimated. Thus, the radiation pressure exerted on the Earth is minuscule (one atmosphere equals about 10^5 N/m^2). Nevertheless, this small pressure due to radiation is important in outer space, since it is responsible for continuously sweeping dust particles out of the solar system. It is quite common for comets to exhibit two separate tails. One (called the "gas tail") consists of ionized gas, and is swept along by the solar wind (a stream of charged particles and magnetic field lines emitted by the Sun). The other (called the "dust tail") consists of uncharged dust particles, and is swept radially outward from the Sun by radiation pressure. Two separate tails are observed if the local direction of the solar wind is not radially outward from the Sun (which is quite often the case).

The radiation pressure from sunlight is very weak. However, that produced by laser beams can be enormous (far higher than any conventional pressure which has ever been produced in a laboratory). For instance, the lasers used in Inertial Confinement Fusion (*e.g.*, the NOVA experiment in Laurence Livermore National Laboratory) typically have energy fluxes of 10^{18} Wm^{-2}. This translates to a radiation pressure of about 10^4 atmospheres. Obviously, it would not be a good idea to get in the way of one of these lasers!

4.15 The Hertzian dipole

Consider two spherical conductors connected by a wire. Suppose that electric charge flows periodically back and forth between the spheres. Let q be the charge on one of the conductors. The system has zero net charge, so the charge on the other conductor is $-q$. Let

$$q(t) = q_0 \sin \omega t. \tag{4.289}$$

We expect the oscillating current flowing in the wire connecting the two spheres to generate electromagnetic radiation (see Section 3.23). Let us consider the simple case where the length of the wire is small compared to the wavelength of the emitted radiation. If this is the case then the current I flowing between the conductors has the same phase along the whole length of the wire. It follows that

$$I(t) = \frac{dq}{dt} = I_0 \cos \omega t, \tag{4.290}$$

where $I_0 = \omega q_0$. This type of antenna is called a Hertzian dipole, after the German physicist Heinrich Hertz.

The magnetic vector potential generated by a current distribution $\boldsymbol{j}(\boldsymbol{r})$ is given by the well known formula

$$\boldsymbol{A}(\boldsymbol{r}, t) = \frac{\mu_0}{4\pi} \int \frac{[\boldsymbol{j}]}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d^3\boldsymbol{r}', \tag{4.291}$$

where

$$[f] = f(\boldsymbol{r}', t - |\boldsymbol{r} - \boldsymbol{r}'|/c). \tag{4.292}$$

Suppose that the wire is aligned along the z-axis and extends from $z = -l/2$ to $z = l/2$. For a wire of negligible thickness we can replace $\boldsymbol{j}(\boldsymbol{r}', t - |\boldsymbol{r} - \boldsymbol{r}'|/c)\, d^3\boldsymbol{r}'$ by $I(\boldsymbol{r}', t - |\boldsymbol{r} - \boldsymbol{r}'|/c)\, dz'\, \hat{\boldsymbol{z}}$. Thus, $\boldsymbol{A}(\boldsymbol{r}, t) = A_z(\boldsymbol{r}, t)\, \hat{\boldsymbol{z}}$ and

$$A_z(\boldsymbol{r}, t) = \frac{\mu_0}{4\pi} \int_{-l/2}^{l/2} \frac{I(z', t - |\boldsymbol{r} - z'\, \hat{\boldsymbol{z}}|/c)}{|\boldsymbol{r} - z'\, \hat{\boldsymbol{z}}|}\, dz'. \tag{4.293}$$

In the region $r \gg l$

$$|\boldsymbol{r} - z'\, \hat{\boldsymbol{z}}| \simeq r \tag{4.294}$$

and

$$t - |\boldsymbol{r} - z'\,\hat{\boldsymbol{z}}|/c \simeq t - r/c. \tag{4.295}$$

The maximum error in the latter approximation is $\Delta t \sim l/c$. This error (which is a time) must be much less than a period of oscillation of the emitted radiation, otherwise the phase of the radiation will be wrong. So

$$\frac{l}{c} \ll \frac{2\pi}{\omega}, \tag{4.296}$$

which implies that $l \ll \lambda$, where $\lambda = 2\pi\, c/\omega$ is the wavelength of the emitted radiation. However, we have already assumed that the length of the wire l is much less than the wavelength of the radiation, so the above inequality is automatically satisfied. Thus, in the "far field" region, $r \gg \lambda$, we can write

$$A_z(\boldsymbol{r}, t) \simeq \frac{\mu_0}{4\pi} \int_{-l/2}^{l/2} \frac{I(z', t - r/c)}{r}\, dz'. \tag{4.297}$$

