Notes for PX263, Electromagnetic Theory and Optics

Nicholas d'Ambrumenil

adapted from earlier notes by Tom Marsh

Updated: May 15, 2022

Contents

1 Introduction and Background 2

1.1 Grad, div and curl 4

1.2 Background physics 9

1.3 The laws 10

2 Maxwell's equations in free space 13

2.1 Derivation of Maxwell's equations from the integral form of the equations 14

2.2 Summary of the equations in free space 19

2.3 Special cases–electrostatic potentials 20

3 Electromagnetic waves in free space 23

3.1 EM wave solutions to Maxwell's equations 23

3.2 Energy in an EM wave 26

4 Maxwell's equations in matter 28

4.1 Polarisation charges and currents 29

4.2 Magnetisation 32

4.3 Maxwell's equations in matter 35

4.4 Boundary conditions on EM fields 39

5 Waves and energy density 44

5.1 EM Waves in dielectrics 44

5.2 Energy density and flux:the Poynting vector II 46

5.2.1 Poynting flux in EM waves in dielectrics 47

5.2.2 Poynting vector in static cases 49

5.3 Radiation pressure 50

6 EM Waves in Matter with Boundaries 51

6.1 Reflection and transmission at dielectric interfaces: Fresnel coefficients . . . 51

6.2 Electromagnetic waves in conductors . . . 58

6.2.1 The skin effect . . . 59

6.2.2 Reflections from conductors . . . 62

7 Geometric Optics 63

7.1 Wavefronts and rays . . . 63

7.2 Fermat's principle of "least time" . . . 64

7.3 Imaging by a spherical interface . . . 68

7.4 Thin lenses . . . 70

7.4.1 Principal rays for visualising lens imaging . . . 71

7.4.2 Ray diagrams for lenses . . . 72

8 Wave Optics 78

8.1 The resolution of optical instruments . . . 78

8.2 Thin film interference . . . 81

8.3 Fabry-Pérot etalons . . . 87

8.4 Conclusion . . . 90

A Index expressions 91

A.1 An introduction to index expressions . . . 91

A.2 Scalar products and the summation convention . . . 92

A.3 Cross products . . . 92

B Grad, div and curl in non-Cartesian coordinates 95

B.1 Introduction . . . 95

B.2 The gradient . . . 96

B.3 The divergence and Laplacian 96
B.4 Curl 98
B.5 Cylindrical coordinates 99

C Electric field of a uniformly polarised sphere **100**

D Potentials **103**
D.1 Some more electrostatic potential problems 103
D.2 The magnetic vector potential 106

E EM and Hamiltonian mechanics **109**

F Polarisation and magnetisation again **111**

G Non-isotropic media, birefringence **113**

H Momentum in EM fields **117**

I Electromagnetism to geometrical optics **119**
I.1 From EM to geometrical optics 119

J Elements of Huygens-Fresnel-Kirchoff diffraction theory **124**
J.1 Introduction 124
J.2 Huygens-Fresnel principle 124
J.3 Kirchoff's diffraction theory 125

K Exact analysis of thin film interference **129**

Chapter 1

Introduction and Background

The module develops the ideas of first year electricity and magnetism into Maxwell's theory of electromagnetism. Maxwell's equations pulled the various laws (Faraday's law, Ampere's law, Lenz's law, Gauss's law and the "law with no name") into one unified and elegant theory. Establishing a complete theory of electromagnetism has proved to be one the greatest achievements of physics. It provided motivation for Einstein to develop special relativity, it has served as the model for subsequent theories of the forces of nature and it has been the basis for all of electronics (radios, telephones, optical fibres, the lot...).

Maxwell's equations established the idea of fields in physics. The equations are local laws (differential equations). They were a major conceptual advance as previously things like Coulomb's law and the Newton's gravitational law implied "action at a distance" and put the emphasis on the force one object had on another. It was the development of Maxwell's theory that showed that it was the fields set up by objects that led to the forces on objects. These acted locally.

We will show that Maxwell's equations in free space have time-dependent solutions, which turn out to be the familiar electromagnetic (EM) waves (light, radio waves, X-rays etc). We will also show how the equations can be adapted to handle the presence of matter and how these affect the propagation of disturbances in EM fields. Finally we will look at optics as a practical example of electromagnetism.

This introductory chapter summarises the notation needed and the background knowledge assumed from modules that you will have taken. The notes for the first year module PX120 *Electricity and Magnetism* and the appropriate mathematics modules should be consulted if you are unsure of any of this content.

The main body of the notes should be taken as being all "examinable" (with very odd exceptions that will be flagged), while the appendices deal with some additional questions. The material of the appendices may not strictly form part of the syllabus for PX263, that's why they are appendices, but don't let that put you off looking at it: it should for the most part be understandable enough.

These notes are adapted from notes written by Tom Marsh. The errors are mine. Please let me know of any you find.

Physical quantities, Symbols & Units
Bold face indicates a vector, e.g. $\boldsymbol{E}$. The usual symbols are listed in Table 1.1, which follows

Quantity	**Symbol**	**Unit**	**SI**
Conductivity	g		$\mathrm{A\,V^{-1}\,m^{-1}} \equiv \Omega^{-1}\,\mathrm{m^{-1}} \equiv \mathrm{S\,m^{-1}}$
Current density	$\boldsymbol{J}$	$\mathrm{A\,m^{-2}}$	$\mathrm{A\,m^{-2}}$
Electric field	$\boldsymbol{E}$		$\mathrm{kg\,m\,s^{-2}\,C^{-1}} \equiv \mathrm{V\,m^{-1}}$
Displacement	$\boldsymbol{D}$		$\mathrm{C\,m^{-2}}$
Electric charge	q	Coulomb, C	A s
Linear charge density	λ		$\mathrm{C\,m^{-1}}$
Surface charge density	σ		$\mathrm{C\,m^{-2}}$
Volume charge density	ρ		$\mathrm{C\,m^{-3}}$
Electric potential	ψ	Volt, V	$\mathrm{kg\,m^2\,s^{-2}\,C^{-1}}$
Electromotive force	$\mathcal{E}$	Volt, V	$\mathrm{kg\,m^2\,s^{-2}\,C^{-1}}$
Electric flux	Φ_E		$\mathrm{kg\,m^2\,s^{-2}\,C^{-1}}$
Magnetic flux density	$\boldsymbol{B}$	Tesla, T	$\mathrm{kg\,s^{-1}\,C^{-1}}$
Magnetic field strength	$\boldsymbol{H}$		$\mathrm{A\,m^{-1}}$
Magnetic flux	Φ_B	Weber, W	$\mathrm{kg\,m^2\,s^{-1}\,C^{-1}}$
Magnetisation	$\boldsymbol{M}$		$\mathrm{A\,m^{-1}}$
Polarisation	$\boldsymbol{P}$		$\mathrm{C\,m^{-2}}$
Position	$\boldsymbol{r}$	metre, m	m
Potential energy	U	Joule, J	$\mathrm{kg\,m^2\,s^{-2}}$
Velocity	$\boldsymbol{v}$		$\mathrm{m\,s^{-1}}$
Vector potential	$\boldsymbol{A}$		$\mathrm{kg\,m\,s^{-2}\,A^{-1}}$

Table 1.1: Physical quantities, and the symbol usually used for them in these notes,

as far as possible the usage of the first year EM module, PX120. The number of symbols needed is such that there may be potential for confusion. In surface integrals $d\boldsymbol{S}$ indicates an element of a surface (rather than $d\boldsymbol{A}$). (There is potential confusion with a quantity called the "vector potential" which is always denoted by $\boldsymbol{A}$. While this quantity is not central this module, it is used widely when looking at the electrodynamics covered in year three and to the role of electromagnetic field in quantum theories.) Later we will loosely refer to two quantities as the magnetic field called, $\boldsymbol{B}$ and $\boldsymbol{H}$. When we need to distinguish between these in words we will refer to them as the "magnetic flux density" and "magnetic field strength" respectively.

Table 1.1 lists the SI unit for each quantity. Although it is the Ampère that is the electromagnetic unit that appears amongst the seven base units of the SI system, we will follow the PX120 notes where a similar table appears, in (mostly) using the unit of charge, the Coulomb, in the right hand column of the table. We will use the Ampère when it seems more natural. A valuable exercise is to convince yourself of the right-hand column for each symbol.

1.1 Grad, div and curl

Setting up Maxwell's equations needs a number of results from vector calculus including the divergence theorem and Stokes's theorem. We will revise these quickly first. You should have studied these either in PX275 Mathematical Methods for Physicists or in a combination of MA259 Multivariable Calculus and PX276 Methods of Mathematical Physics. In the following, locations will be denoted by the position vector $\boldsymbol{r}$, and if, as is sometimes needed, two locations are required, we will use $\boldsymbol{r}$ and $\boldsymbol{r}'$.

Consider a scalar field defined at every point over some region, $f(\boldsymbol{r})$. On changing location from $\boldsymbol{r}$ to $\boldsymbol{r} + d\boldsymbol{r}$, *i.e.* (x, y, z) to $(x + dx, y + dy, z + dz)$, f changes by

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz = \nabla f \cdot d\boldsymbol{r}, \tag{1.1}$$

where in Cartesian coordinates $\nabla f = (\partial_x f, \partial_y f, \partial_z f)$, using the short-hand notation ∂_x for $\partial/\partial x$. (We are assuming that f is differentiable, and will often do so implicitly throughout the module.) The symbol $\nabla = (\partial_x, \partial_y, \partial_z)$ is the "vector derivative" or "gradient" operator, sometimes called "del" or "nabla". The quantity df is the differential of f and is a linear function on the variables, dx, dy and dz.

We can write the rhs of 1.1 as $|\nabla f||d\boldsymbol{r}|\cos\theta$, with θ the angle between ∇f and $d\boldsymbol{r}$. For directions, $d\boldsymbol{r}$, pointing along contours of f (lines of constant f) the change in f is zero. This occurs when $\theta = \pi$. Hence ∇f is perpendicular to contours of f at any given point (see Figure 1.1). The maximum change of f, for a given step length $dr = |d\boldsymbol{r}|$, occurs when $\theta = 0$ $(\cos\theta = 1)$ and the step is parallel to ∇f. The rate of change of f with distance in this direction equals $|\nabla f|$. ∇f is often called the "gradient" of f and sometimes grad(f).

An example of a gradient in a physical context is the temperature gradient ∇T. This points from low to high temperature along the direction of maximum temperature change and $|\nabla T|$ is the rate at which temperature increases with distance along this direction. Heat conducts from high to low temperatures and, in an isotropic medium, the heat flux at any point is given by $-\kappa\nabla T$, where κ is the thermal conductivity. Fig. 1.1 illustrates the concept of the gradient of a scalar field. A final relation associated with gradients follows from $df = \nabla f \cdot d\boldsymbol{r}$. When integrated between two arbitrary points, A and B, it gives the difference in f evaluated at the two points

$$\Delta f = f(B) - f(A) = \int_A^B \nabla f \cdot d\boldsymbol{r}. \tag{1.2}$$

Divergence

The divergence of a vector field $\boldsymbol{W}$ is written div$(\boldsymbol{W}) = \nabla \cdot \boldsymbol{W}$. In terms of Cartesian components

$$\text{div}(\boldsymbol{W}) = \nabla \cdot \boldsymbol{W} = \frac{\partial W_x}{\partial x} + \frac{\partial W_y}{\partial y} + \frac{\partial W_z}{\partial z}. \tag{1.3}$$

The notation suggests thinking of this as a dot product between the operation ∇ and the vector field. It delivers a scalar (number) at each point.

The divergence of a vector field gives a measure of the outward flux of $\boldsymbol{W}$ created per unit volume. If we measure the outward flux through some closed surface, S, it will be the integral over the enclosed volume, V, of what is created inside. This is the content of Gauss's

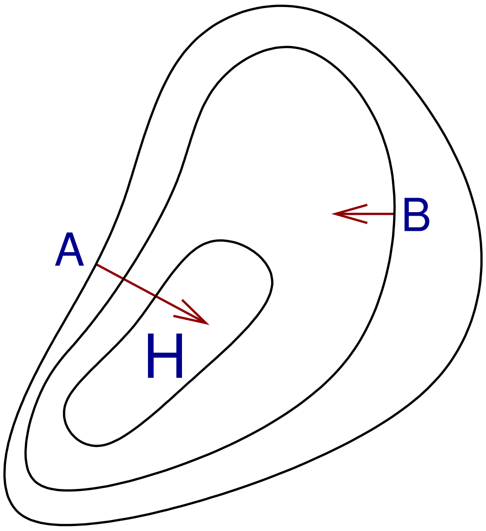

Figure 1.1: The closed loops marks contours of equal intensity of some scalar function. In a physical application these might be temperature or pressure and these lines would be called isotherms or isobars. "H" marks the highest point. Arrows at A and B indicate the direction of the gradients at these two points. They are perpendicular to the contours at these points towards the region of high values. There is a steeper gradient at A than B.

(divergence) theorem

$$\oint_S \boldsymbol{W} \cdot d\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{W}\, dV. \tag{1.4}$$

On the left-hand side is the integral of the outward flux $\boldsymbol{W}$ across the surface, while the right-hand side is the integral over the enclosed volume of what is "created" inside. (A negative value of the divergence would imply that flux is being absorbed or lost.) The circle on the left-hand integral is to show that it is over a "closed" surface. Only one integral sign is used as it should be clear that the left-hand integral is over a surface (in most cases involving EM fields this will be a 2-dimensional or double integral) while the volume integral on the right is 3-dimensional. The dimensionality of the integral is apparent from the domain of integration (S or V).

A pictorial explanation of Gauss's theorem is given in Fig. 1.2. Here a volume is divided into many small cuboids. The flux integral over the outer surface can be approximated as the sum of fluxes emergent from each of the small cuboids. This is because the fluxes emergent from the faces of adjacent cuboids cancel in pairs, leaving only the unpaired surface contributions. The right-hand diagram of Fig. 1.2 shows the flux coming out of one small cuboid (it should be imagined as having depth dz out of the page). The surface areas of the left- and right-hand faces are $dydz$. The flux emerging from the right-hand face $\approx W_x(x+dx,y,z)\,dy\,dz$, where $W_x(x+dx,y,z)$ is the x-component of a vector field $\boldsymbol{W}$ evaluated at the centre of the right-hand face. The flux entering at the left is similarly $\approx W_x(x,y,z)\,dy\,dz$, so the total flux emerging from the left- and right-hand faces is

$$\begin{aligned} &\approx \quad [W_x(x+dx,y,z) - W_x(x,y,z)]\;dy\,dz, \\ &\approx \quad \left(\frac{W_x(x+dx,y,z) - W_x(x,y,z)}{dx}\right) dx\,dy\,dz, \\ &\approx \quad \frac{\partial W_x}{\partial x}\,dV. \end{aligned} \tag{1.5}$$

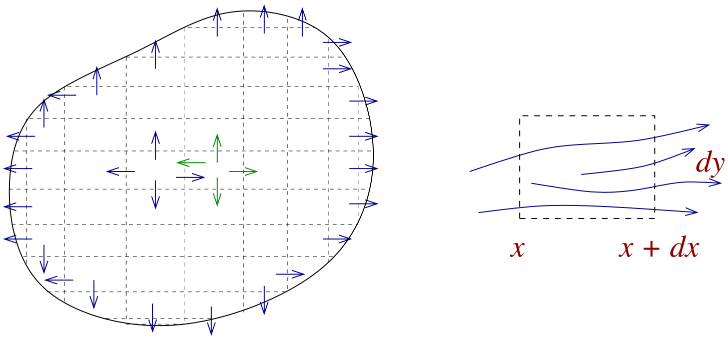

Figure 1.2: Gauss's theorem. Consider a volume sliced up into small cuboids. At the interface to the two central cuboids the fluxes shown in blue and green must have opposite signs. If a flux in the x-direction is emerging from an internal cuboid, see 1.5, it must enter the cuboid adjacent to it. There it is counted with the opposite sign as, if it is exiting one cuboid, it is entering the other. When summing over all cuboids (integrating $\nabla \cdot \boldsymbol{W}$ over the volume), the fluxes across internal faces cancel and we are left with the flux across all the boundaries of the cuboids with the outside region. The diagram on the right focuses upon the flux emergent from a single small cuboid—some of it enters from one side and leaves again (these contribute nothing to $\nabla \cdot \boldsymbol{W}$) while some originates there (and contribute to $\nabla \cdot \boldsymbol{W}$).

With two similar terms from the other pairs of faces, the flux emergent from a small cuboid of volume dV is thus $\approx (\nabla \cdot \boldsymbol{W})\, dV$. Summing over all cuboids in the infinitesimal limit we deduce that

$$\oint_S \boldsymbol{W} \cdot d\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{W}\, dV, \tag{1.6}$$

i.e. Gauss's theorem.

Curl

The curl of a vector field provides a sense of its rotation. The curl of a vector field is given in Cartesian coordinates by the following relations

$$\text{curl}(\boldsymbol{W}) = \nabla \times \boldsymbol{W} = \begin{vmatrix} \hat{\boldsymbol{e}}_x & \hat{\boldsymbol{e}}_y & \hat{\boldsymbol{e}}_z \\ \partial_x & \partial_y & \partial_z \\ W_x & W_y & W_z \end{vmatrix} = \begin{pmatrix} \partial_y W_z - \partial_z W_y \\ \partial_z W_x - \partial_x W_z \\ \partial_x W_y - \partial_y W_x \end{pmatrix}, \tag{1.7}$$

again using the short-hand form of the partial derivatives. Its interpretation as a measure of rotation comes from Stokes's theorem:

$$\oint_C \boldsymbol{W} \cdot d\boldsymbol{\ell} = \int_S \nabla \times \boldsymbol{W} \cdot d\boldsymbol{S}. \tag{1.8}$$

The integral on the left is a line integral around some circuit indicated by C. The integral on the right is a surface integral over some surface S bounded by C. Note that the integral on the right does not have a little circle because the surface S is not closed.

The curl of a vector field can be measured (in principle) by measuring the line integral around a tiny loop embedded in the field, and dividing by its area. One would need to do so in three orientations with the loop oriented perpendicular to the component (e.g. $\hat{\boldsymbol{x}}$) of interest. The curl of the vector field corresponding to solid-body rotation

$$\boldsymbol{v} = \boldsymbol{\omega} \times \boldsymbol{r}, \tag{1.9}$$

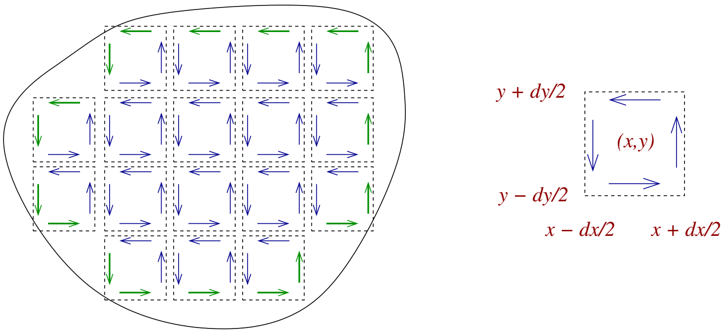

Figure 1.3: A pictorial explanation of Stokes' theorem. A surface spanning a circuit is divided up into small rectangles. The contributions to the line integrals around each of the rectangles cancel at all interior points, leaving just the parts around the edge high-lighted in green. The diagram on the right focusses upon the line integral contributions around one small rectangle of dimensions dx by dy. The traversal direction is anti-clockwise to give the positive z-component of curl.

is given by $\nabla \times \boldsymbol{v} = 2\boldsymbol{\omega}$, hence the association with rotation (you should check this for yourself).

Fig. 1.3 shows a graphical explanation of Stokes' theorem. The surface spanning a circuit is divided up into many small rectangles. The line integral contributions (indicated by the arrows) around each rectangle cancel in pairs at all points except at the outer edge where they align with the circuit. Thus the line integral of a vector field around edge equals the sum of all the small contributions from the rectangles which span the surface. In the limit, the latter is a surface integral. Focusing on the contribution from one small element (right-hand of Fig. 1.3), the line integral can be built from each side, starting from the bottom edge and proceeding anti-clockwise and taking the mid-point value along each edge:

$$W_x(x, y - dy/2, z)\, dx + W_y(x + dx/2, y, z)\, dy - \tag{1.10}$$

$$W_x(x, y + dy/2, z)\, dx - W_y(x - dx/2, y, z)\, dy. \tag{1.11}$$

Expanding to first order in dx and dy this gives

$$\left(\frac{\partial W_y}{\partial x} - \frac{\partial W_x}{\partial y} \right) dx\, dy = (\nabla \times \boldsymbol{W})_z\, dS = \nabla \times \boldsymbol{W} \cdot d\boldsymbol{S}, \tag{1.12}$$

recognising that the area vector corresponding to the small rectangle of Fig. 1.3 points in the positive z-direction. Stokes' theorem then follows.

Curl is a trickier concept than divergence because of the cross product, because it is a vector, and because there are some vector fields that appear to have a sense of rotation and yet have zero curl. An example of this is the magnetic field around a straight wire carrying current I which can be written

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0 I}{2\pi r} \hat{\boldsymbol{\theta}}, \tag{1.13}$$

where r is the perpendicular distance of the point from the wire and $\hat{\boldsymbol{\theta}}$ is a unit vector in the azimuthal direction around the wire. One normally draws the field lines as circles around the wire, but for this field, $\nabla \times \boldsymbol{B} = \boldsymbol{0}$ everywhere except $r = 0$ where it is undefined. One can show this by computing the components of curl. Alternatively we can consider Fig. 1.4 which

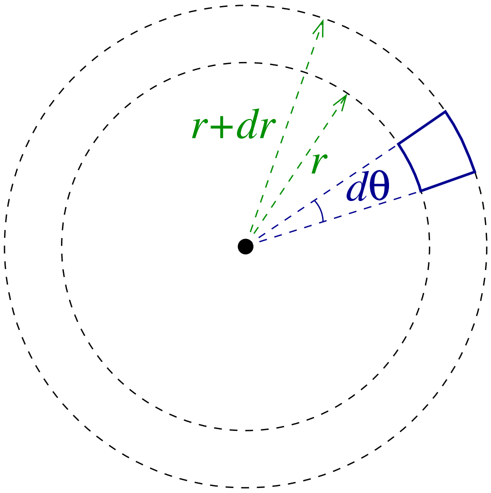

Figure 1.4: Circles of fixed radius around a wire carrying a current out of the page are used to define a small circuit (solid loop) around which Stokes' theorem can be applied to show that curl$(\boldsymbol{B}) = \mathbf{0}$. (See text.)

shows in solid blue a small circuit near a current-carrying wire to which Stokes's theorem can be applied. This will reflect the component of the curl of the magnetic field in the direction of the wire. (We know that the magnetic field runs parallel to the plane of the figure, and does not vary with z. The components of the curl in the xy-plane involve derivatives with respect to z and components of curl$(\boldsymbol{B})$ in the $z-$direction, both of which are zero.) If we consider the small circuit indicated, its straight sides run in a radial direction perpendicular to the magnetic field, so only the contributions along the circular sections are non-zero. Since the magnetic field runs parallel to the circular section, the line integral along these parts is the field strength times the arc lengths. Going around the circuit we obtain two contributions as follows

$$B(r+dr)(r+dr)\,d\theta - B(r)r\,d\theta = \frac{d(rB)}{dr}d\theta dr = \frac{1}{r}\frac{d(rB)}{dr}dS, \tag{1.14}$$

where $dS = r\,d\theta\,dr$ is the area of the small loop. To go from the left-hand to the middle expression, we have used the standard definition of a derivative.