This integral is easy to perform since the current is uniform along the length of the wire. So,

$$A_z(\boldsymbol{r}, t) \simeq \frac{\mu_0 l}{4\pi} \frac{I(t - r/c)}{r}. \tag{4.298}$$

The scalar potential is most conveniently evaluated using the Lorentz gauge condition

$$\nabla \cdot \boldsymbol{A} = -\epsilon_0 \mu_0 \frac{\partial \phi}{\partial t}. \tag{4.299}$$

Now,

$$\nabla \cdot \boldsymbol{A} = \frac{\partial A_z}{\partial z} \simeq \frac{\mu_0 l}{4\pi} \frac{\partial I(t - r/c)}{\partial t} \left(-\frac{z}{r^2 c}\right) + O\left(\frac{1}{r^2}\right) \tag{4.300}$$

to leading order in r^{-1}. Thus,

$$\phi(\boldsymbol{r}, t) \simeq \frac{l}{4\pi\epsilon_0 c} \frac{z}{r} \frac{I(t - r/c)}{r}. \tag{4.301}$$

Given the vector and scalar potentials, Eqs. (4.298) and (4.301), respectively, we can evaluate the associated electric and magnetic fields using

$$\boldsymbol{E} \;=\; -\frac{\partial \boldsymbol{A}}{\partial t} - \nabla\phi,$$

$$\boldsymbol{B} \;=\; \nabla \wedge \boldsymbol{A}. \tag{4.302}$$

Note that we are only interested in radiation fields, which fall off like r^{-1} with increasing distance from the source. It is easily demonstrated that

$$\boldsymbol{E} \simeq -\frac{\omega l I_0}{4\pi\epsilon_0 c^2}\,\sin\theta\,\frac{\sin[\omega(t-r/c)]}{r}\,\hat{\boldsymbol{\theta}} \tag{4.303}$$

and

$$\boldsymbol{B} \simeq -\frac{\omega l I_0}{4\pi\epsilon_0 c^3}\,\sin\theta\,\frac{\sin[\omega(t-r/c)]}{r}\,\hat{\boldsymbol{\phi}}. \tag{4.304}$$

Here, $(r,\,\theta,\,\phi)$ are standard spherical polar coordinates aligned along the z-axis. The above expressions for the far field (*i.e.*, $r \gg \lambda$) electromagnetic fields generated by a localized oscillating current are also easily derived from Eqs. (3.320) and (3.321). Note that the fields are symmetric in the azimuthal angle ϕ. There is no radiation along the axis of the oscillating dipole (*i.e.*, $\theta = 0$), and the maximum emission is in the plane perpendicular to this axis (*i.e.*, $\theta = \pi/2$).

The average power crossing a spherical surface S (whose radius is much greater than λ) is

$$P_{\rm rad} = \oint_S \langle \boldsymbol{u}\rangle \cdot d\boldsymbol{S}, \tag{4.305}$$

where the average is over a single period of oscillation of the wave, and the Poynting flux is given by

$$\boldsymbol{u} = \frac{\boldsymbol{E}\wedge\boldsymbol{B}}{\mu_0} = \frac{\omega^2 l^2 {I_0}^2}{16\pi^2\epsilon_0 c^3}\,\sin^2[\omega(t-r/c)]\,\frac{\sin^2\theta}{r^2}\,\hat{\boldsymbol{r}}. \tag{4.306}$$

It follows that

$$\langle \boldsymbol{u}\rangle = \frac{\omega^2 l^2 {I_0}^2}{32\pi^2\epsilon_0 c^3}\,\frac{\sin^2\theta}{r^2}\,\hat{\boldsymbol{r}}. \tag{4.307}$$

Note that the energy flux is radially outwards from the source. The total power flux across S is given by

$$P_{\rm rad} = \frac{\omega^2 l^2 {I_0}^2}{32\pi^2\epsilon_0 c^3}\int_0^{2\pi} d\phi \int_0^{\pi} \frac{\sin^2\theta}{r^2}\;r^2\sin^2\theta\;d\theta. \tag{4.308}$$

Thus,

$$P_{\rm rad} = \frac{\omega^2 l^2 {I_0}^2}{12\pi\epsilon_0 c^3}. \tag{4.309}$$

The total flux is independent of the radius of S, as is to be expected if energy is conserved.

Recall that for a resistor of resistance R the average ohmic heating power is

$$P_{\rm heat} = \langle I^2 R\rangle = \frac{1}{2}{I_0}^2 R, \tag{4.310}$$

assuming that $I = I_0 \cos\omega t$. It is convenient to define the "radiation resistance" of a Hertzian dipole antenna:

$$R_{\rm rad} = \frac{P_{\rm rad}}{{I_0}^2/2}, \tag{4.311}$$

so that

$$R_{\rm rad} = \frac{2\pi}{3\epsilon_0 c}\left(\frac{l}{\lambda}\right)^2, \tag{4.312}$$

where $\lambda = 2\pi\, c/\omega$ is the wavelength of the radiation. In fact,

$$R_{\rm rad} = 789\left(\frac{l}{\lambda}\right)^2 \text{ ohms}. \tag{4.313}$$

In the theory of electrical circuits, antennas are conventionally represented as resistors whose resistance is equal to the characteristic radiation resistance of the antenna plus its real resistance. The power loss ${I_0}^2 R_{\rm rad}/2$ associated with the radiation resistance is due to the emission of electromagnetic radiation. The power loss ${I_0}^2 R/2$ associated with the real resistance is due to ohmic heating of the antenna.

Note that the formula (4.313) is only valid for $l \ll \lambda$. This suggests that $R_{\rm rad} \ll R$ for most Hertzian dipole antennas; *i.e.*, the radiated power is swamped by the ohmic losses. Thus, antennas whose lengths are much less than that of the emitted radiation tend to be extremely inefficient. In fact, it is necessary to have $l \sim \lambda$ in order to obtain an efficient antenna. The simplest practical antenna

is the "half-wave antenna," for which $l = \lambda/2$. This can be analyzed as a series of Hertzian dipole antennas stacked on top of one another, each slightly out of phase with its neighbours. The characteristic radiation resistance of a half-wave antenna is

$$R_{\rm rad} = \frac{2.44}{4\pi\epsilon_0 c} = 73 \text{ ohms}. \tag{4.314}$$

Antennas can be used to receive 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. It is easily demonstrated that antennas most readily detect electromagnetic radiation incident from those directions in which they preferentially emit radiation. Thus, a Hertzian dipole antenna is unable to detect radiation incident along its axis, and most efficiently detects radiation incident in the plane perpendicular to this axis. In the theory of electrical circuits, a receiving antenna is represented as an e.m.f 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:

$$V_0 \cos\omega t = I_0 \cos\omega t\, (R_{\rm rad} + R_{\rm load}), \tag{4.315}$$

where $I = I_0 \cos\omega t$ is the current induced in the circuit.

The power input to the circuit is

$$P_{\rm in} = \langle V I\rangle = \frac{{V_0}^2}{2(R_{\rm rad} + R_{\rm load})}. \tag{4.316}$$

The power transferred to the load is

$$P_{\rm load} = \langle I^2 R_{\rm load}\rangle = \frac{R_{\rm load}\,{V_0}^2}{2(R_{\rm rad} + R_{\rm load})^2}. \tag{4.317}$$

The power re-radiated by the antenna is

$$P_{\rm rad} = \langle I^2 R_{\rm rad}\rangle = \frac{R_{\rm rad}\,{V_0}^2}{2(R_{\rm rad}+R_{\rm load})^2}. \tag{4.318}$$

Note that $P_{\rm in} = P_{\rm load} + P_{\rm rad}$. The maximum power transfer to the load occurs when

$$\frac{\partial P_{\rm load}}{\partial R_{\rm load}} = \frac{{V_0}^2}{2}\left[\frac{R_{\rm load}-R_{\rm rad}}{(R_{\rm rad}+R_{\rm load})^3}\right] = 0. \tag{4.319}$$

Thus, the maximum transfer rate corresponds to

$$R_{\rm load} = R_{\rm res}. \tag{4.320}$$

In other words, the resistance of the load circuit must match the radiation resistance of the antenna. For this optimum case,

$$P_{\rm load} = P_{\rm rad} = \frac{{V_0}^2}{8R_{\rm rad}} = \frac{P_{\rm in}}{2}. \tag{4.321}$$