For a small circuit, over which we can assume the curl to be be constant, the right-hand side of Stokes's theorem reduces to $(\nabla \times \boldsymbol{B}) \cdot \hat{\boldsymbol{n}}\,dS$. The component of curl parallel to the wire is therefore given by

$$\frac{1}{r}\frac{d(rB)}{dr}. \tag{1.15}$$

You may recognize this from mathematics modules as it appears in the equation for curl when written in cylindrical polar coordinates. In this particular case, since $B \propto 1/r$, even this component of curl reduces to zero, for $r > 0$ (it is undefined for $r = 0$). We have illustrated that it is possible for a vector field to appear to have a sense of rotation about it and yet have zero curl. Note though that any circuit *that encloses the wire* will lead to a non-zero line integral as it must to satisfy Ampère's law.

For further details of vector derivatives in non-Cartesian coordinates, see appendix B. While you don't need to know the details of how to derive the various formulae there, you could be asked to apply them, so please be aware that there are forms other than Cartesian for div, grad and curl.

Other Relations
Many relations can be derived for vectors and vector derivatives. A few of the more useful ones are as follows

$$\begin{aligned}
\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c}) &= (\boldsymbol{a}\cdot\boldsymbol{c})\boldsymbol{b}-(\boldsymbol{a}\cdot\boldsymbol{b})\boldsymbol{c} && (1.16)\\
\nabla\times(\nabla\times\boldsymbol{W}) &= \nabla(\nabla\cdot\boldsymbol{W})-\nabla^2\boldsymbol{W} && (1.17)\\
\nabla\cdot(f\boldsymbol{W}) &= (\nabla f)\cdot\boldsymbol{W}+f\nabla\cdot\boldsymbol{W} && (1.18)\\
\nabla\times(f\boldsymbol{W}) &= (\nabla f)\times\boldsymbol{W}+f\nabla\times\boldsymbol{W}. && (1.19)
\end{aligned}$$

For two vector field, $\boldsymbol{V}$ and $\boldsymbol{W}$,

$$\nabla\cdot(\boldsymbol{V}\times\boldsymbol{W})=\boldsymbol{W}\cdot(\nabla\times\boldsymbol{V})-\boldsymbol{V}\cdot(\nabla\times\boldsymbol{W}), \tag{1.20}$$

When learning vector calculus, just as for all techniques, you should practise. Proving as many of these as you can is a good exercise. See Appendix A for details of index notation which can help in such proofs (like all appendices this covers material beyond what is needed for the module, but may be of interest, and will come up in other modules). Appendix B looks at expressions for ∇f, $\nabla\cdot\boldsymbol{W}$ etc in non-Cartesian coordinates. PX263 is about physics not mathematical techniques, and relationships of this form of any complexity will be given in the rubric of examinations papers for this module.

1.2 Background physics

We need to revise the physics from the first year. This is a reference, not a complete exposition of this background; see the PX120 notes for further information.

The Lorentz force
The force acting on a charge q moving at velocity $\boldsymbol{v}$ in an electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$ is given by

$$\boldsymbol{F}=q\left(\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B}\right), \tag{1.21}$$

which is known as the Lorentz force. The Lorentz force encapsulates what we mean by "electric and magnetic fields": if a stationary charge accelerates, an electric field must be present; if its path starts to curve as it starts to move, there is a magnetic field as well. Always remember that it is $\boldsymbol{E}$ and $\boldsymbol{B}$ that appear in the Lorentz force, not the related variants $\boldsymbol{D}$ and $\boldsymbol{H}$ that we will also encounter.

Coulomb's Law
The electric field at position vector $\boldsymbol{r}$ from a stationary point charge q:

$$\boldsymbol{E}(\boldsymbol{r})=\frac{q\,\hat{\boldsymbol{r}}}{4\pi\epsilon_0 r^2}=\frac{q\,\boldsymbol{r}}{4\pi\epsilon_0 r^3}, \tag{1.22}$$

where $\hat{\boldsymbol{r}}=\boldsymbol{r}/r$ is a unit vector parallel to $\boldsymbol{r}$ and $r=|\boldsymbol{r}|$.

If instead positions are measured with respect to an arbitrary origin rather than the charge itself, then Coulomb's law appears as

$$\boldsymbol{E}(\boldsymbol{r})=\frac{q\left(\boldsymbol{r}-\boldsymbol{r}'\right)}{4\pi\epsilon_0|\boldsymbol{r}-\boldsymbol{r}'|^3}, \tag{1.23}$$

where $\boldsymbol{r}'$ is the position of the charge.

If charge is distributed over a volume with charge density ρ, then a small element of volume dV contains charge $dq = \rho\, dV$, and the electric field at $\boldsymbol{r}$ due to the charges can be calculated as

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r}')(\boldsymbol{r} - \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|^3}\, dV', \tag{1.24}$$

where the dash on dV' indicates an integration over all positions $\boldsymbol{r}'$ within the volume containing charges.

Coulomb potential
In static cases (all time derivatives $\partial_t \equiv 0$), electric fields can derived from potentials via

$$\boldsymbol{E} = -\nabla\psi, \tag{1.25}$$

where ψ, the potential, is a scalar function of position. The potential corresponding to the Coulomb field from a charge q located at the origin is

$$\psi(\boldsymbol{r}) = \frac{q}{4\pi\epsilon_0 r}, \tag{1.26}$$

when $\boldsymbol{r}$ is measured from the origin. If the charge is not at the origin, but is located at position $\boldsymbol{r}'$, then

$$\psi(\boldsymbol{r}) = \frac{q}{4\pi\epsilon_0 |\boldsymbol{r} - \boldsymbol{r}'|}. \tag{1.27}$$

A distribution of charge, with the charge in small volume dV' at position $\boldsymbol{r}'$ represented by $\rho(\boldsymbol{r}')\, dV'$ where ρ is the charge density, will give a total potential at position $\boldsymbol{r}$ of

$$\psi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, dV'. \tag{1.28}$$

Another important potential is the dipole potential, which can be thought of as the potential due to two charges of $\pm q$ separated by $\boldsymbol{\epsilon}$ in the limit $|\boldsymbol{\epsilon}| \to 0$ with the dipole moment $\boldsymbol{p} = q\boldsymbol{\epsilon}$ held constant. Such a dipole located at $\boldsymbol{r}'$ contributes a potential at $\boldsymbol{r}$ given by

$$\psi(\boldsymbol{r}) = \frac{\boldsymbol{p} \cdot (\boldsymbol{r} - \boldsymbol{r}')}{4\pi\epsilon_0 |\boldsymbol{r} - \boldsymbol{r}'|^3}, \tag{1.29}$$

or in spherical polar coordinates when oriented along the z-axis and centred on the origin

$$\psi(r, \theta) = \frac{p\cos\theta}{4\pi\epsilon_0 r^2}. \tag{1.30}$$

1.3 The laws

Gauss's law states:

$$\Phi_E = \oint_S \boldsymbol{E} \cdot d\boldsymbol{S} = \frac{Q}{\epsilon_0} = \frac{1}{\epsilon_0} \int_V \rho\, dV. \tag{1.31}$$

It can be derived for the case of stationary charges from Coulomb's law, if we assume that we can superpose the electric fields due to many point charges. However, Gauss's law also applies in the presence of moving charges while Coulomb's law needs adapting to describe this case. The integral on the left is over an arbitrary closed surface with vector elements of area denoted by $d\boldsymbol{S}$. Q is the total charge enclosed by the surface and the integral on the right is over the volume enclosed by the surface with ρ the charge density (charge per unit volume) and the volume elements denoted by dV. The left-hand integral is the *electric flux* coming out of the surface.

Gauss's law can be used to derive electric fields in cases of high symmetry. Three cases are:

(a) **Spherical Symmetry** The electric field outside a spherically-symmetric distribution of (total) charge q centred on the origin:

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{q\,\hat{\boldsymbol{r}}}{4\pi\epsilon_0 r^2}; \tag{1.32}$$

(b) **Cylindrical Symmetry** The magnitude of the electric field a distance r from an infinite line of charge of linear density λ:

$$E(r) = \frac{\lambda}{2\pi\epsilon_0 r}; \quad \text{and} \tag{1.33}$$

(c) **Mirror Plane Symmetry** The electric field either side of an infinite sheet of charge with surface density σ:

$$E = \frac{\sigma}{2\epsilon_0}. \tag{1.34}$$

The Biot-Savart law

The Biot-Savart law gives the magnetic field at position vector $\boldsymbol{r}$ from a charge q moving at (constant) velocity $\boldsymbol{v}$

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} q \frac{\boldsymbol{v} \times \hat{\boldsymbol{r}}}{r^2}. \tag{1.35}$$

In the same way that Coulomb's law can be extended to a charge distribution, the Biot-Savart law can be extended to a current distribution:

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{r}') \times (\boldsymbol{r} - \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|^3}\, dV', \tag{1.36}$$

or equivalently to a current I flowing in a wire where it becomes a line integral

$$\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi} \int \frac{d\boldsymbol{l} \times (\boldsymbol{r} - \boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|^3}. \tag{1.37}$$

Ampère's law

Ampère's law follows from the Biot-Savart law

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{\ell} = \mu_0 I = \mu_0 \int_S \boldsymbol{J} \cdot d\boldsymbol{S}. \tag{1.38}$$

Here, the left-hand line integral is around a closed loop that encloses the surface that is integrated over on the right-hand side, I is the electric current enclosed by the loop while $\boldsymbol{J}$ is the current density (current per unit area).

Like Gauss's law, Ampère's law can yield the magnetic field in situations of high symmetry. The field distance r from an infinite line of current I can be shown to be

$$B(r) = \frac{\mu_0 I}{2\pi r}, \tag{1.39}$$

while the field inside a long solenoid with N turns per unit length and carrying a current I is

$$B = \mu_0 N I. \tag{1.40}$$

The law with no name

The Biot-Savart law, Eq. 1.35, shows that the magnetic field lines due to a moving charge take the form of closed circles perpendicular to its velocity vector. Since the field lines never start or stop, the magnetic flux emerging from a closed surface is always zero. This is sometimes called the "solenoidal condition" and in integral form says

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{S} = 0. \tag{1.41}$$

for any circuit C. It leads to the only one of Maxwell's equations without the name of some past luminary attached to it, hence the "law with no name", but you can call it the solenoidal condition if you prefer.

The Faraday–Lenz law of induction

If the magnetic flux through a circuit changes an EMF ("electro-motive force") is generated. The EMF is what we would call a "voltage", and is the line integral of the electric field. The voltage generated is proportional to the rate of change of magnetic flux through the circuit (Faraday's law), and it has a direction such that if a current can flow, it opposes the change in flux (Lenz). This is all expressed mathematically as

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{\ell} = -\frac{d\Phi_B}{dt} = -\frac{d}{dt}\int_S \boldsymbol{B} \cdot d\boldsymbol{S}. \tag{1.42}$$

A remarkable feature of the Faraday-Lenz law is that it applies whether the magnetic flux changes simply due to alterations in the field or because of alterations in the circuit C itself.

Chapter 2

Maxwell's equations in free space

The equations of electromagnetism in integral form—Gauss's law, Ampère's law, etc—are general, but when it comes to solving for specific fields the integral forms are not usually a good starting point. Coulomb's law and the Biot-Savart law can be written to allow for arbitrary distributions of charge and current (Eqs. 1.24 and 1.36), but we don't always know what these are in advance. For instance in electrostatics, conductors impose boundary conditions of constant potential, but not a defined charge distribution since the latter is a response of whatever other charges are present. The route to solving such cases is through the partial differential equations known as Maxwell's equations.

In this chapter, we will work through Maxwell's formulation of the laws of electromagnetism in free space. Maxwell found that a term was missing in what had been found beforehand. His formulation of the laws as differential equations (including the missing term which we will derive) are as fundamental to physics as Newton's laws are to classical mechanics. It turned out later that these equations also included relativistic effects correctly and, at the level of wavefunctions for massless bosonic particles (photons in this case), were actually a quantum theory.

Conservation Law
One result we will need is the statement in differential form of the conservation of charge. This is the statement that, if charge is not being created, the net current out of (or into) an arbitrary volume must equal the rate at which the charge inside the volume decreases (or increases).

Consider the total charge in an arbitrary volume

$$Q = \int_V \rho \, dV. \tag{2.1}$$

This changes if there is any net current flow into or out of the volume which gives

$$\frac{dQ}{dt} = \frac{d}{dt} \int_V \rho \, dV = - \oint_S \boldsymbol{J} \cdot d\boldsymbol{S}, \tag{2.2}$$

and is the integral expression of charge conservation. Application of the divergence theorem 1.4 gives

$$\int_V dV \left(\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} \right) = 0. \tag{2.3}$$

Equation 2.3 is true for regions of any shape. This is only possible if the integrand itself is everywhere zero. One can see why this might be true by imagining shrinking the region of integration until it is a tiny region within which $\nabla \cdot \boldsymbol{J}$ and ρ are essentially constant. We can therefore write the continuity equation for electric charge :

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0.} \tag{2.4}$$

The continuity equation 2.4 has the interpretation that a non-zero divergence in $\boldsymbol{J}$ at any point is balanced by a changing local charge density. If $\nabla \cdot \boldsymbol{J}$ is positive, charge is flowing away from the vicinity and the charge density must reduce with time.

Similar relations apply to other cases with conserved quantities with ρ and $\boldsymbol{J}$ replaced by the corresponding densities and fluxes. For fluid flow $\partial_t \rho + \nabla \cdot (\rho \boldsymbol{v}) = 0$, where ρ is the mass density and $\boldsymbol{v}$ the flow velocity, is that statement that mass is conserved. In quantum mechanics, $\boldsymbol{J}$ is the probability current and ρ the probability density, and the relation is the statement that probability is conserved. For the case of particles with charge q and number density n moving at velocity $\boldsymbol{v}$, the electrical current density

$$\boldsymbol{J} = n\, q\, \boldsymbol{v} \equiv \rho\, \boldsymbol{v}. \tag{2.5}$$

To see why this is true, think of a small area, δA, through which the particles are flowing at normal incidence with speed, v. In time δt, $n\, \delta A\, v\, \delta t$ particles cross the surface. The charge, δQ, crossing δA is $\delta Q = q\, n\, \delta A\, v\, \delta t$. The current density is then $J = \delta Q/(\delta A\, \delta t) = n\, q\, v = \rho\, \boldsymbol{v}$ consistent with 2.5[1].

We will see that the conservation law, 2.4, has to be part of the laws of electromagnetism. In fact, Maxwell used this insight to establish the correct form for the equations.

2.1 Derivation of Maxwell's equations from the integral form of the equations

Each of the integral equations can be transformed into a partial differential equation.

GAUSS's LAW
The integral on the left-hand side of Gauss's (physical) *law*, see Equation 1.31,

$$\oint_S \boldsymbol{E} \cdot d\boldsymbol{S} = \frac{1}{\epsilon_0} \int_V \rho\, dV, \tag{2.6}$$

gives the electric flux emergent from the volume of interest. It can be transformed using Gauss's *theorem* to give

$$\oint_S \boldsymbol{E} \cdot d\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{E}\, dV = \frac{1}{\epsilon_0} \int_V \rho\, dV. \tag{2.7}$$

[1] A possible source of confusion with 2.5 relates to the case of currents in materials. Materials are usually neutral, $\rho = 0$, so how do materials carry current? Well, in materials there is more than one sort of charged particle, namely electrons and ions. We should therefore write $\boldsymbol{J} = \sum_i n_i\, q_i\, \boldsymbol{v}_i \equiv \sum_i \rho_i\, \boldsymbol{v}_i$, where the sum is over the different types of particle. In metals, for example, the charge densities cancel, $\rho_{ion} + \rho_e = 0$. Only the electrons move ($\boldsymbol{v}_{ion} = 0$), so $\boldsymbol{J} = \rho_e\, \boldsymbol{v}_e$. We can get away with using 2.5 provided we remember that ρ and $\boldsymbol{v}$ are the values for electrons.

The two volume integrals are over the same region indicated by V so we can write

$$\int_V \left(\nabla \cdot \boldsymbol{E} - \frac{\rho}{\epsilon_0} \right) dV = 0. \tag{2.8}$$

This is only possible if the integrand in 2.8 itself is everywhere zero (using the same argument as we used to derive 2.4). We can conclude that

$$\boxed{\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}.} \tag{2.9}$$

Equation 2.9 is the differential form of Gauss's law and the first of Maxwell's equations. It says that the divergence of the electric field at any point is proportional to the electric charge density at that point.

SOLENOIDAL CONDITION on $\boldsymbol{B}$

Magnetic flux lines in free space are always closed loops; flux is neither created nor destroyed. We should take this as an empirical fact (no sources of field lines, the magnetic counterparts to charges, have been observed). This leads to the solenoidal condition (Eq. 1.41)

$$\oint_S \boldsymbol{B} \cdot d\boldsymbol{S} = 0. \tag{2.10}$$

Following the same procedure as in the previous section we deduce the second of Maxwell's equations

$$\boxed{\nabla \cdot \boldsymbol{B} = 0.} \tag{2.11}$$

The magnetic field is divergence-less. Put another way, there are no magnetic charges (or "monopoles") to produce or consume magnetic flux. Magnetic monopoles could be included in this equation by introducing a term $\mu_0 \rho_m$ on the right hand side of Equation 2.11. As no monopoles have been observed we take $\rho_m = 0$.

FARADAY-LENZ LAW of INDUCTION

Faraday's law in integral form (with Lenz's law for the sign) is

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{\ell} = -\frac{d}{dt} \int_S \boldsymbol{B} \cdot d\boldsymbol{S}. \tag{2.12}$$

Using Stokes's theorem, the left-hand line integral can be written

$$\oint_C \boldsymbol{E} \cdot d\boldsymbol{\ell} = \int_S \nabla \times \boldsymbol{E} \cdot d\boldsymbol{S}, \tag{2.13}$$

and thus

$$\int_S \nabla \times \boldsymbol{E} \cdot dS = -\frac{d}{dt} \int_S \boldsymbol{B} \cdot d\boldsymbol{S}. \tag{2.14}$$

The integral on each side is over the same surface S, so, as before, we can write

$$\int_S \left(\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} \right) \cdot d\boldsymbol{S} = 0. \tag{2.15}$$

We are restricting ourselves to *fixed* surfaces S and boundary circuits C so that the magnetic flux only changes because of changes in the magnetic field at fixed locations rather than because of any motion of the circuit, C, through the field. If the loop, C, were moving then part of the EMF would be generated through the Lorentz force which we are not tracking here[2]. This is why we switch from the ordinary derivative (in 2.14) to the partial derivative in 2.15 with respect to time. The surface is stationary, but otherwise arbitrary. We can conclude that the integrand must vanish (the result in 2.15 has to hold for every circuit, C, including infinitesimally small ones inside and around which the integrand is constant) and deduce the differential version of the Faraday–Lenz law and the third of Maxwell's equations:

$$\boxed{\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}.} \tag{2.16}$$

In words, Equation 2.16 states that the curl of the electric field at any point is equal to minus the partial derivative with respect to time of the magnetic field.

The curl of any vector field, $\boldsymbol{A}$, has zero divergence:

$$\nabla \cdot (\nabla \times \boldsymbol{A}) = 0, \tag{2.17}$$

This identity is similar to the relation $\boldsymbol{a} \cdot \boldsymbol{a} \times \boldsymbol{b} = 0$ for any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and provides a consistency check. Taking the divergence of both sides of 2.16, we obtain

$$0 = \nabla \cdot (\nabla \times \boldsymbol{E}) = \nabla \left(-\frac{\partial \boldsymbol{B}}{\partial t} \right) = -\frac{\partial}{\partial t} \nabla \cdot \boldsymbol{B}. \tag{2.18}$$

We have used the commutativity of the temporal and spatial derivatives:

$$\frac{\partial}{\partial t} \nabla = \nabla \frac{\partial}{\partial t}, \tag{2.19}$$

which follows because all the individual components of ∇ commute with ∂_t. Since we know from 2.11 that $\nabla \cdot \boldsymbol{B} = 0$, its derivative must also be zero. If the solenoidal condition did not hold (*i.e.* there were magnetic monopoles), Faraday's law would require modification similar to the one Maxwell found for Ampère's law and which we are about to look at.

AMPERE's LAW and DISPLACEMENT CURRENT

The fourth and final of Maxwell's equation comes from Ampère's law, Equation 1.38,

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{\ell} = \mu_0 \int_S \boldsymbol{J} \cdot d\boldsymbol{S}. \tag{2.20}$$

Using Stokes's theorem, the line integral can be transformed to a surface integral leading to

$$\oint_C \boldsymbol{B} \cdot d\boldsymbol{\ell} = \int_S \nabla \times \boldsymbol{B} \cdot d\boldsymbol{S} = \mu_0 \int_S \boldsymbol{J} \cdot d\boldsymbol{S}. \tag{2.21}$$

Using the hopefully now familiar argument based upon the arbitrary nature of the surface domain of integration in the last two terms, S, we deduce that

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J}. \tag{2.22}$$

[2] It is remarkable that two effects here combine to give Faraday's simple flux rule – see Feynmann's lectures in physics for further discussion of this. It just seems to be the way the world is.

Equation 2.22 is the differential version of Ampère's law for magneto-statics and *almost* the fourth and final of Maxwell's equations. It is not correct in time-varying situations (hence the lack of high-lighting and the "almost"). To see what is wrong, let's take the divergence of this equation. We obtain

$$\nabla \cdot (\nabla \times \boldsymbol{B}) = 0 = \mu_0 \nabla \cdot \boldsymbol{J}, \tag{2.23}$$

i.e.

$$\nabla \cdot \boldsymbol{J} = 0. \qquad \text{[WRONG!]} \tag{2.24}$$

This cannot be true. If the charge distribution varies in time ($\partial\rho/\partial t \neq 0$) we know from Equation 2.4 that $\nabla \cdot \boldsymbol{J} = -\partial\rho/\partial t \neq 0$. If it were true and $\partial\rho/\partial t = 0$, we could never accumulate charge in any region.

If $\nabla \cdot \boldsymbol{J} \neq 0$ as claimed, what *does* it equal? It is true that $\nabla \cdot \nabla \times \boldsymbol{B} = 0$, as this is an identity. To make things work, we need to modify the right-hand side of Ampère's law to make it divergenceless. The answer comes from substituting Gauss's law, $\rho = \epsilon_0 \nabla . \boldsymbol{E}$, into the continuity equation and writing that

$$\frac{\partial\rho}{\partial t} + \nabla \cdot \boldsymbol{J} = \epsilon_0 \frac{\partial \nabla \cdot \boldsymbol{E}}{\partial t} + \nabla \cdot \boldsymbol{J} = \epsilon_0 \nabla \cdot \frac{\partial \boldsymbol{E}}{\partial t} + \nabla \cdot \boldsymbol{J} = 0. \tag{2.25}$$

Hence

$$\nabla \cdot \left(\boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \right) = 0. \tag{2.26}$$

Equation 2.26 is exactly what we need: a modified form of $\boldsymbol{J}$ which we can put on the right-hand side of the "almost" Maxwell-Ampère equation:

$$\boldsymbol{J} \to \boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}, \tag{2.27}$$

which is divergence-free. This leads to the fourth and final of Maxwell's equations, which is now consistent with the continuity equation 2.4:

$$\boxed{\nabla \times \boldsymbol{B} = \mu_0 \left(\boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \right).} \tag{2.28}$$

The extra term

$$\epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}, \tag{2.29}$$

is known as the displacement current density (we will often just call it the displacement current for short). The term "displacement" carries no meaning in the modern version of electromagnetism and harks back to mechanical concepts Maxwell used to develop the theory. For displacement current to be significant, one requires a rapidly changing electric field. This is why it was not an effect picked up in the quasi-static experiments that led Ampère to his law. Instead it was deduced by Maxwell from the reasoning above, an excellent example of the deduction of a physical law through theoretical reasoning. The displacement current term means that a changing electric field generates a magnetic field, just as Faraday's law means that a changing magnetic field generates an electric field. This is crucial to the propagation of electromagnetic waves.