So, in the optimum case *half* of the power absorbed by the antenna is immediately 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

$$V_0 = E_0 l. \tag{4.322}$$

Here, we have used the fact that the wavelength of the radiation is much longer than the length 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

$$\langle u_{\rm in}\rangle = \frac{\epsilon_0 c {E_0}^2}{2}, \tag{4.323}$$

whereas the power transferred to a properly matched detector circuit is

$$P_{\rm load} = \frac{{E_0}^2 l^2}{8 R_{\rm rad}}. \tag{4.324}$$

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

$$P_{\rm load} = \langle u_{\rm in}\rangle A_{\rm eff}. \tag{4.325}$$

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. Thus,

$$P_{\rm load} = \frac{{E_0}^2 l^2}{8 R_{\rm rad}} = \frac{\epsilon_0 c {E_0}^2}{2}\, A_{\rm eff}, \tag{4.326}$$

giving

$$A_{\rm eff} = \frac{l^2}{4\epsilon_0 c R_{\rm rad}} = \frac{3}{8\pi}\,\lambda^2. \tag{4.327}$$

It is clear that the effective area of a Hertzian dipole antenna is of order the wavelength squared of the incoming radiation.

For a properly aligned half-wave antenna

$$A_{\rm eff} = 0.13\,\lambda^2. \tag{4.328}$$

Thus, the antenna, which is essentially one dimensional with length $\lambda/2$, acts as if it is two dimensional, with width $0.26\,\lambda$, as far as its absorption of incoming electromagnetic radiation is concerned.

4.16 AC circuits

Alternating current (AC) circuits are made up of e.m.f. sources and *three* different types of passive element; resistors, inductors, and capacitors, Resistors satisfy Ohm's law:

$$V = IR, \tag{4.329}$$

where R is the resistance, I is the current flowing through the resistor, and V is the voltage drop across the resistor (in the direction in which the current flows). Inductors satisfy

$$V = L\,\frac{dL}{dt}, \tag{4.330}$$

where L is the inductance. Finally, capacitors obey

$$V = \frac{q}{C} = \int_0^t I\,dt \Big/ C, \tag{4.331}$$

where C is the capacitance, q is the charge stored on the plate with the more positive potential, and $I = 0$ for $t < 0$. Note that any passive component of a real electrical circuit can always be represented as a combination of ideal resistors, inductors, and capacitors.

Let us consider the classic LCR circuit, which consists of an inductor L, a capacitor C, and a resistor R, all connected in series with an e.m.f. source V. The circuit equation is obtained by setting the input voltage V equal to the sum of the voltage drops across the three passive elements in the circuit. Thus,

$$V = I\,R + L\,\frac{dI}{dt} + \int_0^t I\,dt \Big/ C. \tag{4.332}$$

This is an integro-differential equation which, in general, is quite tricky to solve. Suppose, however, that both the voltage and the current oscillate at some angular frequency ω, so that

$$\begin{aligned} V(t) &= V_0 \exp(\mathrm{i}\,\omega t), \\ I(t) &= I_0 \exp(\mathrm{i}\,\omega t), \end{aligned} \tag{4.333}$$

where the physical solution is understood to be the *real part* of the above expressions. The assumed behaviour of the voltage and current is clearly relevant to electrical circuits powered by the mains voltage (which oscillates at 60 hertz).

Equations (4.332) and (4.333) yield

$$V_0 \exp(\mathrm{i}\,\omega t) = I_0 \exp(\mathrm{i}\,\omega t)\,R + L\,\mathrm{i}\,\omega\,I_0 \exp(\mathrm{i}\,\omega t) + \frac{I_0 \exp(\mathrm{i}\,\omega t)}{\mathrm{i}\,\omega C}, \tag{4.334}$$

giving

$$V_0 = I_0 \left(\mathrm{i}\,\omega L + \frac{1}{\mathrm{i}\,\omega C} + R \right). \tag{4.335}$$

It is helpful to define the "impedance" of the circuit;

$$Z = \frac{V}{I} = \mathrm{i}\,\omega L + \frac{1}{\mathrm{i}\,\omega C} + R. \tag{4.336}$$

Impedance is a generalization of the concept of resistance. In general, the impedance of an AC circuit is a *complex* quantity.