The need for the displacement current can be seen in the case of a charging capacitor. Consider the situation shown in Fig. 2.1 which shows a parallel plate capacitor being charged with a

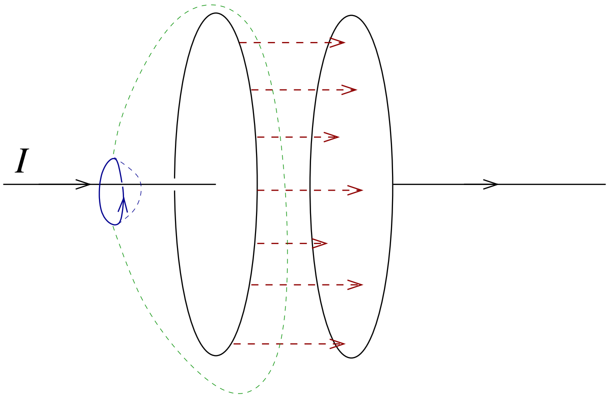

Figure 2.1: A capacitor is charged with a steady current. The small loop with an arrow to the left of the capacitor indicates a circuit to which Ampère's law will be applied. The dashed blue and green lines indicate schematically the nature of the two surfaces spanning the loop to which Ampère's law will be applied.

steady current, I. If we apply Ampère's law to the circuit shown to the left of the capacitor (solid blue loop), we should be able to choose the surface spanning the circuit. If we apply it to the surface indicated by the blue dashed line (shown only in cross-section only), the current in the wire crosses the surface and Ampère's law applies in the usual manner with $\int \boldsymbol{J} \cdot d\boldsymbol{S} = I$. However, if we consider a surface that passes between the capacitor plates but still ends on the solid blue loop (indicated by the dashed green line), the wire does not cross the surface and $\int \boldsymbol{J} \cdot d\boldsymbol{S} = 0$. According to Ampère's law, both of these are equal, namely $\mu_0^{-1} \oint \boldsymbol{B} \cdot d\ell$, but this is impossible (one is equal to I and the other to 0).

The solution of this apparent contradiction is in the displacement current. The electric field between the plates of a parallel plate capacitor is given by $E = \sigma/\epsilon_0$, where σ is the charge density $= Q/A$ if the charge stored is Q and the area of the plates is A. When charged by current I, $dQ/dt = I$, thus $dE/dt = I/\epsilon_o A$, and the displacement current density between the plates $\epsilon_0 \partial E/\partial t = I/A$. The total displacement current between the plates is I, and exactly matches the real current flowing in the wire. Just like a real current, the displacement current leads to a magnetic field between the plates, although in the situation drawn it will be complex to calculate because of edge effects (the space between the plates is very far from the "long cylinder" required for a simple application of Ampère's circuital law).

Another example of the need for a displacement current is a spherically-symmetric outflow of charge (e.g. a charged sphere is placed within an infinite homogeneous conducting medium and allowed to discharge). Current will flow radially outwards from the centre of the charge distribution, but the spherical symmetry means that is hard to see how one can define a magnetic field. (Purely radial fields are ruled out by $\nabla \cdot \boldsymbol{B} = 0$.) The electric field at radius r from the centre of the charge distribution must have magnitude (Gauss's law)

$$E = \frac{Q(r,t)}{4\pi\epsilon_0 r^2}, \tag{2.30}$$

where $Q(r,t)$ is the charge enclosed within radius r at time t. The current density at a given r is

$$J(r,t) = -\frac{1}{4\pi r^2}\frac{\partial Q}{\partial t}, \tag{2.31}$$

(out-flowing current divided by the surface area of a sphere of radius r). The displacement current density is

$$\epsilon_0 \frac{\partial E}{\partial t} = +\frac{1}{4\pi r^2}\frac{\partial Q}{\partial t}. \tag{2.32}$$

The sum of the current density and the displacement current density is zero (the two terms are equal and opposite), so the right hand side in equation 2.28 is zero and $\nabla \times \boldsymbol{B} = \boldsymbol{0}$. No magnetic field is generated.

2.2 Summary of the equations in free space

We have derived the following equations

$$\boxed{\begin{aligned} \nabla \cdot \boldsymbol{E} &= \frac{\rho}{\epsilon_0}, & (2.33)\\ \nabla \cdot \boldsymbol{B} &= 0, & (2.34)\\ \nabla \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, & (2.35)\\ \nabla \times \boldsymbol{B} &= \mu_0 \left(\boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \right). & (2.36) \end{aligned}}$$

These are Maxwell's equations in free space. They are as fundamental as $F = ma$. You should just know them, no ifs, no buts.

In addition it is worth repeating the continuity equation expressing charge conservation

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0.} \tag{2.37}$$

You should also know this equation. It is not an independent equation since it can be deduced by taking the divergence of the fourth of Maxwell's equations and was in fact used to reverse engineer it.

Since the last two of Maxwell's equations are vector relations, there are a total of 8 equations. If you want to calculate with these, you will usually need their component forms in whatever coordinate system you use. For instance, in Cartesian coordinates the first equation is

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \frac{\rho}{\epsilon_0}, \tag{2.38}$$

while the z-component of the last equation is

$$\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \mu_0 \left(J_z + \epsilon_0 \frac{\partial E_z}{\partial t} \right). \tag{2.39}$$

There is no shame in using component forms – you will need them to derive numerical values – however, general relations are often more easily appreciated in terms of ∇. Remember too that the same equations can look different in different coordinate systems. For example the radial component of the Maxwell-Ampère equation in spherical polar "(r, θ, ϕ)" coordinates is

$$\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta B_\phi) - \frac{1}{r\sin\theta}\frac{\partial B_\theta}{\partial\phi} = \mu_0\left(J_r + \epsilon_0\frac{\partial E_r}{\partial t}\right). \tag{2.40}$$

You need not memorise this equation, but you should be able to apply it, and ones like it, for specific forms of B_ϕ, B_θ etc.

2.3 Special cases–electrostatic potentials

Before moving on to study electromagnetic waves in the next chapters, it would be a shame not to see how Maxwell's equations help understand, and derive, some familiar results from electrostatics.

Poisson's and Laplace's Equation
In electrostatics, $\partial/\partial t = \partial_t \equiv 0$. The only non-trivial equations are 2.33 and 2.35:

$$\begin{aligned} \nabla\cdot\boldsymbol{E} &= \rho/\epsilon_0, && (2.41)\\ \nabla\times\boldsymbol{E} &= \boldsymbol{0}. && (2.42)\end{aligned}$$

The second equation, which says that static electric fields have zero curl, is equivalent to $\int_C \boldsymbol{E}\cdot d\ell = 0$, means that $\boldsymbol{E}$ performs zero overall work when a charge is taken around a closed loop. This is the definition of a conservative field. It can be written as the gradient of a potential

$$\boxed{\boldsymbol{E} = -\nabla\psi,} \tag{2.43}$$

where ψ is the electrostatic potential, a scalar field. The negative sign is a convention such that work is done when a positive charge moves from a high to a low potential; ψ is measured in volts (V).

Inserting $\boldsymbol{E} = -\nabla\psi$ into the first of the two equations gives

$$\boxed{\nabla^2\psi = -\rho/\epsilon_0,} \tag{2.44}$$

which is *Poisson's equation* relating the electrostatic potential to the distribution of electric charge.

Specialising to charge-free regions ($\rho = 0$ everywhere) we have

$$\boxed{\nabla^2\psi = 0.} \tag{2.45}$$

This is *Laplace's equation*. You should know some solutions to this equation, e.g. $\psi = -Ex$ (uniform electric-field in the x-direction), and $\psi = q/4\pi\epsilon_0 r$, the Coulomb potential for a charge q at the origin. There are many others. Designing a device to produce a particular electrostatic field is a matter of solving Laplace's equation subject to boundary conditions, e.g. as set by conductors of fixed potential (voltage).

With Laplace's equation we can solve problems that are hard to tackle using Gauss's law in integral form. Here are a couple of examples.

Potential above a flat plate of sinusoidally-varying voltage

Consider a plate in the x–y plane ($z = 0$). The potential on the plate is held at

$$\psi(x, y) = V_0 \cos kx \tag{2.46}$$

with no $y-$dependence. How does the potential vary with z? We need a solution to Laplace's equation satisfying the boundary condition in 2.46. The potential varies equally between positive and negative values. One can guess that it decays away towards zero as one moves away from $z = 0$. As there is no dependence on y, we can specialise Laplace's equation to

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial z^2} = 0. \tag{2.47}$$

At this point one could apply the method of separation of variables, but we can guess a plausible-looking solution:

$$\psi = A\cos(kx)\exp(az), \tag{2.48}$$

where A and a are constants. Substituting into Laplace's equation gives

$$-Ak^2\cos(kx)\exp(az) + Aa^2\cos(kx)\exp(az) = 0, \tag{2.49}$$

which implies $a = \pm k$. $A = V_0$ to match the boundary condition at $z = 0$. We expect the field to decay away from the plane, so $a = -k$ for $z > 0$ and $a = +k$ for $z < 0$. Thus above the plane, the potential takes the form

$$\psi = V_0\cos(kx)\exp(-kz). \tag{2.50}$$

The decay with z is rapid: in the distance of one cycle in the x-direction, kx changes by 2π, so over the same distance in z, the amplitude drops by $\exp(-2\pi) \approx 0.0019$.

A dielectric cylinder in a uniform electric field

An uncharged, uniform, non-conducting cylinder of radius a is made from a material of dielectric constant ϵ_r. The cylinder is placed in a uniform electric field perpendicular to the axis of the cylinder. The cylinder is very long ($L \gg a$), and oriented along the z-axis, while the electric field points along the x-axis. The problem to solve is: what is the potential (and therefore field) inside and outside the cylinder?

The natural coordinates for this are cylindrical polar coordinates (r, θ, z), where $r = \sqrt{x^2 + y^2}$ is the perpendicular distance from the z-axis and is not to be confused with the "r" of spherical polar coordinates which is the distance from the origin, $\sqrt{x^2 + y^2 + z^2}$. A uniform electric field parallel to the x-axis of strength E_0 has corresponding potential

$$\psi = -E_0 x = -E_0 r\cos\theta. \tag{2.51}$$

Since the cylinder is "long", we will assume that we can neglect end effects and that we are in a region where there is no variation with z. Then Laplace's equation in cylindrical polar coordinates reduces to (appendix B) reduces to:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\theta^2} = 0. \tag{2.52}$$

We are interested in solutions that match the uniform field potential, so we look for potentials of the form $\psi = f(r)\cos\theta$. Substituting this into Laplace's equation

$$\frac{\cos\theta}{r}\frac{d}{dr}\left(r\frac{df}{dr}\right) - \frac{f\cos\theta}{r^2} = 0, \tag{2.53}$$

hence

$$r\frac{d}{dr}\left(r\frac{df}{dr}\right) - f = 0. \tag{2.54}$$

The general solution of this equation is

$$f(r) = Ar + Br^{-1}, \tag{2.55}$$

therefore we assume potentials of the form

$$\begin{aligned}\psi_{r\le a} &= \left(A_1 r + \frac{B_1}{r}\right)\cos\theta,\\ \psi_{r\ge a} &= \left(A_2 r + \frac{B_2}{r}\right)\cos\theta,\end{aligned} \tag{2.56}$$

inside and outside the cylinder. Observe that a given solution only holds over a single region.

We are left with 4 undetermined constants to fix in 2.56. We fix these using boundary conditions. First, as $r \to \infty$ ($r \ge a$), we expect to settle back to the original unperturbed field, which is the case if $A_2 = -E_0$, since the term in B_2 decays with r. The interior of the cylinder, $r < a$, includes $r = 0$, and to avoid infinities there, we require $B_1 = 0$ (with $B_1 = 0$, the potential inside the cylinder corresponds to a uniform field of strength $-A_1$.).

We are left with the two coefficients A_1 and B_2, which we fix with boundary conditions at $r = a$. The boundary conditions we have are that the potential must be continuous to avoid infinite gradients and therefore fields, and that there are no free charges. For the other boundary condition we have to borrow a result from the Section 4.4. The final boundary condition is that $\epsilon_0\epsilon_r\boldsymbol{E}_\perp$ is continuous across the boundary ($\epsilon_r = 1$ for $r > a$, and $\boldsymbol{E}_\perp$ is the electric field perpendicular to the interface). If we consider the point $r = a$, $\theta = 0$, then the radial field is perpendicular to the boundary, and we can deduce the condition:

$$-\epsilon_r \left.\frac{d\psi_{r\le a}}{dr}\right|_{r=a} = -\left.\frac{d\psi_{r\ge a}}{dr}\right|_{r=a}. \tag{2.57}$$

This leads to

$$-\epsilon_r A_1 = E_0 + \frac{B_2}{a^2}, \tag{2.58}$$

while the continuity of the potential at $r = a$ gives

$$A_1 a = -E_0 a + \frac{B_2}{a}. \tag{2.59}$$

The two conditions are easily solved for A_1 and B_2:

$$A_1 = -\frac{2}{\epsilon_r + 1}E_0, \tag{2.60}$$

$$B_2 = \left(\frac{\epsilon_r - 1}{\epsilon_r + 1}\right)a^2 E_0. \tag{2.61}$$

So the field strength inside the cylinder is a factor $2/(\epsilon_r + 1)$ times E_0. The term proportional to B_2 perturbs the field outside the cylinder with an angular dependence of $\cos\theta$.

For more on potentials see appendix D.

Chapter 3

Electromagnetic waves in free space

One outstanding result of Maxwell's work was the prediction of electromagnetic waves which, in free space, propagate at the speed of light. The conclusion was that light waves were electromagnetic waves. This unified seemingly disparate areas of physics (electricity and magnetism on the one hand and light and optics on the other). It has to be one of the greatest discoveries in science.

3.1 EM wave solutions to Maxwell's equations

We start from Maxwell's equations in a vacuum with no charges or currents present, *i.e.* we set $\rho = 0$ and $\boldsymbol{J} = \boldsymbol{0}$. Eqs 2.33 to 2.36 then reduce to

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= 0, && (3.1)\\ \nabla \cdot \boldsymbol{B} &= 0, && (3.2)\\ \nabla \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, && (3.3)\\ \nabla \times \boldsymbol{B} &= \mu_0 \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}. && (3.4) \end{aligned}$$

The final two are a pair of coupled partial differential equations: $\boldsymbol{E}$ is generated from a changing $\boldsymbol{B}$ by Faraday's law, while $\boldsymbol{B}$ is generated from a changing $\boldsymbol{E}$ by Maxwell's displacement current term. As we will see, it is this coupling that allows the propagation of EM waves.

Eqs 3.3 and 3.4 can be combined as follows. First take the curl ($\nabla\times$) of both sides of Eq. 3.3:

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla \times \left(-\frac{\partial \boldsymbol{B}}{\partial t} \right) = -\frac{\partial}{\partial t} \nabla \times \boldsymbol{B}, \tag{3.5}$$

where we have used the commutativity of the time and space derivatives as usual. The curl of $\boldsymbol{B}$ on the right-hand side can be replaced using Eq. 3.4,

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \tag{3.6}$$

while the left-hand side can be transformed with a standard identity (Eq. 1.17) to give

$$\nabla (\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E} = -\mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{3.7}$$

With no charge present, the divergence of $\boldsymbol{E}$ is zero (Eq. 3.1) and we are left with

$$\boxed{\nabla^2 \boldsymbol{E} = \mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}.} \tag{3.8}$$

The same procedure, but starting by taking the curl of Eq. 3.4, leads to

$$\boxed{\nabla^2 \boldsymbol{B} = \mu_0 \epsilon_0 \frac{\partial^2 \boldsymbol{B}}{\partial t^2}.} \tag{3.9}$$

(Verify this for yourself.)

Comparing with the standard wave equation $\nabla^2 \psi = (1/v_\phi^2)\partial_t^2 \psi$ for the propagation of some disturbance ψ at wave speed (phase velocity) v_ϕ, Eqs 3.8 and 3.9 are wave equations for the propagation waves in 3D, involving variations of $\boldsymbol{E}$ and $\boldsymbol{B}$, through the vacuum at speed

$$\boxed{c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} = 299\,792\,458\,\mathrm{m\,s^{-1}},} \tag{3.10}$$

matching the speed of light. It was thus natural for Maxwell to conclude that light was a form of electromagnetic wave.

Take a breath. This was a great and unexpected discovery—if you aren't clapping and cheering now, you should be. Maxwell found these wave solutions to his equations and computed their speed. He found it matched the best estimates from experiment for the speed of light. These worked mostly from astrophysical phenomena. One of the first estimates for c (Romer, 1676) was based on the observation that the periods of the moons of Jupiter were shorter when the Earth was approaching and longer when it was moving away from Jupiter.

The wave equations have solutions which are non-dispersive (constant phase velocity, c) and propagate arbitrarily-shaped disturbances unchanged. (As we will see, this is not the case for EM waves in materials.) We will look at the nature of EM plane waves. Consider E- and B-fields of the form

$$\boldsymbol{E} = \boldsymbol{E}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)}, \tag{3.11}$$
$$\boldsymbol{B} = \boldsymbol{B}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)}, \tag{3.12}$$

where $\boldsymbol{k}$ is the wave-vector which points in the direction of propagation. The wavenumber $k = |\boldsymbol{k}| = 2\pi/\lambda$ where λ is the wavelength and ω is the angular frequency of the wave. $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ are amplitude vectors independent of position or time. The term $\boldsymbol{k}\cdot\boldsymbol{r} - \omega t$ is the phase of the wave, which we will denote by ϕ. The amplitudes $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ might contain constant phase factors of the form $e^{i\eta_{E,B}}$, and they should therefore be assumed to be complex. For instance, we might have $E_0 = |E_0|e^{i\eta_E}$, in which case the phase angle for $\boldsymbol{E}$ becomes $\boldsymbol{k}\cdot\boldsymbol{r} - \omega t + \eta_E$. We will see that the η_E and η_B are equal in free space but can differ in the presence of matter.

The phase of the wave is constant for all positions and times for which

$$\phi = \boldsymbol{k}\cdot\boldsymbol{r} - \omega t \tag{3.13}$$

is constant. At a particular instant of time, this means $\boldsymbol{k}\cdot\boldsymbol{r} =$ constant, which is the equation of a plane, hence the name "plane wave". They are idealised since they fill all of space for all time, whereas there is always spatial and temporal confinement in any real system. Nevertheless, there are many circumstances where neglecting this confinement and working with plane waves make little difference to the outcome.

We substitute the assumed form for the fields of Eqs. 3.11 and 3.12 into the sourceless Maxwell equations in a vacuum, Eqs 3.1 to 3.4. The time and spatial derivatives (∂_t, ∂_x, ∂_y, ∂_z), act on the exponentials since the amplitude vectors are constant. Remembering that

$$\boldsymbol{k}\cdot\boldsymbol{r}-\omega t = k_x x + k_y y + k_z z - \omega t, \tag{3.14}$$

we have

$$\begin{aligned} \partial_t\left(\mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\right) &= -i\omega \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \\ \partial_x\left(\mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}\right) &= ik_x \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \end{aligned} \tag{3.15}$$

with similar relations for the y- and z-derivatives. If we substitute plane wave disturbances into Maxwell's equations, we see that the time derivatives are equivalent to multiplication by $-i\omega$, and acting with ∇ is equivalent to acting with $i\boldsymbol{k}$. After dividing out factors of i, Eqs 3.1 to 3.4 become

$$\begin{aligned} \boldsymbol{k}\cdot\boldsymbol{E} &= 0, && (3.16)\\ \boldsymbol{k}\cdot\boldsymbol{B} &= 0, && (3.17)\\ \boldsymbol{k}\times\boldsymbol{E} &= \omega\boldsymbol{B}, && (3.18)\\ \boldsymbol{k}\times\boldsymbol{B} &= -(\omega/c^2)\boldsymbol{E}. && (3.19) \end{aligned}$$

The first two of these relations show that $\boldsymbol{E}$ and $\boldsymbol{B}$ are perpendicular to the wave vector $\boldsymbol{k}$, which means that EM waves in a vacuum are *transverse* waves. The second two equations show that $\boldsymbol{E}$ and $\boldsymbol{B}$ are perpendicular to each other. In free space, EM waves have $\boldsymbol{E}$, $\boldsymbol{B}$ and $\boldsymbol{k}$ mutually perpendicular.

Looking only at what they say about the magnitudes of the vectors, equations 3.18 and 3.19 show that

$$\boxed{\frac{E}{B} = \frac{\omega}{k} = c.} \tag{3.20}$$

The effect of these fields on a test charge is via the Lorentz force $q(\boldsymbol{E}+\boldsymbol{v}\times\boldsymbol{B})$. One can see from 3.20 that for non-relativistic motion, $v \ll c$, the effect of the electric field of an EM wave on test charges is dominant over that of the magnetic field. An EM wave with an electric field amplitude of $1000\,\mathrm{V\,m^{-1}}$ has a magnetic field amplitude of $\approx 3\,\mu\mathrm{T}$.

One more result to note is that no imaginary or complex numbers feature in the above equations (i.e. Eqs 3.16 to 3.19). This means that $\boldsymbol{E}$ and $\boldsymbol{B}$ oscillate *in phase*. (The phase factors η_E and η_B mentioned in the discussion after 3.12 are equal.) The electric amd magnetic fields reach their maximum and minimum strength at the same place at a given time or the same time at a given place.

The characteristic properties of EM waves in a vacuum, that we have noted above, justify the classic visualisation of EM waves shown in Fig. 3.1. This picture illustrates the transverse nature with $\boldsymbol{E}$ and $\boldsymbol{B}$ perpendicular to each other and in phase. (Note that $\boldsymbol{E}$ and $\boldsymbol{B}$ are not in the same units.) The picture is slightly misleading to the extent that it appears to indicate that the fields are only defined on a line, whereas the values indicated at any one point along the propagation direction extend to all points on planes perpendicular to it.

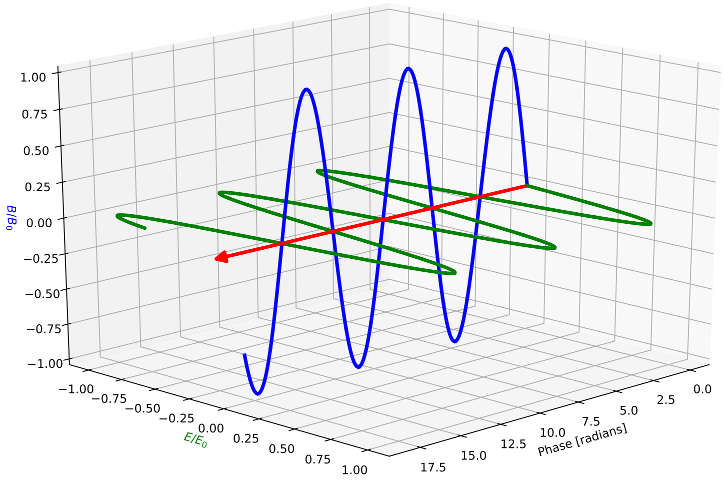

Figure 3.1: An illustration of the relative phases and directions of the electric (green) magnetic (blue) fields in an electromagnetic plane wave, propagating in the direction of the arrow (leftwards and outwards from the page).

3.2 Energy in an EM wave

We should expect to be able to find a statement of the conservation of energy in EM systems. We know that there is energy in light (sunlight warms us up). We need to find an expression for the energy flux. Our statement of energy conservation should involve the energy stored in the electric and magnetic fields, the energy flux in and out of an arbitrary volume and any work done on charges. The equation stating this law will have the form (see 2.4):

$$\frac{\partial u}{\partial t} = -\nabla \cdot \boldsymbol{P} - W, \tag{3.21}$$

with u as the energy density in the EM fields and $\boldsymbol{P}$ as the energy flux of the EM fields out of the local volume. W is the rate at which work is being done per unit volume on any charges present.