The average power output of the e.m.f. source is

$$P = \langle V(t) I(t) \rangle, \tag{4.337}$$

where the average is taken over one period of the oscillation. Let us, first of all, calculate the power using real (rather than complex) voltages and currents. We can write

$$\begin{aligned} V(t) &= V_0 \cos\omega t, \\ I(t) &= I_0 \cos(\omega t - \theta), \end{aligned} \tag{4.338}$$

where θ is the phase lag of the current with respect to the voltage. It follows that

$$\begin{aligned} P &= V_0 I_0 \int_{\omega t=0}^{\omega t=2\pi} \cos\omega t\, \cos(\omega t - \theta)\, \frac{d(\omega t)}{2\pi} \\ &= V_0 I_0 \int_{\omega t=0}^{\omega t=2\pi} \cos\omega t\, (\cos\omega t\, \cos\theta + \sin\omega t\, \sin\theta)\, \frac{d(\omega t)}{2\pi}, \end{aligned} \tag{4.339}$$

giving

$$P = \frac{1}{2} V_0 I_0 \cos\theta, \tag{4.340}$$

since $\langle \cos\omega t\, \sin\omega t \rangle = 0$ and $\langle \cos\omega t\, \cos\omega t \rangle = 1/2$. In complex representation, the voltage and the current are written

$$\begin{aligned} V(t) &= V_0 \exp(\mathrm{i}\,\omega t), \\ I(t) &= I_0 \exp[\mathrm{i}\,(\omega t - \theta)], \end{aligned} \tag{4.341}$$

where I_0 and V_0 are assumed to be real quantities. Note that

$$\frac{1}{2}(V I^* + V^* I) = V_0 I_0 \cos\theta. \tag{4.342}$$

It follows that

$$P = \frac{1}{4}(V I^* + V^* I) = \frac{1}{2}\,\mathrm{Re}(V I^*). \tag{4.343}$$

Making use of Eq. (4.336), we find that

$$P = \frac{1}{2}\,\mathrm{Re}(Z)\,|I|^2 = \frac{1}{2}\frac{\mathrm{Re}(Z)\,|V|^2}{|Z|^2}. \tag{4.344}$$

Note that power dissipation is associated with the *real part* of the impedance. For the special case of an LCR circuit,

$$P = \frac{1}{2} R {I_0}^2. \tag{4.345}$$

It is clear that only the resistor dissipates energy in this circuit. The inductor and the capacitor both store energy, but they eventually return it to the circuit without dissipation.

According to Eq. (4.336), the amplitude of the current which flows in an LCR circuit for a given amplitude of the input voltage is given by

$$I_0 = \frac{V_0}{|Z|} = \frac{V_0}{\sqrt{(\omega L - 1/\omega C)^2 + R^2}}. \tag{4.346}$$

The response of the circuit is clearly resonant, peaking at $\omega = 1/\sqrt{LC}$, and reaching $1/\sqrt{2}$ of the peak value at $\omega = 1/\sqrt{LC} \pm R/2L$ (assuming that $R \ll \sqrt{L/C}$). In fact, LCR circuits are used in radio tuners to filter out signals whose frequencies fall outside a given band.

The phase lag of the current with respect to the voltage is given by

$$\theta = \arg(Z) = \tan^{-1}\left(\frac{\omega L - 1/\omega C}{R}\right). \tag{4.347}$$

The phase lag varies from $-\pi/2$ for frequencies significantly below the resonant frequency, to zero at the resonant frequency ($\omega = 1/\sqrt{LC}$), to $\pi/2$ for frequencies significantly above the resonant frequency.

It is clear that in conventional AC circuits the circuit equation reduces to a simple algebraic equation, and the behaviour of the circuit is summed up by the impedance Z. The real part of Z tells us the power dissipated in the circuit, the magnitude of Z gives the ratio of the peak current to the peak voltage, and the argument of Z gives the phase lag of the current with respect to the voltage.

4.17 Transmission lines

The central assumption made in the analysis of conventional AC circuits is that the voltage (and, hence, the current) has the same phase throughout the circuit. Unfortunately, if the circuit is sufficiently large and the frequency of oscillation ω is sufficiently high then this assumption becomes invalid. The assumption of a constant phase throughout the circuit is reasonable if the wavelength of the oscillation $\lambda = 2\pi\, c/\omega$ is much larger than the dimensions of the circuit. This is generally not the case in electrical circuits which are associated with *communication*. The frequencies in such circuits tend to be high and the dimensions are, almost by definition, large. For instance, leased telephone lines (the type you attach computers to) run at 56 kHz. The corresponding wavelength is about 5 km, so the constant phase approximation clearly breaks down for long distance calls. Computer networks generally run at about 10 MHz, corresponding to $\lambda \sim 30$ m. Thus, the constant phase approximation also breaks down for the computer network in this building, which is certainly longer than 30 m. It turns out that you need a special sort of wire, called a transmission line, to propagate signals around circuits whose dimensions greatly exceed the wavelength λ. Let us investigate transmission lines.