What are u and $\boldsymbol{P}$ in terms of $\boldsymbol{E}$ and $\boldsymbol{B}$? We will look for a relation of the type 3.21 involving the electric and magnetic fields and identify which terms are playing the role of the energy density and the energy flux. If there are any charges flowing in the system, the work done on these charges per unit volume is given by

$$W = \boldsymbol{E} \cdot \boldsymbol{J}. \tag{3.22}$$

(Integrating over an element of surface turns $\boldsymbol{J}$ into a current and integrating along an element of length in the direction of the flow turns $\boldsymbol{E}$ into a voltage drop to give the usual $I \times V$ expression familiar from circuits.)

Inserting 3.22 into 2.37 and using the Maxwell equation 2.36 together with the identity 1.20[1]

[1] Setting $\boldsymbol{V} = \boldsymbol{E}$ and $\boldsymbol{W} = \boldsymbol{B}$ we have $\nabla \cdot (\boldsymbol{E} \times \boldsymbol{B}) = \boldsymbol{B} \cdot (\nabla \times \boldsymbol{E}) - \boldsymbol{E} \cdot (\nabla \times \boldsymbol{B})$

gives

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{P} = -\frac{1}{\mu_0}\boldsymbol{E} \cdot (\nabla \times \boldsymbol{B}) + \epsilon_0 \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t} \tag{3.23}$$

$$= \frac{1}{\mu_0}\nabla \cdot (\boldsymbol{E} \times \boldsymbol{B}) - \frac{1}{\mu_0}\boldsymbol{B} \cdot (\nabla \times \boldsymbol{E}) + \epsilon_0 \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t}. \tag{3.24}$$

With the Maxwell equation 2.35 we obtain

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{P} = \frac{1}{\mu_0}\nabla \cdot (\boldsymbol{E} \times \boldsymbol{B}) + \frac{1}{\mu_0}\boldsymbol{B} \cdot \frac{\partial \boldsymbol{B}}{\partial t} + \epsilon_0 \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t} \tag{3.25}$$

$$= \frac{1}{\mu_0}\nabla \cdot (\boldsymbol{E} \times \boldsymbol{B}) + \frac{\partial}{\partial t}\left(\frac{B^2}{2\mu_0} + \frac{\epsilon_0 E^2}{2}\right). \tag{3.26}$$

This is what we are looking for. We identify the energy flux, $\boldsymbol{P}$, and energy density, u, in the EM fields

$$u = \left(\frac{B^2}{2\mu_0} + \frac{\epsilon_0 E^2}{2}\right), \tag{3.27}$$

$$\boldsymbol{P} = \frac{1}{\mu_0}(\boldsymbol{E} \times \boldsymbol{B}). \tag{3.28}$$

We call $\boldsymbol{P}$ the Poynting vector after its discoverer.

You may ask is it correct to identify terms in an equation with physical quantities just because they have the same form. This is a good question. It is discussed in Feynmann Vol II (section 27-3). Roughly speaking, Feynman's answer is that the predictions based on these identifications have been tested against what can be measured and seem to be right. They are now universally accepted to be so.

Looking at the Poynting vector in the case of the plane wave given in 3.11 and 3.12, we have from 3.16 to 3.19 that (remember $\epsilon_0\mu_0 = 1/c^2$)

$$\boldsymbol{P} = \frac{1}{\mu_0\omega}\boldsymbol{E} \times (\boldsymbol{k} \times \boldsymbol{E}) = \frac{1}{\mu_0\omega}E^2\boldsymbol{k} = \frac{\epsilon_0}{\epsilon_0\mu_0}\frac{k}{\omega}E^2\hat{\boldsymbol{k}} = c(\epsilon_0 E^2)\,\hat{\boldsymbol{k}} = c\frac{B^2}{\mu_0}\,\hat{\boldsymbol{k}}. \tag{3.29}$$

This is an intuitively sensible result. The Poynting vector for the plane wave is in the direction of $\hat{\boldsymbol{k}}$ with magnitude proportional to the intensity of the light multiplied by its speed. (Note that, when inserting the expressions for $\boldsymbol{E}$ and $\boldsymbol{B}$ from 3.11 and 3.12, one should take their real parts before computing any actual values.)

We will come back to the Poynting vector after looking at EM Theory in matter.

Chapter 4

Maxwell's equations in matter

When electric fields are applied to matter, the distribution of their electric charge changes. The positively-charged nuclei are pulled in the direction of the applied field relative to the negatively-charged electrons. So long as the applied field is not too large, individual atoms and molecules in the material remain neutral, but the relative movement of positive and negative charges means that the atoms and molecules acquire electric dipole moments which generate their own contribution to the electric field. The electric field inside an object is then not the same as the applied external field. Moreover, the fields generated by the dipole moments can affect the field external to the object as well. In some substances, most famously water, the molecules may have permanent electric dipole moments owing to the distribution of electronic charge, resulting in a large response to applied fields which tend to align the dipoles (thermal excitation tends to randomise their orientation).

When a magnetic field is applied to matter, currents may be generated. The constituent atoms become magnetised as a result and then generate their own magnetic field. In certain substances, interactions between atoms can cause long-term ordering. An example is "ferromagnetism". This is where a material is magnetised even in the absence of an applied magnetic field.

The phenomena associated with the polarisation and magnetisation of matter are of immense practical importance and we need to be able to describe them. To do this, we will develop a set of equations which, while entirely equivalent to Maxwell's equations in free space, are more suited to describing electromagnetic fields in the presence of matter. We introduce the *polarisation*, $\boldsymbol{P}$, which is the electric dipole moment per unit volume or electric dipole moment density, and the *magnetisation*, $\boldsymbol{M}$, which is the magnetic dipole moment per unit volume or the magnetic dipole moment density. Electric polarisation and magnetisation are the result of charge redistribution and motion. If we took account of these charges in Maxwell's equations everything would be correct. However, as polarisation and magnetisation are naturally described in terms of electric and magnetic dipoles and dipole densities, it is conceptually useful to treat them separately. We do this by treating the effects of the bound charges (electrons bound inside neutral matter) separately from those of free charges.

Electric polarisation and magnetisation are slightly "fuzzy" concepts. We have to imagine averaging over small volumes of matter but still large enough that they contain enough atoms that the averages are well defined. This is called taking the continuum limit. This limit is physically reasonable only if the variation in the averaged quantities of interest, such as $\boldsymbol{P}$

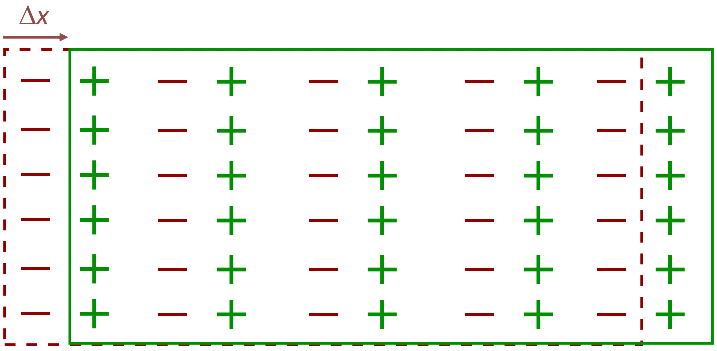

Figure 4.1: The positive charges in a neutral uniform cuboid are shifted by an amount Δx to the right relative to the negative charges (cross-section in x-y plane shown). A uniform polarisation is generated within the bulk of the cuboid, but there is no overall charge there, however surface charges are generated. The green and red boxes represent the positive and negative charges.

and $\boldsymbol{M}$, is on length scales large compared to atomic separations. For the propagation of light, this variation occurs on the scale of the wavelength of the light, λ. For visible light, λ is actually much larger than the atomic scale and the continuum limit gives a valid description of the behaviour of the system. The typical wavelength of optical light encompasses ~ 5000 atoms linearly and thus there are $> 10^{10}$ atoms in a volume of λ^3. However, at high photon energies, e.g. X-rays, the wavelengths are much smaller and it is no longer possible to assume a smooth distribution of electric charge.

4.1 Polarisation charges and currents

Polarisation Charges

An ideal electric dipole with dipole moment p consists of two equal and opposite charges $\pm q$ a distance d apart in the limit $d \to 0$, $q \to \infty$ with $qd = p$ fixed. The picture of a sea of closely-separated equal-but-opposite charges is a useful way to visualise an electrically polarised material. (Be careful to distinguish the electrical polarisation we are talking about here from the more commonly met polarisation of light: the two are entirely distinct but unfortunately since they both crop up in electromagnetism, they can be simultaneously encountered.)

Consider a uniform cuboid aligned with Cartesian axes in which there are balancing charge densities $\pm\rho$, so that it is electrically neutral overall. Suppose that initially the positive and negative charges lie on top of each other. Now imagine displacing all the positive charges by the distance Δx in the positive x-direction. Any small volume δV splits into negative and positive charges of magnitude $q = \rho\,\delta V$, and acquires a dipole moment $p = \rho\Delta x\,\delta V$. The resulting polarisation (dipole moment per unit volume) is in the x-direction and has magnitude

$$P_x = \rho\Delta x. \tag{4.1}$$

Fig. 4.1 illustrates this idea. The charges within the bulk of the cuboid cancel so there is no overall charge density in this region. However, charge imbalances are created on each of the faces perpendicular to the displacement. In the case shown, a surface charge density, $\sigma = \rho x = P_x$, is placed on the right-hand face and $-P_x$ on the left-hand face. The top and

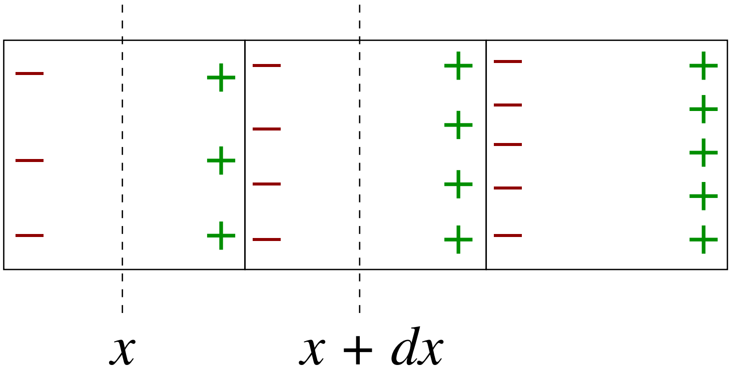

Figure 4.2: In this figure the polarisation (x-component only) increases in strength with x, as represented by the increasing number of charges from left to right. Splitting the material into boxes each with surface charges of $\pm P_x$, at each boundary we see an overall excess of negative charge with leads to a volume polarisation charge density.

bottom faces, which are parallel to the displacement, have no charge. A point on the surface, where the polarisation is $\boldsymbol{P}$ and where the outward-pointing normal unit vector is $\hat{\boldsymbol{n}}$, will have a surface charge density

$$\boxed{\sigma_P = \boldsymbol{P} \cdot \hat{\boldsymbol{n}}.} \tag{4.2}$$

The subscript P indicates that these are polarisation charges to distinguish them from "free" charges. In the case shown in Fig. 4.1, the right-hand face has $\hat{\boldsymbol{n}} = \hat{\boldsymbol{x}}$ and $\sigma_P = P_x$, while the left-hand face has $\hat{\boldsymbol{n}} = -\hat{\boldsymbol{x}}$ and $\sigma_P = -P_x$.

We also need to be able to handle spatially varying polarisations. We can do this by breaking up an object, in which the polarisation is non-uniform, into small cuboids each of which can be said to be nearly-uniformly polarised, allowing us to apply the surface charge formulation of 4.2.

Consider a situation in which the polarisation is non-uniform (see Fig. 4.2) and approximate the non-uniform polarisation by breaking the object into a series of small cuboids of length dx along the x axis, each with individually uniform polarisation. The cuboid centred on x has polarisation charge density $\pm P_x(x)$ on its two faces. The next one, centred on $x + dx$, has polarisation charge density $\pm P_x(x + dx)$. At the interface between them, there is a charge excess of $A(P_x(x) - P_x(x + dx))$, where A is cross-sectional area of the cuboid perpendicular to x. Dividing by the volume per cuboid, $A\,dx$, gives a polarisation charge density:

$$\rho_P = \frac{AP_x(x) - AP_x(x+dx)}{A\,dx} = -\frac{P_x(x+dx) - P_x(x)}{dx} = -\frac{\partial P_x}{\partial x} \tag{4.3}$$

in the limit $dx \to 0$. Adding similar terms for y and z, we find that

$$\boxed{\rho_P = -\nabla \cdot \boldsymbol{P},} \tag{4.4}$$

is the volume density of polarisation charges. For a more mathematical derivation of the same results, see appendix F.

If the polarisation in a material changes with time, charges must move leading to currents. These "polarisation currents" will generate magnetic fields which we will need to account for. Returning to Fig. 4.1, imagine that the displacement s is a function of time, then the

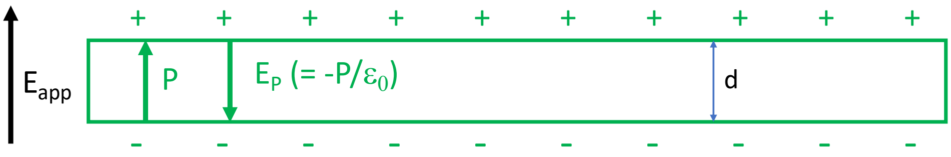

Figure 4.3: A slab placed in an applied electric field, $\boldsymbol{E}_0$, is polarised. In an isotropic medium, the polarisation, $\boldsymbol{P}$ is parallel to $\boldsymbol{E}_0$ and leads to a charge density on its surfaces $\sigma_P = P$ (see 4.2 with $\boldsymbol{P}$ parallel to the surface normal). The electric fields from the charges on the two surfaces generate the polarisation field inside the slab $\boldsymbol{E}_P = \boldsymbol{P}/\epsilon_0$. The total field is the sum of the applied field and polarisation fields.

corresponding current density is $\rho(ds/dt) = \partial P_x/\partial t$ (as $P_x = \rho s$). Allowing for components in the other directions, the polarisation current density is is given by

$$\boxed{\boldsymbol{J}_P = \frac{\partial \boldsymbol{P}}{\partial t}.} \tag{4.5}$$

This obeys the continuity equation

$$\frac{\partial \rho_P}{\partial t} + \nabla \cdot \boldsymbol{J}_P = 0. \tag{4.6}$$

We can verify this by substituting $\rho_P = -\nabla\cdot\boldsymbol{P}$ and $\boldsymbol{J}_P = \partial\boldsymbol{P}/\partial t$. The left-hand-side becomes

$$-\frac{\partial \nabla \cdot \boldsymbol{P}}{\partial t} + \nabla \cdot \frac{\partial \boldsymbol{P}}{\partial t}, \tag{4.7}$$

which vanishes because $\partial/\partial t$ and ∇ commute. We will use Eq. 4.5 later in this chapter.

Excursion on Uniformly Polarised Matter

In 4.2 and 4.4, we have the equations we need to describe electrostatics. Before moving on to study polarisation currents and magnetisation, we will look at polarisation in simple electrostatic problems ($\nabla \cdot \boldsymbol{P} = 0$) involving finite objects and their surface charge. The simplest of these by far is a uniform slab in a perpendicular applied field, E_{app}, see Figure 4.3.

The material polarises (has a non-zero polarisation, $\boldsymbol{P}$), which we will assume to be uniform. The resulting surface charge densities on the two surfaces generate the polarisation field inside the material, $\boldsymbol{E}_P = -\boldsymbol{P}/\epsilon_0$ (the magnitude is set by the charge density on the faces of the slab - a problem dealt with in PX120). The total electric field is the sum of the applied and polarisation fields:

$$\boldsymbol{E} = \boldsymbol{E}_{\mathrm{app}} + \boldsymbol{E}_P = \boldsymbol{E}_{\mathrm{app}} - \frac{\boldsymbol{P}}{\epsilon_0}. \tag{4.8}$$

To go further we would need to know something about the material, namely how much polarisation an electric field generates. This will depend on the material. We assume that the response to the electric field is linear in $\boldsymbol{E}$ and write

$$\boxed{\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E}.} \tag{4.9}$$

Linear means that χ does not depend on the field. χ is called the polarisability or susceptibility[1]. The factor ϵ_0 is included in the definition to make χ dimensionless. For isotropic materials it

[1] The term susceptibility is used quite generally to denote a response to a perturbing field, which could be electric or magnetic.

is a number and, as in 4.9, the polarisation is parallel to the electric field. This is an important equation. It is what we will assume about all materials we will look at (we will only look at isotropic materials). Note that the system responds to the actual electric field, which is the sum of the applied and polarisation fields.

Taking the total electric field from 4.8 and rearranging gives

$$\boldsymbol{P} = \frac{\chi}{1+\chi}\epsilon_0 \boldsymbol{E}_{\text{app}} \quad \text{and} \quad \boldsymbol{E} = \frac{1}{1+\chi}\boldsymbol{E}_{\text{app}} \equiv \frac{1}{\epsilon_r}\boldsymbol{E}_{\text{app}}. \tag{4.10}$$

The quantity ϵ_r is called the relative permittivity. Equation 4.10 shows that the electric field inside the slab is reduced by the factor $\epsilon_r = (1+\chi)$. The larger the polarisability, χ, the less the field becomes. We sometimes say that the field has been 'screened'.

Was it right to assume constant uniform polarisation? As we found a solution, the answer is yes. Maxwell's equations are linear PDEs, if our answer satisfies the boundary conditions and the equations, the solution is unique. If we can find a solution we know it is the right one.

In other geometries, it is not as simple to find the electric field and polarisation as it is for the slab. Even if we can assume an isotropic medium (see 4.9), the shape of the boundary comes into play. For some simple cases like spheres, ellipsoids and rods, the problem can be solved analytically. Equation 4.8 becomes

$$\boldsymbol{E} = \boldsymbol{E}_{\text{app}} + \boldsymbol{E}_P = \boldsymbol{E}_{\text{app}} - \alpha\frac{\boldsymbol{P}}{\epsilon_0}, \tag{4.11}$$

where α is called the depolarisation factor and is some number. The fields $-\alpha\boldsymbol{P}/\epsilon_0$ are sometimes called "depolarisation fields". Equations 4.10 then become

$$\boldsymbol{P} = \frac{\chi}{1+\alpha\chi}\epsilon_0 \boldsymbol{E}_{\text{app}} \quad \text{and} \quad \boldsymbol{E} = \frac{1}{1+\alpha\chi}\boldsymbol{E}_{\text{app}}. \tag{4.12}$$

Known examples include the sphere ($\alpha = 1/3$, this is handled in appendix C) and a thin rod ($\alpha = 0$). For ellipsoids there are different factors for applied fields aligned along the three principal axes (α_x, α_y and α_z, with $\alpha_x + \alpha_y + \alpha_z = 1$). In other cases, one may have to resort to numerics to solve for $\boldsymbol{P}$ and $\boldsymbol{E}$. Note, though, that the constituent relation 4.9 is a local relation between the total field and the polarisation at that point. It does not depend on geometry.

4.2 Magnetisation

As electric polarisation can be thought of in terms of distributed electric dipoles, magnetised materials can be pictured as having a distribution of small current loops. We call these loops magnetic dipoles. We use the density of these dipoles, which we call the magnetisation, to account for the effects of the motion of bound charges in matter. Where uniform electric polarisation leads to surface charges, uniform magnetisation leads to surface currents and where non-uniform polarisation leads to a volume charge distribution (see 4.4), non-uniform magnetisation leads to volume currents.

We define the magnetic moment of a current loop to have magnitude equal to the current around the edge of a surface times the area of the surface. Its direction is defined to be the

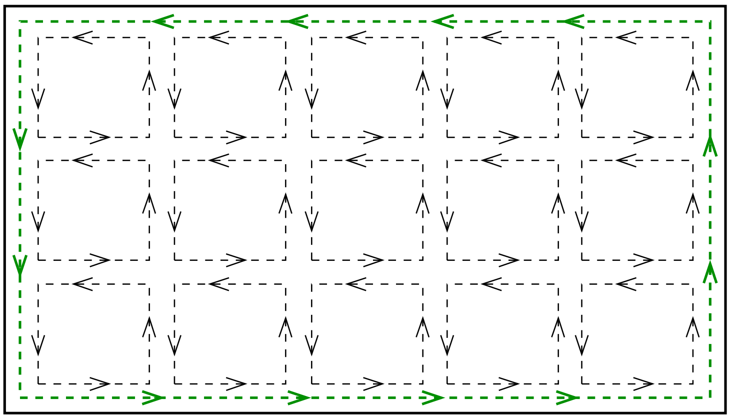

Figure 4.4: A magnetised block, with the magnetisation pointing upwards out of the page, can be thought of as an array of small current loops. When the magnetisation is uniform, the currents from neighbouring loops cancel, except at the surface where a circulating current (green) remains.

surface normal. Incidentally, the understanding of the connection between circulating currents at the microscopic level and magnetisation was hard to come by before a theory of matter came along. This needed to take account of electrons, nuclei and quantum effects—see discussion in Feynmann around Eq 32.17.

We assume that we can attribute the magnetic properties of matter to currents of charges inside the system. Fig. 4.4 shows a magnetised block viewed as a set of current loops which we will take to be responsible for any magnetic field generated by the block. The current runs along the sides rising out of the page. These have height Δz so that one can think of the current per unit length, $I/\Delta z$. We imagine that current loops are spaced in a regular array by $(\Delta x, \Delta y, \Delta z)$ in the (x, y, z) directions, that they are oriented parallel to and fill the x–y plane. Each loop has area $\Delta x \Delta y$ (in Fig. 4.4 they are shown spaced apart for clarity). Each block carries current I, and we define its magnetic moment $m = I\Delta x\Delta y$, leading to magnetisation

$$M = \frac{m}{V} = \frac{I\Delta x\Delta y}{\Delta x\Delta y\Delta z} = \frac{I}{\Delta z}. \tag{4.13}$$

By the right-hand rule, the magnetic field corresponding to the current loop points upwards out of the page, which we are calling the z-axis, i.e. $\boldsymbol{M} = (I/\Delta z)\hat{\boldsymbol{z}}$. We associate the direction of this magnetic field with the direction of the magnetisation. When the magnetisation is uniform, the current flowing in each loop I is fixed and the currents from neighbouring loops cancel in the interior of the block.

The only unbalanced currents (of strength I) run around the outside of faces of the blocks, which are spaced every Δz in the z-direction. This corresponds to a *surface current* density $I/\Delta z = M$, measured in Ampères per unit metre.[2] The surface current runs perpendicular both to $\boldsymbol{M}$ and to the normal vector to the face across which it is flowing. For a block shown in Figure 4.4, it flows in the positive x-direction on the lower face, for which the outward pointing normal vector $\hat{\boldsymbol{n}} = -\hat{\boldsymbol{y}}$. These results can be written as

$$\boxed{\boldsymbol{j}_M = \boldsymbol{M} \times \hat{\boldsymbol{n}}} \tag{4.14}$$

[2] A surface current density is a current per unit length, while a current density is a current per unit area.