An idealized transmission line consists of two parallel conductors of uniform cross-sectional area. The conductors possess a capacitance per unit length C, and an inductance per unit length L. Suppose that x measures the position along the line.

Consider the voltage difference between two neighbouring points on the line, located at positions x and $x+\delta x$, respectively. The self-inductance of the portion of the line lying between these two points is $L\,\delta x$. This small section of the line can be thought of as a conventional inductor, and therefore obeys the well-known equation

$$V(x,t) - V(x+\delta x,t) = L\,\delta x\,\frac{\partial I(x,t)}{\partial t}, \tag{4.348}$$

where $V(x,t)$ is the voltage difference between the two conductors at position x and time t, and $I(x,t)$ is the current flowing in one of the conductors at position x and time t [the current flowing in the other conductor is $-I(x,t)$]. In the limit $\delta x \rightarrow 0$, the above equation reduces to

$$\frac{\partial V}{\partial x} = -L\,\frac{\partial I}{\partial t}. \tag{4.349}$$

Consider the difference in current between two neighbouring points on the line, located at positions x and $x+\delta x$, respectively. The capacitance of the portion of the line lying between these two points is $C\,\delta x$. This small section of the line can be thought of as a conventional capacitor, and therefore obeys the well-known equation

$$\int_0^t I(x,t)\,dt - \int_0^t I(x+\delta x,t) = C\,\delta x\,V(x,t), \tag{4.350}$$

where $t=0$ denotes a time at which the charge stored in either of the conductors in the region x to $x+\delta x$ is zero. In the limit $\delta x \rightarrow 0$, the above equation yields

$$\frac{\partial I}{\partial x} = -C\,\frac{\partial V}{\partial t}. \tag{4.351}$$

Equations (4.349) and (4.351) are generally known as the "telegrapher's equations," since an old fashioned telegraph line can be thought of as a primitive transmission line (telegraph lines consist of a single wire; the other conductor is the Earth.)

Differentiating Eq. (4.349) with respect to x, we obtain

$$\frac{\partial^2 V}{\partial x^2} = -L\,\frac{\partial^2 I}{\partial x \partial t}. \tag{4.352}$$

Differentiating Eq. (4.351) with respect to t yields

$$\frac{\partial^2 I}{\partial x \partial t} = -C \frac{\partial^2 V}{\partial t^2}. \tag{4.353}$$

The above two equations can be combined to give

$$LC \frac{\partial^2 V}{\partial t^2} = \frac{\partial^2 V}{\partial x^2}. \tag{4.354}$$

This is clearly a wave equation with wave velocity $v = 1/\sqrt{LC}$. An analogous equation can be written for the current I.

Consider a transmission line which is connected to a generator at one end ($x = 0$) and a resistor R at the other ($x = l$). Suppose that the generator outputs a voltage $V_0 \cos \omega t$. If follows that

$$V(0, t) = V_0 \cos \omega t. \tag{4.355}$$

The solution to the wave equation (4.354), subject to the above boundary condition, is

$$V(x, t) = V_0 \cos(\omega t - kx), \tag{4.356}$$

where $k = \omega \sqrt{LC}$. This clearly corresponds to a wave which propagates from the generator towards the resistor. Equations (4.349) and (4.356) yield

$$I(x, t) = \frac{V_0}{\sqrt{L/C}} \cos(\omega t - kx). \tag{4.357}$$

For self-consistency, the resistor at the end of the line must have a particular value;

$$R = \frac{V(l, t)}{I(l, t)} = \sqrt{\frac{L}{C}}. \tag{4.358}$$

The so-called "input impedance" of the line is defined

$$Z_{\rm in} = \frac{V(0, t)}{I(0, t)} = \sqrt{\frac{L}{C}}. \tag{4.359}$$

Thus, a transmission line terminated by a resistor $R = \sqrt{L/R}$ acts very much like a conventional resistor $R = Z_{\rm in}$ in the circuit containing the generator. In fact, the transmission line could be replaced by an effective resistor $R = Z_{\rm in}$ in the circuit diagram for the generator circuit. The power loss due to this effective resistor corresponds to power which is extracted from the circuit, transmitted down the line, and absorbed by the terminating resistor.