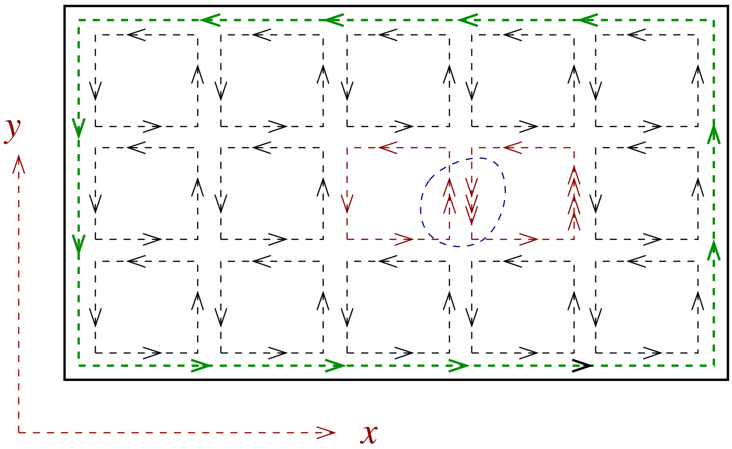

Figure 4.5: A magnetised block, when the sections highlighted in red indicate a region of increasing magnetisation (increasing with x, i.e. $\partial_x M_Z > 0$), with corresponding increase in the loop current I with x. The interior currents no longer cancel for the red regions leaving a net current flowing in the negative y-direction.

where $\boldsymbol{M}$ is the magnetisation and $\hat{\boldsymbol{n}}$ the outward-point normal at the point in question. This is the general expression for the surface current density induced by a magnetisation $\boldsymbol{M}$. It is the magnetic counterpart to $\sigma_P = \nabla \cdot \boldsymbol{P}$ (see 4.4) for electric polarisation.

If magnetisation varies with position inside a material, the currents of neighbouring loops will no longer cancel. A volume current density arises. Fig. 4.5 illustrates an example. A variable magnetisation can be modelled with the same lattice of small loops if we allow the current in each loop to vary with position. In the case shown, I and therefore $M = I/\Delta z$, increase with x. If two neighbouring loops are centred at x and $x + \Delta x$, then where their edges align a residual current

$$\Delta I = I(x + \Delta x) - I(x) \approx \frac{\partial I}{\partial x}\Delta x, \tag{4.15}$$

flows in the negative y-direction. The residual currents form a lattice every Δx by Δz in the x–z plane and lead to a *volume* current density $J = \Delta I/(\Delta x \Delta z)$:

$$J_y = -\frac{\Delta I}{\Delta x \Delta z} = -\frac{1}{\Delta z}\frac{\partial I}{\partial x} = -\frac{\partial M_z}{\partial x}. \tag{4.16}$$

You should recognise this as part of the y-component of the curl of $\boldsymbol{M}$.

After taking account of how the various surface currents depend on position, one finds that the general expression for the volume magnetisation current is

$$\boxed{\boldsymbol{J}_M = \nabla \times \boldsymbol{M}.} \tag{4.17}$$

You could derive this yourself. If this seems daunting, you should certainly verify that this expression gives the correct value for the situation just analysed (I_y depending on x).

As an example of magnetisation currents, consider a permanent magnet in the shape of a cylinder much longer than it is wide. Assuming that it is uniformly magnetised along its length, it will have surface currents circulating around the the cylinder perpendicular to its length of strength $j = M$. The field produced will be the same as that produced by a long solenoid, i.e. $\mu_0 NI/\ell$, where N is the number of turns and ℓ its length. NI/ℓ is equivalent to the surface current density of the solenoid, so by analogy the magnetic flux density in the middle of the permanent magnet is $\mu_0 j = \mu_0 M$. The field at either end of a long solenoid is half the value in the middle, so the magnetic flux density at the poles of the permanent magnet will be $B = \mu_0 M/2$. Assuming $B = 1$ T, typical of modern permanent magnets, then $M \approx 2 \times 10^6\,\mathrm{A\,m^{-1}}$. A 2 cm magnet then has a magnetisation current of order 40 000 A running round it, which is not trivial to replicate with wires and currents.

Appendix F contains a more formal derivation of the magnetisation current formulae which you may prefer, although as with polarisation, the justification given here is a good way to think about what is happening.

4.3 Maxwell's equations in matter

We are now in position to return to Maxwell's equations to see how to formulate them when materials are present. We will begin with polarisation and Gauss's law, which leads to the definition of a new field, the "displacement", $\boldsymbol{D}$. Although we have already come across this in the Maxwell-Ampère equation as the "displacement current", the next section is the better way to think about displacement, which we will later see appears in time derivative form as the displacement current in a modified form of the Maxwell-Ampère equation.

Displacement, $\boldsymbol{D}$
Returning to Maxwell's equations, consider Gauss's law

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{4.18}$$

and separate the total charge density ρ into the sum of free and polarisation charges:

$$\rho = \rho_f + \rho_P = \rho_f - \nabla \cdot \boldsymbol{P}, \tag{4.19}$$

by 4.4. We make this division because free charges are those that we can (in principle) manipulate, while the polarisation charges come with the territory when you apply electric fields and there is matter present – their response is controlled by the properties of the matter. Very often polarisation charges are called "bound" charges as the opposite of "free", but we will call them polarisation charges. They are perfectly real charges, and not in any way imaginary or fictitious or a clever "device".

Taking the divergence of polarisation over to the left-hand side and multiplying both sides by ϵ_0 gives

$$\nabla \cdot (\epsilon_0 \boldsymbol{E} + \boldsymbol{P}) = \rho_f. \tag{4.20}$$

It is then convenient to define a new quantity $\boldsymbol{D}$, known as the "displacement" for historical reasons, by

$$\boxed{\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}.} \tag{4.21}$$

This leads to the modified form of Gauss's law:

$$\boxed{\nabla \cdot \boldsymbol{D} = \rho_f,} \tag{4.22}$$

with the f to remind us that the charge density on the right is the free charge density only.

By analogy with the correspondence between the integral and differential forms of Gauss's law, we can write down the integral form corresponding to Eq. 4.22, namely

$$\oint \boldsymbol{D} \cdot d\boldsymbol{S} = \int \rho_f \, dV = Q_f, \tag{4.23}$$

where Q_f is the free charge enclosed by the surface of integration on the left. This makes it possible to solve problems involving dielectrics using the same arguments involving symmetry and Gaussian surface as used for the free space case.

To solve a problem we need to know $\boldsymbol{P}$ for a given $\boldsymbol{E}$. This in general is a complicated (although interesting) question that depends upon the particular material. A full quantum theory would be needed to describe the properties of any material properly and is not considered part of EM theory. However, there are a number of useful instances where, for small enough electric fields, the induced polarisation is parallel to and linearly proportional to the strength of the field so that we can write

$$\boxed{\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E},} \tag{4.24}$$

where the dimensionless constant χ is called the "electric susceptibility". This is the same as 4.9. Therefore from Eq. 4.21

$$\boxed{\boldsymbol{D} = \epsilon_0 (1 + \chi) \boldsymbol{E} = \epsilon_0 \epsilon_r \boldsymbol{E},} \tag{4.25}$$

where

$$\boxed{\epsilon_r = 1 + \chi} \tag{4.26}$$

is another dimensionless quantity and is known as the "relative permittivity". It should be understood that $\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E}$ is an approximation. In general there must always be some level of non-linearity in which χ will depend on $\boldsymbol{E}$. For example, *all* materials will suffer electrical breakdown for large enough fields.

There are many cases in which, even if the relation 4.9 is well-described as linear (χ independent of $\boldsymbol{E}$), $\boldsymbol{P}$ is not parallel to $\boldsymbol{E}$. This happens when a material is anisotropic (more polarisable in one direction than it is in another). In such cases there is a tensor relation between $\boldsymbol{P}$ and $\boldsymbol{E}$. Anisotropy is responsible for the phenomenon of "birefringence"; calcite is a well-known crystal for which this effect is prominent (see appendix G if you want to know more about this off-syllabus but interesting phenomenon). Although anisotropic crystals are not uncommon, we will assume the linear, scalar form of χ and ϵ_r, which is a good approximation in many instances (e.g. glass, crystalline materials of high structural symmetry).

A final comment on χ (and therefore ϵ_r) is that it is in general frequency dependent, reflecting how fast a material can respond to a changing field. This is the reason why the refractive index of glass changes with wavelength, as we will see later. Water provides a good example of this. At frequencies below ~ 1 GHz, there is enough time for the water molecules to orient themselves in response to electric fields, leading to a strong polarisation and large relative permittivity $\epsilon_r \sim 70$ (the exact value is temperature dependent). At optical frequencies, however, only the electron distribution within the molecules can respond fast enough, and ϵ_r drops below 2.

Example 4.1. A point charge q is placed at the centre of a uniform, uncharged, hollow spherical shell of dielectric material of inner radius a, outer radius b and relative permittivity ϵ_r. What are the electric field $\boldsymbol{E}$ and displacement $\boldsymbol{D}$ strengths as a function of distance r from the charge? Discuss the distribution of polarisation charge in the dielectric.

Answer. Consider a Gaussian surface at radius r. By symmetry $\boldsymbol{D}$ is everywhere radial and has the same magnitude everywhere on the surface of constant radius. We then have

$$4\pi r^2 D(r) = q, \tag{4.27}$$

so

$$D(r) = \frac{q}{4\pi r^2}. \tag{4.28}$$

For $r < a$ and $r > b$, we are in the free space case with no polarisation and $\boldsymbol{D} = \epsilon_0 \boldsymbol{E}$, therefore

$$E(r) = \frac{q}{4\pi\epsilon_0 r^2} \text{ for } r < a \text{ and } r > b, \tag{4.29}$$

whereas in the shell $(a < r < b)$

$$E(r) = \frac{D(r)}{\epsilon_0\epsilon_r} = \frac{q}{4\pi\epsilon_0\epsilon_r r^2}. \tag{4.30}$$

The strength of E drops in the dielectric. This is the result of polarisation charges. In the body of the dielectric there is no polarisation charge since $\nabla\cdot\boldsymbol{P} \propto \nabla\cdot\boldsymbol{E} = 0$ (a $1/r^2$ radial field has no divergence except at the origin). There are therefore just polarisation charges of (surface) density $\sigma_P = P = \epsilon_0\chi E = \epsilon_0(\epsilon_r - 1)E$, with negative charges on the inner surface and positive charges at the outer surface thus

$$\begin{aligned} \sigma(r = a) &= -\frac{\epsilon_r - 1}{\epsilon_r}\frac{q}{4\pi a^2}, \\ \sigma(r = b) &= +\frac{\epsilon_r - 1}{\epsilon_r}\frac{q}{4\pi b^2}. \end{aligned}$$

From these we see that the total polarisation charge on the inner and outer surfaces $= \pm(1 - 1/\epsilon_r)q$. It is these charges that partly screen the innermost free charge q inside the dielectric. If the polarisability of the material is large, $\chi \gg 1$ and hence $\epsilon_r \gg 1$, the screening is very effective and the electric field inside the dielectric is much lower than outside it. A conductor can be approximated as an infinitely polarisable dielectric $(\chi \to \infty)$, and supports no internal electric fields.

Magnetic Field Strength $\boldsymbol{H}$

We carry out a similar modification of Ampère's law. In this case we split the total current density $\boldsymbol{J}$ that appears in the free space equation into three components: a free current density that is (in principle) under our control $\boldsymbol{J}_f$, a term due to magnetisation currents $\boldsymbol{J}_M = \nabla\times\boldsymbol{M}$ (see 4.17), and a term due to changing polarisation $\boldsymbol{J}_P = \partial\boldsymbol{P}/\partial t$ (see 4.5):

$$\boldsymbol{J} = \boldsymbol{J}_f + \boldsymbol{J}_M + \boldsymbol{J}_P = \boldsymbol{J}_f + \nabla\times\boldsymbol{M} + \frac{\partial\boldsymbol{P}}{\partial t}. \tag{4.31}$$

Substituting this into Ampère's law (Eq. 2.28), moving the curl of the magnetisation over to the left hand side and dividing through by μ_0 gives

$$\nabla\times\left(\frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M}\right) = \boldsymbol{J}_f + \frac{\partial\boldsymbol{P}}{\partial t} + \epsilon_0\frac{\partial\boldsymbol{E}}{\partial t}. \tag{4.32}$$

We define a new vector field $\boldsymbol{H}$, the "magnetic field strength" (cf the "magnetic flux density" $\boldsymbol{B}$) by

$$\boxed{\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M} \qquad \text{or} \quad \boldsymbol{B} = \mu_0 \left(\boldsymbol{H} + \boldsymbol{M}\right).} \tag{4.33}$$

The left hand side of Equation 4.32 can be written $\nabla \times \boldsymbol{H}$, while the final two terms on the right hand side are $\partial \boldsymbol{D}/\partial t$. We obtain the Maxwell–Ampère equation in the form

$$\boxed{\nabla \times \boldsymbol{H} = \boldsymbol{J}_f + \frac{\partial \boldsymbol{D}}{\partial t}.} \tag{4.34}$$

Finally, we can see why the final term is called the "displacement current density".

As a consistency check on the form of the Maxwell-Ampére equation 4.34, let's take its divergence:

$$\nabla \cdot \nabla \times \boldsymbol{H} = \nabla \cdot \left(\boldsymbol{J}_f + \frac{\partial \boldsymbol{D}}{\partial t}\right). \tag{4.35}$$

The divergence of a curl (left-hand side) is zero, so we need

$$\nabla \cdot \left(\boldsymbol{J}_f + \frac{\partial \boldsymbol{D}}{\partial t}\right) = \nabla \cdot \boldsymbol{J}_f + \frac{\partial \nabla \cdot \boldsymbol{D}}{\partial t} = \nabla \cdot \boldsymbol{J}_f + \frac{\partial \rho_f}{\partial t} = 0. \tag{4.36}$$

This must be correct. It is the statement, using the continuity equation for the free charges (see 2.4), that charge is conserved. We could have started from $\nabla \times \boldsymbol{H} = \boldsymbol{J}_f$ and deduced the need for the displacement current term as we did for the free space version of the equation.

We can often say that

$$\boldsymbol{M} = \chi_m \boldsymbol{H} \quad \text{and} \quad \boldsymbol{B} = \mu_0 (1 + \chi_m) \boldsymbol{H} \equiv \mu_0 \mu_r \boldsymbol{H}. \tag{4.37}$$

Here χ_m is called the magnetic susceptibility and μ_r the relative permeablility. It is strictly correct only for small applied fields $\boldsymbol{H}$. For larger applied fields, there are non-linear corrections. Materials can also be non-isotropic, in which case the response of the system (its magnetisation) is not parallel to the applied field. Materials can have negative χ_m, typically with $\chi_m \sim -10^{-6}$. If so they are called diamagnetic. If they have positive χ_m, with $\chi_m \sim 10^{-4}$, they are called paramagnetic. There also ferromagnetic materials, which have $\mu_r \sim 10^2 - 10^4$, and will order magnetically in small applied fields. These materials are used in permanent magnets.

Summary

The solenoidal condition and the Maxwell–Faraday equation do not require changing as they involve no source terms (charge, current) and hence we arrive at

$$\nabla \cdot \boldsymbol{D} = \rho_f \tag{4.38}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{4.39}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{4.40}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}_f + \frac{\partial \boldsymbol{D}}{\partial t} \tag{4.41}$$

along with

$$\boxed{\begin{aligned} \boldsymbol{D} &= \epsilon_0 \boldsymbol{E} + \boldsymbol{P} \\ \boldsymbol{B} &= \mu_0 \left(\boldsymbol{H} + \boldsymbol{M}\right) \end{aligned}} \qquad \begin{aligned} (4.42) \\ (4.43) \end{aligned}$$

> These are Maxwell's equations in matter. As with their free space equivalents, you should just know them.

Maxwell's equations in matter may appear no more complex than they do in a free space, but this is an illusion because in addition to these equations we also require "constitutive" relations between $\boldsymbol{P}$ and $\boldsymbol{E}$ and $\boldsymbol{M}$ and $\boldsymbol{H}$ (or equivalently between $\boldsymbol{D}$ and $\boldsymbol{E}$ and between $\boldsymbol{B}$ and $\boldsymbol{H}$). We will often assume simple linear, isotropic relations of the form of Eq. 4.9, but things can be more complex in practice and can even depend upon the past history of the application of $\boldsymbol{E}$ or $\boldsymbol{H}$ (the phenomenon of hysteresis in ferroelectric and ferromagnetic materials). One cannot always assert that $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{H})$ and $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{E})$.

Studying the response of materials to applied fields, both magnetic and electric, and to time-dependent em fields (spectroscopy), are amongst the most important tools we have for studying the properties of materials.

4.4 Boundary conditions on EM fields

Maxwell's equations are partial differential equations (PDEs). If we want to describe phenomena involving EM fields we will need to solve them. As with almost all discussions of PDEs, we are actually more interested in boundary value problems (PDEs with boundary conditions) than just the PDEs. Simple physical examples include light passing from air to glass, or a magnetic field emerging from a ferromagnet, or a dielectric sphere in an externally applied field. Usually we write sets of relations between the fields in each medium derived from Maxwell's equations and use the conditions at the boundaries to match solutions in the different media. You will have encountered a similar approach in first year modules when handling the reflection of waves, and should recognise the mathematics used later on in this module.

Integrating the Maxwell–Faraday equation (Eq. 2.16) around a fixed, closed loop gives us back its integral form

$$\oint_S \nabla \times \boldsymbol{E} \cdot d\boldsymbol{S} = \int_C \boldsymbol{E} \cdot d\boldsymbol{\ell} = -\oint_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot d\boldsymbol{S}. \tag{4.44}$$

Now consider a small thin rectangular loop that crosses the boundary between two media (Fig. 4.6) with its long sides running parallel to the boundary and short sides perpendicular to the boundary. Let the short sides tend to zero length, so that only the long sides will contribute to the integral:

$$\int \boldsymbol{E} \cdot d\boldsymbol{\ell} = (E_{2,\|} - E_{1,\|})L + \mathcal{O}(d). \tag{4.45}$$

Here L is length of along the interface. The last term indicates corrections of order d or higher where d is the depth of the box across the boundary. A "small loop" means that L is small enough that both $\boldsymbol{E}$ and the time-derivative of $\boldsymbol{B}$ field can be taken as constant along L.

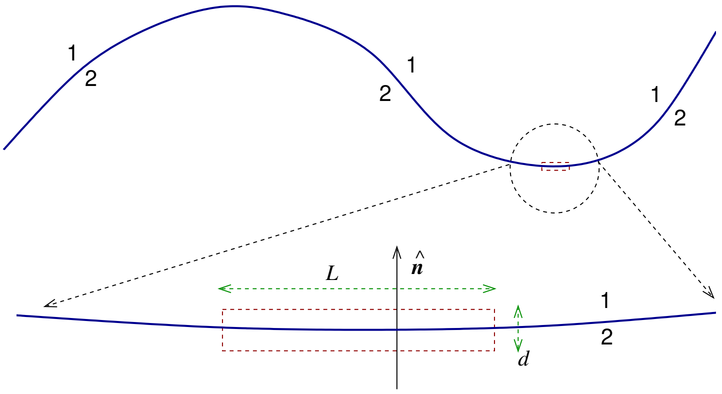

Figure 4.6: The boundary between two homogeneous media with a small rectangular loop arranged to run parallel to the boundary. The depth d should be thought as being small enough that any contributions to the line integral perpendicular to the boundary are negligible. L should be sufficiently small that the field can be taken to be constant along interface. The black arrow defines the direction of the unit vector $\hat{\boldsymbol{n}}$ perpendicular to the boundary which we will take to point from medium 2 into medium 1.

This gives

$$(E_{1,\parallel} - E_{2,\parallel})L = -Ld\frac{\partial B_\parallel}{\partial t} + \mathcal{O}(d). \tag{4.46}$$

The subscript $\parallel$ denotes the field component parallel to the interface. The component of $\boldsymbol{B}$ is parallel to the interface but in the direction of the normal to the loop. This is perpendicular both to the interface and the electric field components. For finite time derivative of $\boldsymbol{B}$, the limit $d \to 0$ gives

$$\boxed{\boldsymbol{E}_{2,\parallel} = \boldsymbol{E}_{1,\parallel}.} \tag{4.47}$$

The result 4.47 relates vectors $\boldsymbol{E}_\parallel$ parallel to the interface. The condition applies in any direction within the (2D) boundary. There are therefore two independent conditions. In words the result says that the components of the electric fields either side of and *parallel* to a boundary are the same. If you don't like the extra notation, "$\parallel$" in 4.47, the statement

$$\boxed{\hat{\boldsymbol{n}} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) = \boldsymbol{0},} \tag{4.48}$$

is equivalent and does not need this added notation. ($\hat{\boldsymbol{n}}$ is a unit vector perpendicular to the boundary, as shown in Fig. 4.6.)

The boundary condition on $\boldsymbol{D}$ is set by Gauss's law written in the form of Eq. 4.38, which we integrate over volume using Gauss's theorem:

$$\oint \boldsymbol{D} \cdot d\boldsymbol{S} = \int \rho_f \, dV = Q_f. \tag{4.49}$$

Now consider a pillbox[3] straddling the boundary as shown in Fig. 4.7. If the box is made squat enough, the only significant flux is that entering or leaving via the top and bottom sides, so

$$\oint \boldsymbol{D} \cdot d\boldsymbol{S} = (D_{1,\perp} - D_{2,\perp})\, A = \rho_f A h + \sigma_f A. \tag{4.50}$$

[3] You should be thinking of a very squat cylinder, flat on its top and bottom, much shorter than it is wide

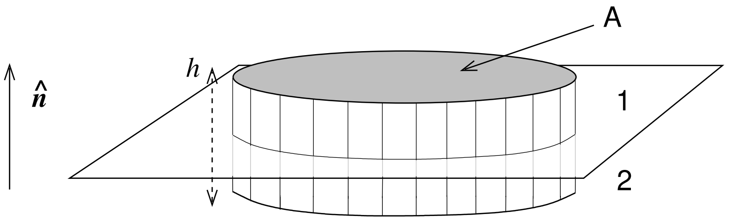

Figure 4.7: A squat cylinder ("pillbox") of height h and cross-sectional area A straddles the boundary between two homogeneous media with its flat faces oriented parallel to the boundary. The height h should be thought of in the limit $h \to 0$.

A is the area of the top and bottom faces of the box, h its depth. The symbol $\perp$ denotes the component perpendicular to the surface. We have split the free charges in the box into a smooth volume distribution and a surface contribution. Letting $h \to 0$ gives

$$\boxed{D_{1,\perp} - D_{2,\perp} = \sigma_f.} \tag{4.51}$$

The result 4.51 shows that the change in the component of $\boldsymbol{D}$ perpendicular to a boundary is determined by the surface density of free surface charge. The order of D_1 and D_2 in this expression implies we are measuring the components in a specific direction. In particular, we are considering flux of D entering the pillbox on the medium 1 side of the boundary, and leaving on the medium 2 side of the boundary—we are resolving in the upwards direction of Fig. 4.7. A more concise statement which some may prefer is

$$\boxed{\hat{\boldsymbol{n}} \cdot (\boldsymbol{D}_1 - \boldsymbol{D}_2) = \sigma_f,} \tag{4.52}$$

where the unit vector $\hat{\boldsymbol{n}}$ is shown in Fig. 4.7.

The condition on $D_\perp$ is a single scalar condition (there is only one component of $\boldsymbol{D}$ perpendicular to the interface). There is an equivalent condition on $E_\perp$, but with the "σ" on the right hand side being the total surface charge density. This would include both free and polarisation charges. While in many problems the free charges can be taken to be zero, making the condition on $D_\perp$ straightforward, this may not be true for polarisation charge density which we would then have to compute. The condition on $D_\perp$ is therefore one most often quoted and (almost always) used.

The pillbox shape of Fig. 4.7 and the argument used to derive the condition on $\boldsymbol{D}_\perp$ can also be applied to deduce the boundary condition on $\boldsymbol{B}$. We start with 4.39: $\nabla \cdot \boldsymbol{B} = 0$ (this is the same equation as $\nabla \cdot \ D = \rho_f$ but with the right hand side equal to zero). By analogy with 4.51 and 4.52 we obtain

$$\boxed{B_{1,\perp} = B_{2,\perp} \qquad \text{or} \quad \hat{\boldsymbol{n}} \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) = 0.} \tag{4.53}$$

The number of flux lines of $\boldsymbol{B}$ per unit area crossing the boundary is the same on both sides.The component of $\boldsymbol{B}$ perpendicular to the boundary therefor matches on each side.

The boundary condition on $\boldsymbol{H}$ can be derives in a similar way to that used for $\boldsymbol{E}$. We start with the Maxwell–Ampère equation $\nabla \times \boldsymbol{H} = \boldsymbol{J}_f + \partial_t \boldsymbol{D}$ and consider a rectangular loop oriented parallel to the boundary as in Fig. 4.6. When showing that $\boldsymbol{E}_{2,\parallel} = \boldsymbol{E}_{1,\parallel}$, we were able to drop the term $-Ld\partial_t B_{\|}$ by allowing the distance across the boundary $d \to 0$, on the (implicit) understanding that $\partial_t B_{\|}$ in the boundary is finite. For $\boldsymbol{H}$ we can make the same assumption for $\partial_r \boldsymbol{D}$, but not necessarily for $\boldsymbol{J}_f$ as we have to allow for surface currents $\boldsymbol{j}_f$. These should be visualised as sheets of current running parallel to the boundary. For instance they might describe the situation when superconductors are involved. We need also to account for the vector directions and a little thought shows that

$$\boxed{\hat{\boldsymbol{n}} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{j}_f.} \tag{4.54}$$

In this case the concise notation is to be preferred, because it expresses the vector nature of the condition best, but in terms of magnitudes it says that the component of $\boldsymbol{H}$ parallel to the boundary changes by an amount equal to the surface current density.

You should note that there is really no such thing as a surface current sheet; such currents always have a finite thickness, even in superconductors. However there are cases where currents are confined to layers much thinner than other interesting scales in the problem (the wavelength of EM waves perhaps) making it convenient to picture a zero-thickness current sheet.

Charge–current boundary condition

Finally the conservation equation $\nabla \cdot \boldsymbol{J}_f + \partial_t \rho_f = 0$ for free charges leads, via another pillbox straddling the boundary similar to that shown in Fig. 4.7, to

$$\hat{\boldsymbol{n}} \cdot (\boldsymbol{J}_{2f} - \boldsymbol{J}_{1f}) + \frac{d\sigma_f}{dt} = 0. \tag{4.55}$$

This states that, if the current densities perpendicular to the boundary don't match, then there must be a rate of change of the charge density on the boundary. As with the boundary condition on $\boldsymbol{D}$, we have chosen the direction of $\hat{\boldsymbol{n}}$ is upwards in Fig. 4.7, i.e. we are considering the positive direction of the current flow to be from medium 1 to medium 2. If $J_{2f,\perp} > J_{1f,\perp}$ then $d\sigma_f/dt < 0$ as the above relation implies.

At least one of the two media involved here should be conducting for this condition to be of any interest. Currents in conductors are driven by the electric field and for homogeneous isotropic conductors, $\boldsymbol{J}$ and $\boldsymbol{E}$ are often linearly related to a good approximation which we will write as

$$\boldsymbol{J}_f = g\boldsymbol{E}, \tag{4.56}$$

where g is the conductivity (not "σ" as is often seen to avoid confusion with σ representing surface charge). We can thus replace the $\boldsymbol{J}$s that crop up in the charge–current boundary condition with equivalent $g\boldsymbol{E}$ terms. Also since σ_f is related to the change in the perpendicular component of $\boldsymbol{D}$, and assuming a simple linear relation between $\boldsymbol{D} = \epsilon\boldsymbol{E}$ we can replace the σ_f as well to deduce

$$\hat{\boldsymbol{n}} \cdot \left(g_2 \boldsymbol{E}_2 + \epsilon_2 \frac{d\boldsymbol{E}_2}{dt} - g_1 \boldsymbol{E}_1 - \epsilon_1 \frac{d\boldsymbol{E}_1}{dt} \right) = 0, \tag{4.57}$$

which says in words that the flux of $g\boldsymbol{E} + \epsilon\partial_t \boldsymbol{E}$ is conserved perpendicular to boundaries between (potentially) conducting media. This complicates the treatment of reflections at arbitrary angles from conductors. We will not look at such cases and restrict ourselves to the consideration of reflections from conductors at normal incidence only.

Summary

The EM fields on either side of a boundary should satisfy the following conditions:

$$\boxed{\begin{aligned} \hat{\boldsymbol{n}} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) &= \boldsymbol{0}, \\ \hat{\boldsymbol{n}} \cdot (\boldsymbol{D}_1 - \boldsymbol{D}_2) &= \sigma_f, \\ \hat{\boldsymbol{n}} \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) &= 0, \\ \hat{\boldsymbol{n}} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) &= \boldsymbol{j}_f. \end{aligned}} \qquad \begin{aligned} &(4.58)\\ &(4.59)\\ &(4.60)\\ &(4.61) \end{aligned}$$

There is also the charge conservation conditions, although these will be of less importance to what follows in these notes. In practice one does not need to apply all of these, as will be seen in the next chapter.

You need to remember these boundary conditions. Sometimes it helps to think of these in words such as "the components of $\boldsymbol{E}$ parallel to the boundary match" or "the perpendicular components of $\boldsymbol{D}$ are discontinuous by the surface charge density". If you can remember the definition of $\hat{\boldsymbol{n}}$ (pointing perpendicular to the boundary from medium 1 to medium 2), you might find it easier to remember the vector forms in full.

Chapter 5

Waves and energy density

The properties of plane EM waves in homogeneous media can be derived as they were in free space, see Section 3.1. We will find that Maxwell's equations adapted to macroscopic media allow us to understand some well-known properties of light in water and glass including the refractive index. We will also revisit the derivation of the Poynting vector and look at what it says about the energy stored and transmitted by waves in matter and in other cases.

5.1 EM Waves in dielectrics

Maxwell's equations as listed in Eqs 4.38 to 4.41, without any free charges or currents ($\rho_f = 0$, $\boldsymbol{J}_f = \boldsymbol{0}$), give:

$$\boxed{\nabla \cdot \boldsymbol{D} = 0, \qquad \nabla \cdot \boldsymbol{B} = 0, \qquad \nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \qquad \nabla \times \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t}.} \tag{5.1}$$

We will assume linear, isotropic polarisation and magnetisation so that

$$\boldsymbol{D} = \epsilon_0 \epsilon_r \boldsymbol{E}, \qquad \text{and} \qquad \boldsymbol{B} = \mu_0 \mu_r \boldsymbol{H}. \tag{5.2}$$

We take the curl of both sides of the third relation in 5.1 (as we did when deriving 3.8 and 3.9 in Chapter 3). We use the equation for $\nabla \times \boldsymbol{B}$ and the assumed relations between $\boldsymbol{H}$ and $\boldsymbol{B}$ and between $\boldsymbol{E}$ and $\boldsymbol{D}$. We find that all the fields obey the wave equation, but with the wave speed (phase velocity) given by

$$\boxed{\frac{1}{\sqrt{\mu_0 \mu_r \epsilon_0 \epsilon_r}} = \frac{c}{\sqrt{\mu_r \epsilon_r}}.} \tag{5.3}$$

We will always use c to denote the speed of light in a vacuum. The speed of light in a medium other than the vacuum is then expressed in terms of the refractive index, n, via $v_\phi = c/n$. We see from 5.3 that

$$\boxed{n = \sqrt{\mu_r \epsilon_r}.} \tag{5.4}$$

Most transparent media are only weakly magnetisable so that $\mu_r \approx 1$ and $n \approx \sqrt{\epsilon_r}$. In diamond for example we have $\epsilon_r \approx 5.7$ so we would expect $n \approx \sqrt{5.7} = 2.39$. The measured value is 2.417 which agree quite well with the prediction. In water, on the other hand, $\epsilon_r \approx 80 \Rightarrow n \approx 9$ while the measured value is $n \approx 1.33$. We need to be careful. The statements above assume that ϵ_r and μ_r are constants. This is far from the case because the response of media to applied fields is not instantaneous. In water the large value of ϵ_r is a low frequency result and reflects the polarisation following from the reorientation of the polar molecules. The refractive index for visible light characterises the behaviour at optical frequencies around 6×10^{14}Hz. At these high frequencies the molecules cannot reorient on the timescales involved.

As ϵ_r and μ_r are actually frequency dependent, we shouldn't have treated them as constants. Instead we should consider plane waves which have a well-defined frequency (waveform $\sim e^{-i\omega t}$). Any wave pulse can be constructed as a linear superposition of plane waves. We assume plane wave solutions:

$$\boldsymbol{E} = \boldsymbol{E}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \quad \boldsymbol{B} = \boldsymbol{B}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)} \quad \boldsymbol{D} = \boldsymbol{D}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \quad \boldsymbol{H} = \boldsymbol{H}_0 \mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}. \tag{5.5}$$

We can replace all derivatives (see 3.15), ∂_t, by $-i\omega$, and the action of ∇ by that of $i\boldsymbol{k}$ to find

$$\boxed{\begin{aligned} \boldsymbol{k}\cdot\boldsymbol{D} &= 0, & (5.6)\\ \boldsymbol{k}\cdot\boldsymbol{B} &= 0, & (5.7)\\ \boldsymbol{k}\times\boldsymbol{E} &= \omega\boldsymbol{B}, & (5.8)\\ \boldsymbol{k}\times\boldsymbol{H} &= -\omega\boldsymbol{D}. & (5.9)\end{aligned}}$$

With the linear isotropic relations between $\boldsymbol{D}$ and $\boldsymbol{E}$ and between $\boldsymbol{H}$ and $\boldsymbol{B}$, but with frequency dependent ϵ_r and μ_r in 5.2, we can conclude that EM waves in media are transverse. $\boldsymbol{E}$ and $\boldsymbol{D}$ are perpendicular to $\boldsymbol{B}$ and $\boldsymbol{H}$, and both are perpendicular to $\boldsymbol{k}$.

The ratio E/B generalises slightly from the result 3.20 to become:

$$\frac{E}{B} = \frac{\omega}{k} = v_\phi = \frac{c}{n}, \tag{5.10}$$

where n is the refractive index. A related version of this is the ratio of E to H. In terms of magnitudes, for linear isotropic media we can write $kE = \omega B = \mu_0\mu_r\omega H$, so

$$\frac{E}{H} = \mu_0\mu_r\frac{\omega}{k} = \mu_0\mu_r v_\phi \equiv Z = \sqrt{\frac{\mu_0\mu_r}{\epsilon_0\epsilon_r}} = Z_0\sqrt{\frac{\mu_r}{\epsilon_r}}, \tag{5.11}$$

where Z is called the "impedance". The quantity Z_0,

$$\boxed{Z_0 = \mu_0 c = \sqrt{\frac{\mu_0}{\epsilon_0}} = 376.73\,\Omega,} \tag{5.12}$$

is known as the "impedance of free space". It has units of resistance (Ω) because the electric field $\boldsymbol{E}$ has units of V m^{-1} while the magnetic field strength $\boldsymbol{H}$ has units of A m^{-1}.

The quantity Z_0 appears regularly, particularly in formulae connected to dipole radiation. The dimensionless factor $\sqrt{\mu_r/\epsilon_r}$ in 5.12 means that the impedance Z changes across boundaries (at least one of ϵ_r and μ_r changes when moving from one medium to another). You should have come across the notion of "impedance" in the first year module on Foundations as a quantity that determines reflectance at boundaries. We will see it plays the same role for light. In terms of the refractive index, n, the impedance can be expressed as

$$Z = \frac{\mu_r}{n} Z_0, \tag{5.13}$$

which, commonly for transparent materials for which $\mu_r \approx 1$, reduces to $Z \approx Z_0/n$.

Light Polarisation

Since they are transverse, EM waves have two separate modes. For a wave travelling in the $\hat{\boldsymbol{z}}$ direction, the electric field can have components along the x- or y-axes. In practice this is seen through the phenomenon of "polarisation" (*not* to be confused with dielectric polarisation!). We say that there are two independent components of polarisation. Polarised light interacts with interfaces and optical devices in different ways according to the state of polarisation and the device orientation. A full description of polarised (and unpolarised) light depends on the temporal variation of phase within the two components and goes beyond what we need to cover here (look up "Stokes parameters" and "coherence" for more on this interesting topic). Our main interest in polarisation will be how it affects reflection at interfaces. We will see that the results depend upon field orientation and can induce polarisation in previously unpolarised light.

5.2 Energy density and flux:the Poynting vector II

The idea of energy stored in electromagnetic fields was introduced in Chapter 3, section 3.2. This short section is really the continuation of what was covered there. The derivation of the Poynting vector, which gives the energy flux in the EM field, for systems described by the version of Maxwell's equations involving matter strictly contains no new information. However the formulation in terms of the additional fields, $\boldsymbol{D}$ and $\boldsymbol{H}$, helps understand a number of useful and amusing phenomena. Results we will look at include treatments of capacitors and inductors, and currents in metals.

Work done by electromagnetic fields

At a point where a current density density $\boldsymbol{J}_f$ flows and there is an electric field $\boldsymbol{E}$, the electric field performs work at a rate of $\boldsymbol{E} \cdot \boldsymbol{J}_f$ per unit volume on the free charges. The law of conservation of energy, 3.24, becomes

$$\begin{aligned}
\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{S} = -\boldsymbol{E} \cdot \boldsymbol{J}_f = -\boldsymbol{E} \cdot \left(\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} \right) && (5.14) \\
= \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) - \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} && (5.15) \\
= \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t} + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} && (5.16)
\end{aligned}$$

We have used the identity 1.20[1] (to obtain line 2) and the Maxwell equation 4.40 (to obtain line 3) just as we did to obtain 3.25. Expressions for u and $\boldsymbol{S}$ are what we want to find. They are energy density and energy flux associated with the EM fields.

[1] Setting $\boldsymbol{V} = \boldsymbol{E}$ and $\boldsymbol{W} = \boldsymbol{H}$ in 1.20 we have $\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) = \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) - \boldsymbol{E} \cdot (\nabla \times \boldsymbol{H})$

If the material is linear ($\boldsymbol{D} \propto \boldsymbol{E}$ and $\boldsymbol{H} \propto \boldsymbol{B}$) then $\boldsymbol{E} \cdot d\boldsymbol{D} = d(\boldsymbol{E} \cdot \boldsymbol{D})/2$ and $\boldsymbol{H} \cdot d\boldsymbol{B} = d(\boldsymbol{H} \cdot \boldsymbol{B})/2$, and

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{S} = \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) + \frac{\partial}{\partial t}\frac{1}{2}(\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{H} \cdot \boldsymbol{B}). \tag{5.17}$$

We identify

$$\boxed{u = \frac{1}{2}\boldsymbol{E} \cdot \boldsymbol{D} + \frac{1}{2}\boldsymbol{H} \cdot \boldsymbol{B},} \tag{5.18}$$

as the *energy density* (units of energy per unit volume) and

$$\boxed{\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H},} \tag{5.19}$$

as the Poynting vector. Eq. 5.17, along with the particular forms for u and $\boldsymbol{S}$, is Poynting's theorem. Note that it is invariant to changing $\boldsymbol{S}$ to $\boldsymbol{S} + \nabla \times \boldsymbol{W}$ where $\boldsymbol{W}$ is some arbitrary vector field, since the divergence of a curl is zero, but $\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$ is the conventional choice.

Be careful to avoid confusion with the symbol for area element $d\boldsymbol{S}$. Note also that for linear isotropic media $\boldsymbol{E} \cdot \boldsymbol{D}/2$ can be expressed alternatively as $\epsilon E^2/2 = D^2/2\epsilon$ and $\boldsymbol{B} \cdot \boldsymbol{H}/2 = B^2/2\mu = \mu H^2/2$.

Finally, we need to be wary of the difference between the result for the Poynting flux in 5.19 and 3.28. We have used $\boldsymbol{S}$ rather than $\boldsymbol{P}$, which we used in Section 3.2, for the Poynting vector. The two coincide in free space where $\boldsymbol{B} = \mu_0 \boldsymbol{H}$, whereas in matter they differ. In 5.18 and 5.19, the energy stored in the motion and polarisation of bound charges has been attributed to the fields. Equations 5.18 and 5.19 are the normal statements of the results and $\boldsymbol{S}$ is the usual symbol. It is how Poynting originally derived them. These results are 'boxed' and you should learn them. In free space (with no bound charges) these boxed results are safe to use as $\boldsymbol{B} = \mu_0 \boldsymbol{H}$ and $\boldsymbol{D} = \epsilon_0 \boldsymbol{E}$.

5.2.1 Poynting flux in EM waves in dielectrics

The identification of $\boldsymbol{E} \times \boldsymbol{H}$ as the energy flux in EM fields, finds its most useful application in the study of EM waves. We learnt in Section 5.1 that $\boldsymbol{E}$, $\boldsymbol{B}$ and the wave vector $\boldsymbol{k}$ are mutually perpendicular. The Poynting vector $\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$ is therefore parallel to $\boldsymbol{k}$ as we found for the vacuum case (see 3.29). (Once we assume the linear relation between the fields $\boldsymbol{D} \propto \boldsymbol{E}$ and $\boldsymbol{H} \propto \boldsymbol{B}$ this is actually guaranteed by 3.29.)

Having established that the direction of $\boldsymbol{S}$ is that of $\boldsymbol{k}$, we can work with field magnitudes. Since $\boldsymbol{E}$ and $\boldsymbol{H}$ are perpendicular,

$$S = EH = \frac{E^2}{Z} = ZH^2, \tag{5.20}$$

where we have used the impedance $Z = E/H$. Remembering also that E and H vary sinusoidally, i.e. at any fixed point they vary with time as $E = E_0 \sin(\omega t)$, $H = H_0 \sin(\omega t)$, the time-averaged flux in an EM wave is given by

$$\bar{S} = \langle S \rangle = \frac{E_0^2}{2Z} = \frac{ZH_0^2}{2}. \tag{5.21}$$

We have used $\left\langle \sin^2(\omega t) \right\rangle = 1/2$. Here E_0 and H_0 are the amplitudes of the electric and magnetic fields.

It is also common to see the formulae of 5.21 expressed in terms of "root-mean-square" or "RMS" amplitudes, such as $E_{\mathrm{RMS}} = E_0/\sqrt{2}$, giving

$$\bar{S} = \frac{E_{\mathrm{RMS}}^2}{Z} = Z H_{\mathrm{RMS}}^2. \tag{5.22}$$

The standard AC mains voltage in the UK, 230 V, is also an RMS value so that its amplitude is $\sqrt{2} \times 230 =$ 325 V. RMS values are also used in more complex cases such as sunlight which is the superposition of an infinity of different frequencies such that one can no longer define "amplitude" as a maximum value. In fact, the field strength measured at multiple times on a timescale much longer than the typical wave period at any point in sunlight has a Gaussian distribution, with no easily definable maximum. The time average of the squared field strength, on the other hand, is well-defined so that $E_{\mathrm{RMS}} = \sqrt{\langle E^2 \rangle}$ retains meaning, as do the above formulae involving RMS values.

Example 5.1. Estimate the electric field strength of sunlight at Earth.

Answer. The energy flux of sunlight at Earth is $\bar{S} =$ 1300 W/m^2. The impedance of free space (similar to that of air) is $Z_0 =$ 377 Ω, thus $E_{\mathrm{RMS}} = \sqrt{Z_0 \bar{S}} =$ 700 V m^{-1}.

Example 5.2. A free electron oscillates in an EM wave of frequency 1 GHz and intensity 10 kW m^{-2}. Calculate the amplitude of its motion.

Answer. Newton's second law (will need to verify that we are OK to do so) gives:

$$m\ddot{x} = qE_0 e^{i\omega t}. \tag{5.23}$$

Setting $x = ae^{i\omega t}$, we find

$$a = \frac{qE_0}{m\omega^2}, \tag{5.24}$$

where q is the charge on the electron, m is its mass and E_0 the amplitude of the electric field. We determine E_0 using $\bar{P} = E_0^2/2Z$ to find $E_0 =$ 2476 V m^{-1} (assuming $Z = Z_0$), and thus

$$a = \frac{1.6 \times 10^{-19} \times 2746}{9.1 \times 10^{-31} \times (2\pi \times 10^9)^2} = 1.22 \times 10^{-5}\,\mathrm{m}. \tag{5.25}$$

The speed of the electron $v = a\omega =$ 7.7 × 10^4 m s^{-1} $\ll c$, so we are well below relativistic speeds. This also ensures that we can neglect the magnetic part of the Lorentz force ($q\boldsymbol{v} \times \boldsymbol{B}$) relative to $q\boldsymbol{E}$. $B = E/c$ in an EM wave in a vacuum (see 3.20), so $|\boldsymbol{v} \times \boldsymbol{B}| = (v/c)E \ll E$ as $v \ll c$. A second effect we are ignoring here is radiation by the accelerated electron. This will be a small effect in this case, but it is also beyond the remit of PX263 and instead a topic for next year: PX384, "Electrodynamics".

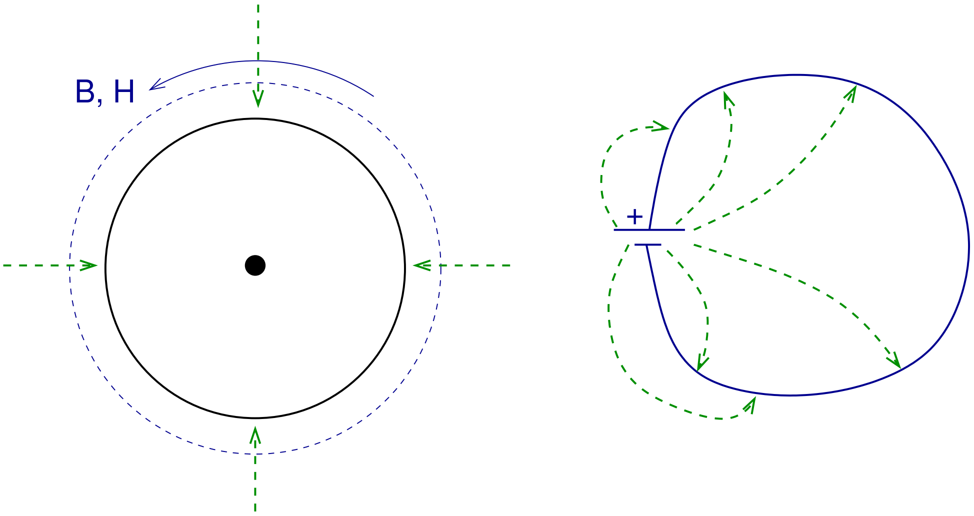

Figure 5.1: *Left:* A wire of circular cross-section with a current flowing up out of the page. The electric field needed to drive the current also points out of the page. The magnetic field points as shown. The Poynting vector flux, $\boldsymbol{E} \times \boldsymbol{H}$, points in towards the wire as indicated by the green arrows. *Right:* Expanding out to see the complete circuit shows a battery driving a current through a loop of wire. The Poynting flux flows out of the battery (where the electric field reverses direction compared to the wire) and then into the wire where heat is dissipated.

5.2.2 Poynting vector in static cases

For static fields, the Poynting vector can lead to some counterintuitive results. Consider the electric and magnetic fields around a wire carrying a current (Fig. 5.1). The Poynting flux points inwards at the surface of the wire, and looking on a larger scale, flows out of the source of power (a battery in this case). This might seem rather unexpected but the total Poynting flux pointing into the wire can be shown to match the rate of heat dissipation. We can show this explicitly for the case of a cylindrical length of wire L of resistance R and radius a with voltage V applied along it. A current $I = V/R$ will flow and the rate of dissipation in the wire $= I^2R = V^2/R$. The electric field strength $E = V/L$. The surface magnetic flux density $B = \mu_0 I/2\pi a$, so that $H = I/2\pi a$ (as can be deduced from Ampère's law based upon $\boldsymbol{H}$). $\boldsymbol{E}$ and $\boldsymbol{H}$ are perpendicular so the magnitude of the Poynting flux $P = EH$ is given by

$$S = EH = \frac{V}{L}\frac{I}{2\pi a} = \frac{V^2}{2\pi LaR}. \tag{5.26}$$

The energy flux flows inwards over the curved cylindrical surface of the wire, which has area $= 2\pi La$. Thus the total energy flux flowing into the wire is $= V^2/R$, which is the dissipation rate calculated earlier.