The most commonly occurring type of transmission line is a co-axial cable, which consists of two co-axial cylindrical conductors of radii a and b (with $b > a$). We have already shown that the capacitance per unit length of such a cable is (see Section 4.5)

$$C = \frac{2\pi\epsilon_0}{\ln(b/a)}. \tag{4.360}$$

Let us now calculate the inductance per unit length. Suppose that the inner conductor carries a current I. According to Ampère's law, the magnetic field in the region between the conductors is given by

$$B_\theta = \frac{\mu_0 I}{2\pi r}. \tag{4.361}$$

The flux linking unit length of the cable is

$$\Phi = \int_a^b B_\theta \, dr = \frac{\mu_0 I}{2\pi} \ln(b/a). \tag{4.362}$$

Thus, the self-inductance per unit length is

$$L = \frac{\Phi}{I} = \frac{\mu_0}{2\pi} \ln(b/a). \tag{4.363}$$

The speed of propagation of a wave down a co-axial cable is

$$v = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\epsilon_0\mu_0}} = c. \tag{4.364}$$

Not surprisingly, the wave (which is a type of electromagnetic wave) propagates at the speed of light. The impedance of the cable is given by

$$Z_0 = \sqrt{\frac{L}{C}} = \left(\frac{\mu_0}{4\pi^2\epsilon_0}\right)^{1/2} \ln\left(b/a\right) = 60 \, \ln\left(b/a\right) \text{ ohms}. \tag{4.365}$$

We have seen that if a transmission line is terminated by a resistor whose resistance R matches the impedance Z_0 of the line then all the power sent down the line is absorbed by the resistor. What happens if $R \neq Z_0$? The answer is that some of the power is reflected back down the line. Suppose that the beginning of the line lies at $x = -l$ and the end of the line is at $x = 0$. Let us consider a solution

$$V(x,t) = V_0 \exp[\mathrm{i}\,(\omega t - kx)] + K V_0 \exp[\mathrm{i}\,(\omega t + kx)]. \tag{4.366}$$

This corresponds to a voltage wave of amplitude V_0 which travels down the line and is reflected, with reflection coefficient K, at the end of the line. It is easily demonstrated from the telegrapher's equations that the corresponding current waveform is

$$I(x,t) = \frac{V_0}{Z_0} \exp[\mathrm{i}\,(\omega t - kx)] - \frac{K V_0}{Z_0} \exp[\mathrm{i}\,(\omega t + kx)]. \tag{4.367}$$

Since the line is terminated by a resistance R at $x = 0$ we have, from Ohm's law,

$$\frac{V(0,t)}{I(0,t)} = R. \tag{4.368}$$

This yields an expression for the coefficient of reflection,

$$K = \frac{R - Z_0}{R + Z_0}. \tag{4.369}$$

The input impedance of the line is given by

$$Z_{\rm in} = \frac{V(-l,t)}{I(-l,t)} = Z_0 \frac{R\cos kl + \mathrm{i}\, Z_0 \sin kl}{Z_0 \cos kl + \mathrm{i}\, R \sin kl}. \tag{4.370}$$

Clearly, if the resistor at the end of the line is properly matched, so that $R = Z_0$, then there is no reflection (*i.e.*, $K = 0$), and the input impedance of the line is Z_0. If the line is short circuited, so that $R = 0$, then there is total reflection at the end of the line (*i.e.*, $K = -1$), and the input impedance becomes

$$Z_{\rm in} = \mathrm{i}\, Z_0 \tan kl. \tag{4.371}$$

This impedance is purely imaginary, implying that the transmission line absorbs no net power from the generator circuit. In fact, the line acts rather like a pure

inductor or capacitor in the generator circuit (*i.e.*, it can store, but cannot absorb, energy). If the line is open circuited, so that $R \rightarrow \infty$, then there is again total reflection at the end of the line (*i.e.*, $K = 1$), and the input impedance becomes

$$Z_{\rm in} = {\rm i}\, Z_0 \tan(kl - \pi/2). \tag{4.372}$$

Thus, the open circuited line acts like a closed circuited line which is shorter by one quarter of a wavelength. For the special case where the length of the line is exactly one quarter of a wavelength (*i.e.*, $kl = \pi/2$), we find

$$Z_{\rm in} = \frac{{Z_0}^2}{R}. \tag{4.373}$$

Thus, a quarter wave line looks like a pure resistor in the generator circuit. Finally, if the length of the line is much less than the wavelength (*i.e.*, $kl \ll 1$) then we enter the constant phase regime, and $Z_{\rm in} \simeq R$ (*i.e.*, we can forget about the transmission line connecting the terminating resistor to the generator circuit).