This result is what Poynting's theorem says, but it does raise some questions. If one places a conductor near (but not touching) the wire, disturbing the electric field (which must be zero inside the conductor), won't the Poynting flux be affected? Partly the answer is yes: as a conductor is moved into place, there will be some effect upon the current in the circuit through inductive effects, but once it settles down, there should not be any permanent change. The concept of the Poynting flux is less helpful in this case because the total flux has to be constant regardless of what is placed external to wire in the Poynting "flow".

5.3 Radiation pressure

Momentum is carried with the energy flow represented by the Poynting flux. In the case of EM waves, the easiest way to see this is the relativistic result $E = pc$, which implies that an energy flux $\boldsymbol{S}$ carries a momentum flux $\boldsymbol{S}/c = \boldsymbol{E} \times \boldsymbol{H}/c$ (this result does not generalise to all types of fields). What we are calling the momentum flux has dimensions of momentum per unit time per unit area, which is the same as pressure and we call this flux the radiation pressure.

Radiation pressure is important in very massive stars. A limit can be reached in which the effects of the force (mainly on electrons) from the radiation pressure can exceed that of gravity (which remains the dominant force on protons), leading to rapid mass loss and limits the overall power of the star. Radiation pressure effects similarly limit the maximum rate at which material can fall onto black holes, and reduce their electromagnetic luminosity. At some point the rate of release of gravitational potential energy liberated in the form of radiation is so large that the radiation pressure exceeds the gravitational attraction and the supply of material is truncated. Look up the "Eddington limit" to see more on this.

Sunlight arriving at the Earth carries an energy flux of $\approx 1300\,\mathrm{W\,m^{-2}}$. The corresponding radiation pressure $1300/c = 4.3 \times 10^{-6}\,\mathrm{N\,m^{-2}}$, which won't exactly knock you off your feet, but can have a significant effect on interplanetary dust grains.[2]

One can show that EM fields carry an associated momentum density, i.e. a momentum per unit volume given by

$$\boldsymbol{p} = \frac{\boldsymbol{S}}{c^2} = \frac{\boldsymbol{E} \times \boldsymbol{H}}{c^2}. \tag{5.27}$$

The overall momentum of EM fields in a volume V is given by

$$\boldsymbol{p} = \int_V \frac{\boldsymbol{E} \times \boldsymbol{H}}{c^2}\, dV. \tag{5.28}$$

Some of the rather complicated motivation for this is discussed in appendix H. An interesting static case to consider is a point charge in a uniform magnetic field. If you think about the Poynting flow and the momentum density formula for this case, you should be able to convince yourself that the fields contain angular momentum around an axis parallel to the magnetic field running through the position of the charge. Now consider moving a charge from far away from a region of magnetic field into the field. If the EM field ends up with angular momentum, where has it come from? You shouldn't spend too long worrying about this though; as said before, the Poynting vector is of most use for EM waves, and beyond this seems to lead to odd toy puzzles of this sort in static cases which don't help you solve many problems of practical interest.

[2] Look up "Crook's radiometer" for a thing once erroneously ascribed to radiation pressure, but which is nothing of the sort. You may have seen one of these at some point in your life.

Chapter 6

EM Waves in Matter with Boundaries

The importance of boundary conditions (see Section 4.4) for matching solutions in different regions is perhaps never better illustrated than the case of light impinging on a dielectric interface. This is of great practical importance as it applies to all optical systems. You could be thinking of light moving from air to water, water to air, air to glass. We know the broad-brush result of course: some light is transmitted from the first to the second medium, undergoing a change of direction in general (refraction) and some is reflected. By applying Maxwell's equations we can understand this quantitatively. Rather remarkably, the formulae we are about to derive were first obtained by Fresnel in the early 1800s before Maxwell's equations had appeared.

You may ask why waves in metals first appear in this chapter about EM waves in systems with boundaries. Why not earlier? In conductors, we will see that waves can only propagate close to a boundary with a dielectric or with free space. Without a boundary in metals, there are no EM waves. This is because the conduction electrons in the metals absorb the energy from the wave. The electric field can only be non-zero within a region close to the boundary of the metal. The thickness of this region is called the skin depth and is determined by the conductivity of the material.

6.1 Reflection and transmission at dielectric interfaces: Fresnel coefficients

Consider an infinite, plane wave travelling in dielectric medium 1, towards a plane boundary with dielectric medium 2. The angle of incidence (measured with respect to the normal to the boundary) will be denoted i, the angle of reflection r and the angle of transmission t (Fig. 6.1). You probably know from other modules that $i = r$ while $n_1 \sin i = n_2 \sin t$ (Snell's law), but we will derive this here. In each medium plane wave solutions can exist. We need them to match at the boundary, at all places and for all time, by enforcing the boundary conditions 4.58 to 4.61 and by ensuring that the phases keep in step.

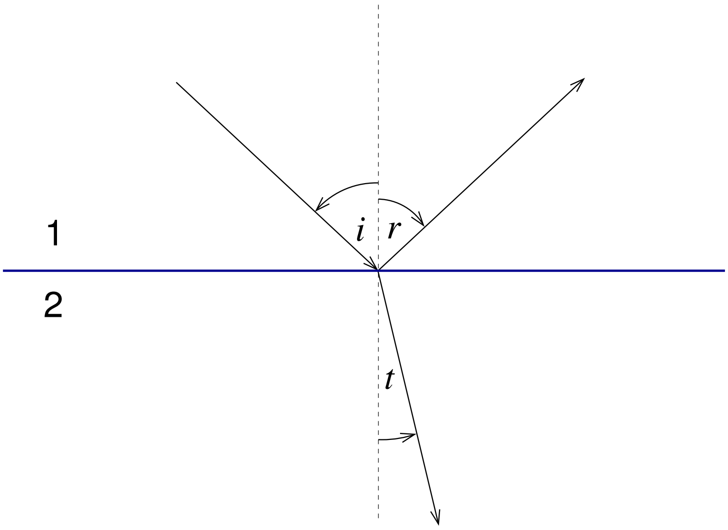

Figure 6.1: Definition of the angles of incidence i, reflection r and transmission t for reflection and refraction at the boundary between two homogeneous media. The incident wave comes in from the upper-left. The diagram shows a single ray impinging upon a single point on the boundary, but it should be imagined that the waves extend over the boundary.

Matching phases—Snell's Law

The incident, reflected and transmitted waves take the form

$$\boldsymbol{E}_i \exp i(\boldsymbol{k}_i \cdot \boldsymbol{r} - \omega_i t), \tag{6.1}$$

$$\boldsymbol{E}_r \exp i(\boldsymbol{k}_r \cdot \boldsymbol{r} - \omega_r t), \tag{6.2}$$

$$\boldsymbol{E}_t \exp i(\boldsymbol{k}_t \cdot \boldsymbol{r} - \omega_t t). \tag{6.3}$$

where their spatial and time dependence is contained within the exponential phase factors. Assuming we can match the fields at some point on the boundary at some particular time, it is the phases that determine whether the conditions will match at other places and other times. This is a generic feature of *all* plane waves, not just EM waves, and thus the results we are about to deduce are generic (they apply to sound and water waves too). We need $\omega_i t = \omega_r t = \omega_t t$ for any time t. We deduce that

$$\omega_i = \omega_r = \omega_t. \tag{6.4}$$

The frequency is therefore unchanged by reflection or transmission, and we will therefore drop the $-i\omega t$ terms in the exponents.

We also require that $\boldsymbol{k}_i \cdot \Delta\boldsymbol{r} = \boldsymbol{k}_r \cdot \Delta\boldsymbol{r} = \boldsymbol{k}_t \cdot \Delta\boldsymbol{r}$ for any change in position $\Delta\boldsymbol{r}$ that lies in the plane boundary itself. Therefore

$$(\boldsymbol{k}_r - \boldsymbol{k}_i) \cdot \Delta\boldsymbol{r} = (\boldsymbol{k}_t - \boldsymbol{k}_i) \cdot \Delta\boldsymbol{r} = 0, \tag{6.5}$$

for all $\Delta\boldsymbol{r}$ parallel to the boundary, which implies that the difference vectors $\boldsymbol{k}_r - \boldsymbol{k}_i$ and $\boldsymbol{k}_t - \boldsymbol{k}_i$ are perpendicular to the boundary, as shown diagrammatically in Fig. 6.2. This is the requirement that

$$\hat{\boldsymbol{n}} \times \boldsymbol{k}_i = \hat{\boldsymbol{n}} \times \boldsymbol{k}_r = \hat{\boldsymbol{n}} \times \boldsymbol{k}_t, \tag{6.6}$$

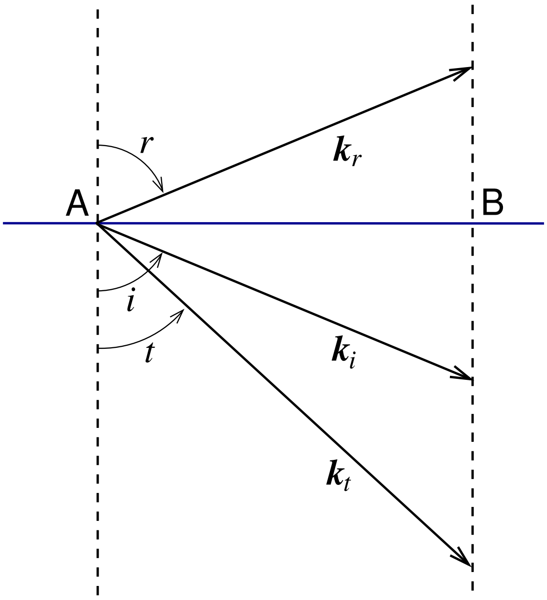

Figure 6.2: The figure shows the condition on the incident, reflected and transmitted wave vectors imposed by the need to match phases across the entire boundary: the difference between any two of the wave vectors is perpendicular to the interface.

where $\hat{\boldsymbol{n}}$ is the usual vector perpendicular to the boundary. The three wavevectors ($\boldsymbol{k}_i, \boldsymbol{k}_i$ and $\boldsymbol{k}_i$) have the same component parallel to the interface. This implies that

$$k_i \sin i = k_r \sin r = k_r \sin t. \tag{6.7}$$

However since $\omega/k = c/n$, where n is the refractive index, and ω the same in all cases. $k \propto n$, and

$$n_1 \sin i = n_1 \sin r = n_2 \sin t. \tag{6.8}$$

Equation 6.7 allows us deduce the laws of reflection and refraction

$$\boxed{\begin{gathered} i = r, \\ n_1 \sin i = n_2 \sin t. \end{gathered}} \tag{6.9}$$

All three wave vectors lie in the plane defined by $\boldsymbol{k}_i$ and $\hat{\boldsymbol{n}}$ (they have the same component parallel to the boundary, see 6.6). This is the plane shown in Figs 6.1 and 6.2.

If the boundary conditions can be met at one place on the boundary at one time, then satisfying the conditions 6.9 ensures that the boundary conditions can be met at all places on the boundary for all time. We will therefore drop the exponential phase factors and concentrate on matching the boundary conditions.

Matching the Fields—Fresnel Relations

We have established Snell's law but would like to find the reflection and transmissions amplitudes. We will use the boundary conditions on $\boldsymbol{E}$ and $\boldsymbol{H}$ to find the relations between the amplitude vectors $\boldsymbol{E}_i, \boldsymbol{E}_r$ and $\boldsymbol{E}_t$ in 6.3 for the case that there is no sheet of free current flowing at the interface. The components of $\boldsymbol{H}$ as well as $\boldsymbol{E}$ parallel to the interface must match on either side (see 4.58 to 4.61). In dielectrics this is always the case as there are no free currents. However it is also true for all conductors other than an ideal conductor. A

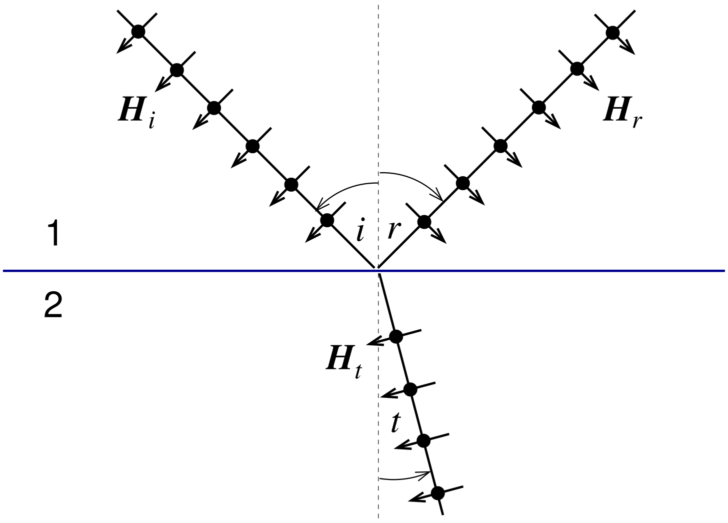

Figure 6.3: The figure defines the field directions for what is called the s-component when the electric field vector is perpendicular to the plane of incidence. Here the dots represent the electric field vectors pointing out of the page for each of the incident, reflected and transmitted waves. The directions of the magnetic field strength vectors are then defined by the direction of $\boldsymbol{k}$ and $\boldsymbol{E}$ in each case.

current *sheet* would have to flow through zero cross-sectional area which would have infinite resistance in any material with non-zero resistivity. It follows that

$$\hat{\boldsymbol{n}} \times (\boldsymbol{H}_2 - \boldsymbol{H}_1) = \boldsymbol{0}, \tag{6.10}$$

$$\hat{\boldsymbol{n}} \times (\boldsymbol{E}_2 - \boldsymbol{E}_1) = \boldsymbol{0}. \tag{6.11}$$

These are 4 independent conditions, matching the 4 degrees of freedom we have for the two polarisation components of each of the reflected and transmitted waves. We will treat the two polarisation components separately, taking our directions with reference to the direction of the incident wave and the boundary between the media.

Fig. 6.3 defines the field directions in the case the electric field is perpendicular to the plane of incidence, which is the plane containing the incident, reflected and transmitted wave vectors. This is known as the s-component from *senkrecht*, the German for perpendicular. The field $\boldsymbol{E}$ is (arbitrarily) defined to point out of the page for each of the three waves. We can assume this to be the case for the incident wave, which is ours to define, and will assume that it is true for the other two waves. If we can satisfy the boundary conditions with this assumption it will be correct (there is only one solution to a well-posed problem). We will find that it does work!

The directions of $\boldsymbol{E}$ and $\boldsymbol{k}$ for each wave automatically define the direction of $\boldsymbol{H}$ through the relation 5.8

$$\boldsymbol{k} \times \boldsymbol{E} = \omega \boldsymbol{B} = \mu_o \mu_r \omega \boldsymbol{H}. \tag{6.12}$$

The corresponding directions for $\boldsymbol{H}$ are indicated in Fig. 6.3. Matching parallel components of $\boldsymbol{E}$ and $\boldsymbol{H}$ on the boundary gives

$$E_i + E_r = E_t, \tag{6.13}$$

$$H_i \cos i - H_r \cos r = H_t \cos t. \tag{6.14}$$

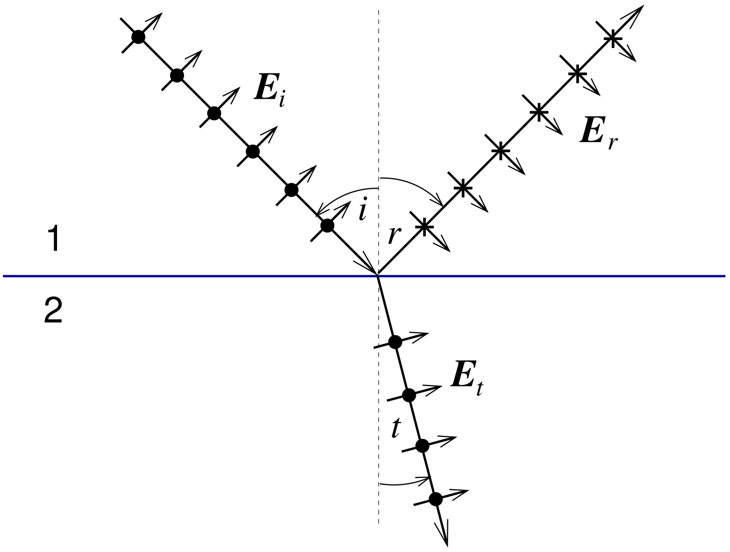

Figure 6.4: The field directions for the p-component when the electric field vector is in the plane of incidence. Now the electric field directions are chosen so that in the case of normal incidence, $i = r = t = 0$, all electric vectors are in the same direction. As before once the $\boldsymbol{E}$-field vectors are defined, then the $\boldsymbol{H}$-vector directions are defined. They point out of the page on the incident and transmitted rays (dots – "arrows" coming towards you) and inwards on the reflected ray (crosses – "arrows" going away from you).

Using the impedance $Z = E/H$ and $r = i$, we can solve these equations for

$$r_S = \frac{E_r}{E_i} = \frac{Z_2 \cos i - Z_1 \cos t}{Z_1 \cos t + Z_2 \cos i}, \quad \text{and} \quad t_S = \frac{E_t}{E_i} = \frac{2Z_2 \cos i}{Z_1 \cos t + Z_2 \cos i}. \tag{6.15}$$

With $\mu_r = 1$ (valid for most dielectrics), $Z = Z_0/n$. We can then write:

$$r_S = \frac{E_r}{E_i} = \frac{n_1 \cos i - n_2 \cos t}{n_1 \cos i + n_2 \cos t}, \quad \text{and} \quad t_S = \frac{E_t}{E_i} = \frac{2n_1 \cos i}{n_1 \cos i + n_2 \cos t}. \tag{6.16}$$

These are Fresnel's relations for the case where the electric field of the incident ray is perpendicular to the plane of incidence (and in this case, parallel to the interface). If the μ_r were significantly different from 1, we would use the results 6.15.

The reflectance, the fraction of incident energy reflected at the boundary, $(E_r/E_i)^2$, is given by

$$R_S = \left(\frac{n_1 \cos i - n_2 \cos t}{n_1 \cos i + n_2 \cos t} \right)^2. \tag{6.17}$$

Conservation of energy implies that a fraction $T_S = 1 - R_S$ is transmitted. Note that $T_S \neq Z_1 |t_s|^2 / Z_2$ as the energy flux varies as $EH = E^2/Z$ (see 5.20). We will look at the implications of these results after we have looked at the p-component of polarisation in which the electric field lies parallel to the plane of incidence.

For the p-component of polarisation, we define the field directions in Figure 6.4. Applying the boundary conditions (6.11) we see that

$$H_i - H_r = H_t, \tag{6.18}$$

$$E_i \cos i + E_r \cos r = E_t \cos t, \tag{6.19}$$

and hence that (taking $\mu_r = 1$)

$$\begin{aligned} n_1 E_i - n_1 E_r &= n_2 E_t, & (6.20) \\ E_i \cos i + E_r \cos r &= E_t \cos t. & (6.21) \end{aligned}$$

These have solutions

$$r_P = \frac{E_r}{E_i} = \frac{n_1 \cos t - n_2 \cos i}{n_1 \cos t + n_2 \cos i}, \quad \text{and} \quad t_P = \frac{E_t}{E_i} = \frac{2n_1 \cos i}{n_1 \cos t + n_2 \cos i}. \tag{6.22}$$

There are alternative conventions for the positive directions of the fields. A common one is to take the positive sense of the $\boldsymbol{H}$'s pointing in the same direction so $H_i + H_r = H_t$ (then there would be a sign change on E_r). One has to choose the sense of the field directions and remain consistent.

Consequences of the Fresnel relations, $n_2 > n_1$

The case, $n_2 > n_1$, applies to air-to-glass and air-to-water reflections. At normal incidence $i = r = t = 0$, $\cos i = \cos t = 1$ and the reflection coefficients become

$$r_S = r_P = \frac{n_1 - n_2}{n_1 + n_2}, \tag{6.23}$$

with reflectance

$$R_S = R_P = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2. \tag{6.24}$$

You should have encountered these relations in Foundations in the first year. For air ($n_1 = 1$) to (typical) glass ($n_2 = 1.5$), the reflectance is 4%; for air to water ($n_2 = 1.33$) it is 2%. Looking through a window, the two sides cause $\sim 8\%$ of the light to be reflected.

The amplitude coefficients r_S and r_P are negative. This means there is a phase shift of π when light from one medium reflects off another of higher refractive index. Although the human eye is not sensitive the the phase of optical light, such phase shifts are significant if there is any interference.

As the angle of incidence i increases towards 90°, $\cos i \to 0$, but the angle of transmission t does not reach 90°, and $\cos t > 0$. Thus for $i = 90°$ one finds that (see 6.16)

$$\begin{aligned} r_S(i = 90°) &= -1, & (6.25) \\ r_P(i = 90°) &= +1. & (6.26) \end{aligned}$$

The reflectance is 100% at grazing incidence. This is the reason for the bright reflections of the sunlight when the Sun is seen low over the sea or a wet road surface.

It is of interest that r_P become positive at large angles of incidence. Since $r_P < 0$ for $i = 0°$, it has to pass through zero at some intermediate angle of incidence. From the expression for r_P this happens when

$$n_1 \cos t = n_2 \cos i. \tag{6.27}$$

Squaring both sides and using $\cos^2\theta + \sin^2\theta = 1$ and Snell's law to replace $\sin t$ gives

$$n_1^2 \left(1 - \frac{n_1^2}{n_2^2} \sin^2 i \right) = n_2^2 \left(1 - \sin^2 i \right). \tag{6.28}$$

Multiplying through by n_2^2 and collecting similar terms

$$\left(n_2^4 - n_1^4\right) \sin^2 i = n_2^2 \left(n_2^2 - n_1^2\right). \tag{6.29}$$

This gives

$$\sin^2 i = \frac{n_2^2}{n_1^2 + n_2^2}, \tag{6.30}$$

or, after a little manipulation,

$$\boxed{\tan i_B = \frac{n_2}{n_1}.} \tag{6.31}$$

The angle i_B is known as *Brewster's angle*. For air/glass (1/1.5) reflections $i_B =$ 56.3°, while air/water (1/1.33) gives $i_B =$ 53.1°. At Brewster's angle, r_P goes to zero while r_S remains non-zero, thus the polarisation component with the electric vector in the plane of incidence is eliminated and the resultant reflected light is 100% linearly polarised with its electric vector perpendicular to the plane of incidence. This effect is the basis for wearing polarised sunglasses. Most of the reflected light off roads or water is polarised with $\boldsymbol{E}$ perpendicular to the plane of incidence (because of the Brewster angle effect). This can then be eliminated by using polarisers which block this polarisation. As most reflecting surfaces are horizontal or near horizontal, such polarised glasses are effective at reducing glare, and for instance can be useful when trying to look into water from the outside.

The P-component is perfectly transmitted at Brewster's angle. This is used in "Brewster windows" of gas lasers to avoid absorbing the light which could prevent lasing altogether if too much is lost. A "pile of plates" (e.g. a stack of microscope slides) can also act as a polariser if set at the right angle with the multiple interfaces removing the S-component. For air-glass about 26% of the S-component is removed per plate, and after ten plates only $(1-0.26)^{10} = 0.049$ or $\approx$ 5% remains compared to 100% of the P-component.