Suppose that we want to build a radio transmitter. We can use a half wave antenna to emit the radiation. We know that in electrical circuits such an antenna acts like a resistor of resistance 73 ohms (it is more usual to say that the antenna has an impedance of 73 ohms). Suppose that we buy a 500 kW generator to supply the power to the antenna. How do we transmit the power from the generator to the antenna? We use a transmission line, of course. (It is clear that if the distance between the generator and the antenna is of order the dimensions of the antenna (*i.e.*, $\lambda/2$) then the constant phase approximation breaks down, so we have to use a transmission line.) Since the impedance of the antenna is fixed at 73 ohms we need to use a 73 ohm transmission line (*i.e.*, $Z_0 = 73$ ohms) to connect the generator to the antenna, otherwise some of the power we send down the line is reflected (*i.e.*, not all of the power output of the generator is converted into radio waves). If we wish to use a co-axial cable to connect the generator to the antenna, then it is clear from Eq. (4.365) that the radii of the inner and outer conductors need to be such that $b/a = 3.38$.

Suppose, finally, that we upgrade our transmitter to use a full wave antenna (*i.e.*, an antenna whose length equals the wavelength of the emitted radiation). A full wave antenna has a different impedance than a half wave antenna. Does this mean that we have to rip out our original co-axial cable and replace it by

one whose impedance matches that of the new antenna? Not necessarily. Let Z_0 be the impedance of the co-axial cable, and Z_1 the impedance of the antenna. Suppose that we place a quarter wave transmission line (*i.e.*, one whose length is one quarter of a wavelength) of characteristic impedance $Z_{1/4} = \sqrt{Z_0 Z_1}$ between the end of the cable and the antenna. According to Eq. (4.373) (with $Z_0 \rightarrow \sqrt{Z_0 Z_1}$ and $R \rightarrow Z_1$) the input impedance of the quarter wave line is $Z_{\rm in} = Z_0$, which matches that of the cable. The output impedance matches that of the antenna. Consequently, there is no reflection of the power sent down the cable to the antenna. A quarter wave line of the appropriate impedance can easily be fabricated from a short length of co-axial cable of the appropriate b/a.

4.18 Epilogue

Unfortunately, our investigation of the many and varied applications of Maxwell's equations must now come to and end, since we have run out of time. Many important topics have been skipped in this course. For instance, we have hardly mentioned the interaction of electric and magnetic fields with matter. It turns out that atoms polarize in the presence of electric fields. Under many circumstances this has the effect of increasing the effective permittivity of space; *i.e.*, $\epsilon_0 \rightarrow \epsilon\epsilon_0$, where $\epsilon > 1$ is called the *relative permittivity* or *dielectric constant* of matter. Magnetic materials (*e.g.*, iron) develop net magnetic moments in the presence of magnetic fields. This has the effect of increasing the effective permeability of space; *i.e.*, $\mu_0 \rightarrow \mu\mu_0$, where $\mu > 1$ is called the *relative permeability* of matter. More interestingly, matter can reflect, transmit, absorb, or effectively slow down, electromagnetic radiation. For instance, long wavelength radio waves are reflected by charged particles in the ionosphere. Short wavelength waves are not reflected and, therefore, escape to outer space. This explains why it is possible to receive long wavelength radio transmissions when the transmitter is over the horizon. This is not possible at shorter wavelengths. For instance, to receive FM or TV signals the transmitter must be in the line of sight (this explains the extremely local coverage of such transmitters). Another fascinating topic is the generation of extremely short wavelength radiation, such as microwaves and radar. This is usually done by exciting electromagnetic standing waves in conducting cavities, rather than by using antennas. Finally, we have not men-

tioned relativity. It turns out, somewhat surprisingly, that Maxwell's equations are invariant under the Lorentz transformation. This is essentially because magnetism is an intrinsically relativistic phenomenon. In relativistic notation the whole theory of electromagnetism can be summed up in just two equations.