Consequences of the Fresnel relations, $n_2 < n_1$
Reflection from interfaces in which the medium of incidence has a higher refractive index than the transmitting medium (e.g. light reflecting from a glass/air or water/air interface), features the well-known effect of *total internal reflection*, which is useful for lossless reflections in optical devices such as binoculars. The elementary way to understand this is to consider Snell's law

$$n_1 \sin i = n_2 \sin t. \tag{6.32}$$

As the angle of incidence is increased from 0° and $\sin i \to 1$, there comes a point beyond which $\sin t > 1$, i.e. when $\sin i > n_2/n_1$. This isn't physical so there is no transmitted light. For glass/air, this happens for $i >$ 41.8°, which isar $<$ 45° making right-angled 45° prisms good reflectors. In fact, there is more to total internal reflection than this explanation might suggest and interestingly it turns out to be similar to quantum tunnelling is significant respects.

It should seem strange to you that there can be *no* transmitted light. Can we really match the boundary conditions on both $\boldsymbol{E}$ and $\boldsymbol{H}$ using the reflected wave only, particularly since total internal reflection occurs over a large range of incident angles? In fact *there is* a transmitted wave but it is *evanescent*, with an amplitude that decays exponentially with distance from the interface on the transmission side of the interface. In the steady state there is no overall flow of energy across the interface. To see this, we return to the origin of Snell's law in terms of the preservation of the component of the wave vector parallel to the interface:

$$k_{\|} = k_i \sin i = k_t \sin t. \tag{6.33}$$

From Pythagoras's theorem we must have

$$k_t^2 = k_{t,\|}^2 + k_{t,\perp}^2 = k_i^2 \sin^2 i + k_{t,\perp}^2. \tag{6.34}$$

Setting $k = (\omega/c)n = (2\pi/\lambda_0)n$ where λ_0 is the wavelength of the light in a vacuum, we find for $n_1 \sin i > n_2$

$$k_{t,\perp} = \pm i \left(\frac{2\pi}{\lambda_0}\right) \sqrt{n_1^2 \sin^2 i - n_2^2}. \tag{6.35}$$

The argument of the square root is positive.

The imaginary wave vector, $k_{t,\perp}$ in 6.35, occurs in the factor $\exp i(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)$. Measuring positions in the plane of incidence by x along the interface and y perpendicular to the interface, the spatial phase factor becomes

$$\exp i \left(k_{t,\|} x + k_{t,\perp} y\right) = \exp\left[\pm \left(\frac{2\pi}{\lambda_0}\right) \left(n_1^2 \sin^2 i - n_2^2\right)^{1/2} y\right] \exp\left[i \left(\frac{2\pi}{\lambda_0}\right) n_1 \sin(i)\, x\right]. \tag{6.36}$$

Physically, the nature of the solution has changed dramatically. Rejecting the solution that grows exponentially with y, we now have a rapid decay of the field with distance y away from the interface. The change of phase is in the x-direction, parallel to the interface only. In the y-direction, the field decays by a factor $1/\mathrm{e}$ over a distance

$$\frac{\lambda_0}{2\pi\sqrt{n_1^2 \sin^2 i - n_2^2}}. \tag{6.37}$$

For any value of incidence angle much above the critical angle at which total internal reflection starts to occur, this is less than one vacuum wavelength. For glass/air with $i = 45°$, the field decays by $1/\mathrm{e}$ in about $0.45\lambda_0$, around 0.2 μm for green light, on the air side of the interface.

The existence of the evanescent field is demonstrated in a phenomenon known as *frustrated internal reflection* when another prism is brought up to one in which light is undergoing internal reflection. Once close enough, partial transmission of the light into the second prism starts to take place. "Close enough" is less than a wavelength of light typically. The mathematics of this, with travelling wave solutions in the glass on either side of the air gap and exponential decay within the air gap, is similar to the phenomenon of quantum tunnelling whereby a particle can penetrate the classically "forbidden" part where its kinetic energy is negative. In the optical case, the air gap plays the role of the forbidden region in the tunnelling case.

6.2 Electromagnetic waves in conductors

A standard result of electrostatics is that the electric field in a conductor is zero. If it weren't, charges would move until the field decayed to zero. This idea lives on in the case of electrodynamics in the form of the "perfect conductor" approximation in which, even in the presence of time varying fields, one asserts that the electric field in the conductor is zero. This immediately implies a 100% reflectivity because the electric field of any incident wave must be completely cancelled by the reflected wave at the surface of a conductor to ensure zero field inside the conductor. This is actually a good approximation in some cases—radio waves on copper for example. The magnetic field even in a perfect conductor is not zero and currents are generated at the surface of the conductor. Real conductors are not perfect and we should move beyond the "perfect conductor" approximation consider the nature of the EM waves in conductors more closely.

6.2.1 The skin effect

We start from Maxwell's equations in matter (Eqs 4.38 to 4.41) now including a current density, $\boldsymbol{J}_f$, but still ignoring any free charge density as such densities quickly decay to zero:

$$\nabla\cdot\boldsymbol{D}=0, \qquad \nabla\cdot\boldsymbol{B}=0, \qquad \nabla\times\boldsymbol{E}=-\frac{\partial\boldsymbol{B}}{\partial t}, \qquad \nabla\times\boldsymbol{H}=\boldsymbol{J}_f+\frac{\partial\boldsymbol{D}}{\partial t}. \tag{6.38}$$

Assuming isotropic media ($\boldsymbol{B}=\mu\boldsymbol{H}$, $\boldsymbol{D}=\epsilon\boldsymbol{E}$, $\mu=\mu_0\mu_r$, $\epsilon=\epsilon_0\epsilon_r$), and we set

$$\boldsymbol{J}_f=g\boldsymbol{E}, \tag{6.39}$$

where g is the conductivity[1]. Equation 6.39 is equivalent to Ohm's law. For a linear isotropic medium, g is a scalar ($\boldsymbol{J}_f \parallel \boldsymbol{E}$) and is independent of $\boldsymbol{E}$.

Taking the curl of the third equation in 6.38 and using the relation 1.17, we find

$$\begin{aligned}\nabla^2\boldsymbol{E}&=\frac{\partial(\nabla\times\boldsymbol{B})}{\partial t}=\mu g\frac{\partial\boldsymbol{E}}{\partial t}+\mu\frac{\partial^2\boldsymbol{D}}{\partial t^2}\\&=\mu g\frac{\partial\boldsymbol{E}}{\partial t}+\mu\epsilon\frac{\partial^2\boldsymbol{E}}{\partial t^2}.\end{aligned} \tag{6.40}$$

This is the generalisation of 3.8 to the case of conductors (but in the absence of free charge density). When we look for plane wave solutions ($\boldsymbol{E}=\boldsymbol{E}_0\mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \boldsymbol{H}=\boldsymbol{H}_0\mathrm{e}^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}$, see 5.5) we make the substitutions $\nabla\rightarrow i\boldsymbol{k}$ and $\partial_t\rightarrow -i\omega$ and find

$$k^2\boldsymbol{E}=i\mu g\omega\boldsymbol{E}+\mu\epsilon\omega^2\boldsymbol{E}. \tag{6.41}$$

The dipersion relation is then

$$k^2=\mu\epsilon\omega^2+i\mu g\omega, \tag{6.42}$$

meaning that k is complex. The plane wave substitution into Maxwell's equations gives

$$\boldsymbol{k}\times\boldsymbol{E}=\mu\omega\boldsymbol{H} \tag{6.43}$$

which shows that $\boldsymbol{E},\boldsymbol{H}$ and $\boldsymbol{k}$ are mutually orthogonal in metals as in dielectrics. It also means that the impedance, which is the ratio E/H (see 5.12), becomes complex (it involves the ratio ω/k and k is now complex). A complex impedance, $Z=|Z|e^{i\phi}$ means that that $\boldsymbol{H}$ is phase-shifted by $-\phi$ with respect to $\boldsymbol{E}$ (remember that $H=E/Z$).

[1] Often σ is used for the conductivity, but g is a common fall-back which avoids confusion with σ denoting surface charge density

The "good conductor" approximation

One could struggle on and work with the square root of the right-hand side of Eq. 6.42 to find an explicit form for k. The algebra is greatly simplified, while capturing the essential physics, if we make the "good conductor" or "metallic" approximation. This is a step up from the perfect conductor $\boldsymbol{E} = \mathbf{0}$ approximation.

A good conductor has large conductivity g and, in the case where $g \gg \epsilon\omega$, we can neglect the first term in Eq. 6.42 and write

$$k^2 \approx i\mu g\omega, \tag{6.44}$$

or

$$k = \pm(1+i)\sqrt{\frac{\mu g\omega}{2}} = \pm\frac{1+i}{\delta}. \tag{6.45}$$

Here

$$\boxed{\delta = \sqrt{\frac{2}{\mu g\omega}},} \tag{6.46}$$

is called the *skin depth*.

For a wave propagating in the $+x$ direction, the spatial and time dependence of a wave in a metal is

$$\mathrm{e}^{i(kx-\omega t)} = \mathrm{e}^{-x/\delta}\mathrm{e}^{i(x/\delta-\omega t)}. \tag{6.47}$$

This is a travelling, exponentially-damped, wave. Its amplitude drops by 1/e in distance δ, hence the term "skin depth". Since $1/\delta$ is also effectively the wavenumber in the travelling wave part, the wavelength of the wave in a conductor is given by

$$\lambda = 2\pi\delta. \tag{6.48}$$

This can be much smaller than the vacuum wavelength, as an example will show. For good conductors, the skin depth can be small at moderate frequencies. In the case of $g \to \infty$, then $\delta \to 0$ and we recover the result for the perfect conductor. The currents flowing in this case take the form of a current sheet. Although there is no such thing as an infinitely thin sheet of current, it can get close to this ideal in superconductors.

Example 6.1. Calculate the skin depth of copper ($g = 5.96 \times 10^7\,\Omega^{-1}\,\mathrm{m}^{-1}$) for EM waves of frequency $f = 2\,\mathrm{MHz}$ (radio).

Show that the neglect of $\mu\epsilon\omega^2$ compared to $\mu\sigma\omega$ is justified.

Answer. $\omega = 2\pi f = 4\pi \times 10^6\,\mathrm{rad\,s^{-1}}$. Copper is weakly (dia-)magnetic, $\mu \approx \mu_0 \approx 4\pi \times 10^{-7}$, and

$$\delta = \sqrt{\frac{2}{\mu\sigma\omega}} = \sqrt{\frac{2}{4\pi \times 10^{-7} \times 5.96 \times 10^7 \times 4\pi \times 10^6}} = 4.61 \times 10^{-5}\,\mathrm{m}. \tag{6.49}$$

The ratio of the magnitudes of the two terms on the rhs of 6.42 is

$$\frac{\mu\epsilon\omega^2}{\mu\sigma\omega} = \frac{\epsilon\omega}{\sigma} = \epsilon_r\frac{8.85 \times 10^{-12} \times 4\pi \times 10^6}{5.96 \times 10^7} = 1.87 \times 10^{-12}\epsilon_r. \tag{6.50}$$

This is much less than one for normal values of the relative permittivity ϵ_r. The vacuum wavelength for these waves is 150 m, and the wavelength is dramatically shorter in the conductor. The wave speed is correspondingly slow, a sluggish 13 830 m s^{-1} in this case.

Submarine communication

The answer of the example shows that high frequency waves rapidly damp in conductors. Even in sea water, with a conductivity about 10^7 times lower than that of copper, for the same frequency of 2 MHz, the skin depth is only 15 cm. This rules out normal radio communications in water, and one requires extremely low frequencies and high powers to be able to communicate with submarines when they are at their operating depths. The submarines themselves can only receive and not transmit. Operating frequencies of $\sim$ 80 Hz are discussed, requiring electrodes driven into the ground tens of miles apart and entire dedicated power plants to radiate a few Watts of power (radiation at such a low frequency is inefficient). I am unclear whether this is still employed, or whether there is now a reliance on alternatives such as sonar beacons spread around the oceans or carrier pigeons.

High frequency currents

The skin depth is important when one tries to pass high frequency currents through conductors. The current is confined to a thin skin on the surface of the conductor. In the example above, only an $\approx$ 50 μm layer at the surface of a copper wire carries a 2 MHz signal. At high frequencies, it makes sense to use flattened wires to avoid unused central parts. Careful use of multi-stranded wires can help as well. At high frequencies (microwaves), wires don't work well at all, and one uses wave-guides (metal pipes of usually rectangular cross-section) with thin interior layers of good conductors to minimise losses.

It is often said that the skin effect means that high frequency voltage sources are not as dangerous to humans as direct current because the current is confined to the skin, avoiding the heart. An oft-qoted example is the Tesla coil. It would be unwise to put too much faith in this however as the conductivity of the human body is not homogeneous.

The skin depth can apply in other situations which may not appear to be examples of EM waves. Consider the following:

Example 6.2. A 10 cm wide block of copper immersed in Earth's magnetic field is rotated. Estimate the spin frequency at which significant exclusion of Earth's field occurs.

Answer. From the perspective of the copper block, the magnetic field at the surface switches back and forth at the spin frequency, as it would in the case of an EM wave. We want to find the frequency at which the skin depth becomes comparable to the size of the block.

Working from the previous example (2 MHz gave $\delta = 4.61 \times 10^{-5}$ m), we obtain

$$f \sim 2 \times 10^6 \times \left(\frac{4.6 \times 10^{-5}}{0.1}\right)^2 = 0.4\,\text{Hz}, \tag{6.51}$$

a surprisingly low frequency. If spun at high speed, the Earth's field would be screened by surface currents and almost completely excluded from the interior of the block.

Induction cookers

The skin depth $\sqrt{2/\mu\sigma\omega}$ is particularly small in ferromagnetic materials which have relative permeabilities in the thousands. This is used in induction cookers to concentrate currents into a thin layer at the bottom of cooking vessels, increasing their effective resistance. This is why these cookers do not work with copper or aluminium saucepans. On the other hand, steel is never used for power cables not only because it is has a lower conductivity than copper or aluminium, but because its high magnetic permeability greatly reduces the skin depth and leading to a much higher resistance per unit length of wire, even at mains frequencies.

6.2.2 Reflections from conductors

We will consider an EM wave normally incident upon a conductor[2]. The impedance of conductors $Z = E/H$ (with $\boldsymbol{E}$ and $\boldsymbol{k}$ perpendicular to each other) follows from Eq. 6.43:

$$Z_C = \frac{E}{H} = \frac{\mu\omega}{k} = \frac{\mu\omega\delta}{1+i} = \frac{\mu\omega\delta}{2}(1-i) \tag{6.52}$$

The relation between E and H in a conductor is complex, meaning that there is a phase shift between them of $\pi/4$ radians or $45°$.

For the example of the previous section, $\mu\omega\delta \approx 7.3 \times 10^{-4}\,\Omega$, much less than the impedance of free space $Z_0 = 377\,\Omega$. We will see that this large impedance mismatch between free space (and air) and good conductors means that there is near-100% reflectance with a $\approx \pi$ phase-shift of light incident on a conductor. The electric fields on the free space side of the boundary nearly cancel while the magnetic fields $\boldsymbol{H}$ combine allowing them to match their values in the metal.

At normal incidence, the parallel components of E and H match at the boundary just as with dielectrics (no current sheets, although there are volume currents). We can use 6.14 with $\cos i = \cos r = 1$. The formula for the reflection coefficient in terms of the impedances (6.15) is the same as for the dielectric case:

$$r = \frac{Z_C - Z_0}{Z_C + Z_0}. \tag{6.53}$$

Dividing through by Z_0 gives

$$r = -\frac{1 - Z_C/Z_0}{1 + Z_C/Z_0} \approx -\left(1 - 2\frac{Z_C}{Z_0}\right). \tag{6.54}$$

$|Z_C/Z_0| \ll 1$. The result is a strong reflection with a phase shift close to π, similar to what we asserted was the case for perfect conductors at the start of the chapter. With the explicit formula for a conductor one could calculate the true reflectivity and phase shift which would not be precisely 100% and π respectively.

[2] The case of oblique reflection from conductors is complicated – even the rather ferocious classic EM textbook "Jackson" steers clear of it.

Chapter 7

Geometric Optics

Light is an EM wave, so optics is a sub-branch of electromagnetism. The full Maxwell equations are hard to solve in situations of any complexity, and, where necessary, tend to be solved using computers. However, in many cases, problems in optics such as lenses can be dealt with through approximation. One level of approximation is (scalar) wave optics which stems from Huygens principle (refined by Kirchoff). This states that each point on a wavefront acts a source of secondary waves. What are called far-field or Fraunhofer and near-field or Fresnel diffraction are encompassed by this approach. Far-field diffraction is very much bound up with Fourier analysis and is covered in mathematics modules, and we will not cover it here. It is itself non-trivial to implement in many circumstances. Instead we will fall fall back on a still-more elementary approximation, that of "geometric optics". Although approximate, it is of immense practical importance as it forms the basis of the designs of optical instruments, such as telescopes, cameras, binoculars and microscopes.

7.1 Wavefronts and rays

When the wavelength of light is much smaller than the size of any object it impinges on, light of wavelength $\lambda = 500\,\mathrm{nm}$ compared to a lens of diameter 5 cm for example, we can usually ignore the wave nature of light. This is the regime of geometrical optics. There are three key rules of governing geometrical optics:

1. Light travels in a straight line in a homogeneous medium.
2. On reflection, the angle of incidence equals the angle of reflection.
3. On refraction, Snell's law applies $n_1 \sin i = n_2 \sin t$.

Wavefronts and rays are still important concepts in geometric optics. We will take a pictorial approach to these; for a more mathematical approach (non-examinable), see appendix I if only to meet a word you may never have heard before: the "eikonal". A first example is "light from infinity", beloved concept in all standard examples of diffraction. You could be thinking of light from a distant star for instance. This can be represented by plane wavefronts, with corresponding rays defined by the direction of travel perpendicular to the wavefronts, as

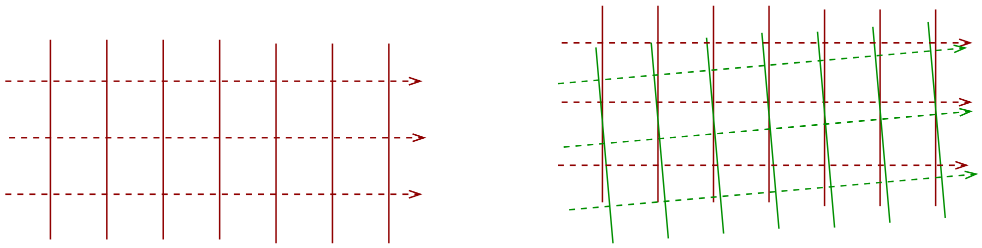

Figure 7.1: *Left:* By the time it reaches us, light from a distant star, off the left of the plot, consists of a series of plane wavefronts. The dashed arrows represent a corresponding set of rays. The light here comes from a single direction and when looked at by eye appears as a single point on the sky. *Right:* Two points sources separated by 5° on the sky give rise to two sets of wavefronts and rays as shown. One would see these as two points of light. Although it may appear that interference effects occur, this would be only be the case if the two sources had a fixed phase with respect to each other. More normally separate sources of light are "incoherent" and only the total fluxes from each need be added. Interference and diffraction are not part of geometrical optics.

illustrated in Fig. 7.1. Light from two distant stars separated by a small angle on the sky gives two sets of wavefronts and rays tilted with respect to each other by whatever the angular separation is (right of fig. 7.1). At finite distances, from point sources, the wavefronts are spherical (Fig. 7.2) with the corresponding light rays diverging from the source. Spherical waves can also converge on a point. From this point of view, a lens is a device that can convert an expanding set of wavefronts with a diverging set of rays into a contracting series of wavefronts and a converging set of rays. Fig. 7.3 illustrates this.

7.2 Fermat's principle of "least time"

In this section we are interested in the path that a ray takes. One way to approach this is to consider small segments, take them one step at a time and apply the laws of refraction and reflection to work out changes in angle, and then move onto the next step etc. A different (but equivalent) approach is contained in "Fermat's principle of least time" which says that the path taken by a light ray between any two points, A and B, is the one that takes the least time (but see below because this is in fact not quite right). This seems a little odd at first: how does a light ray "know" that the path it is taking is the minimum? Fig. 7.4 shows a

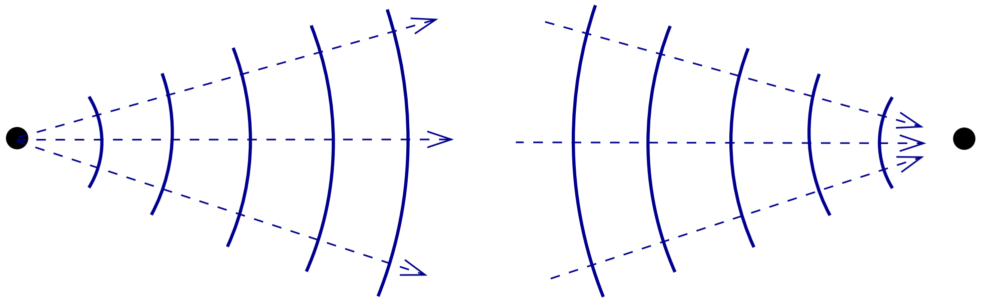

Figure 7.2: Spherical waves can expand away from a point (left) or converge towards a point (right).

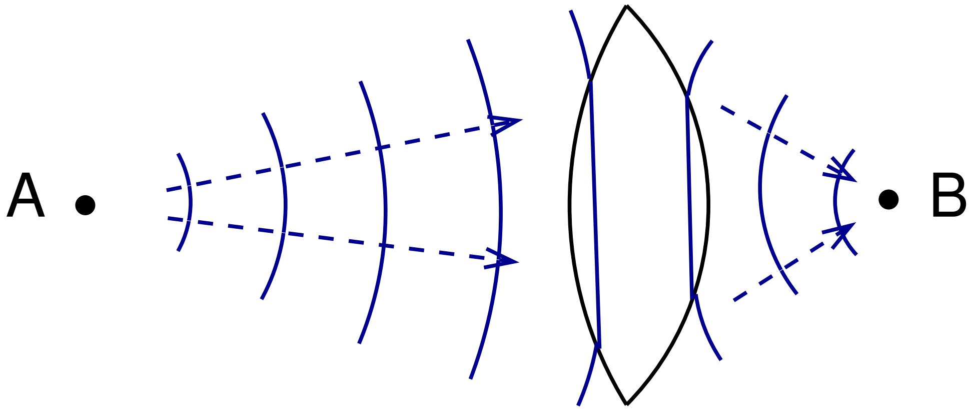

Figure 7.3: Expanding spherical waves from a point source A on the left are converted into a contracting set on the right which converge on point B. B is the image of A. Notice that if you follow any ray marking the path taken by the light from A to B, it cuts through the same number of wavefronts.

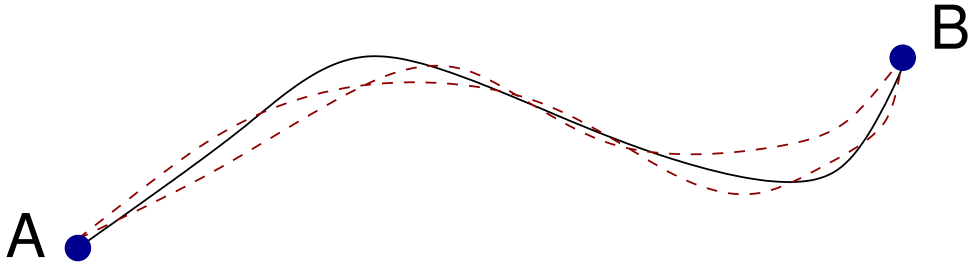

Figure 7.4: Light in some medium travels from point A to B along the solid black line. The two dashed lines show two of an infinite number of alternative nearby paths. Because of the Huygens-Fresnel principle, each of these is traversed by the light, but the one that wins out is the path for which any small perturbation makes no difference to the overall phase change between A and B because then neighbouring paths re-enforce each other rather than cancelling.

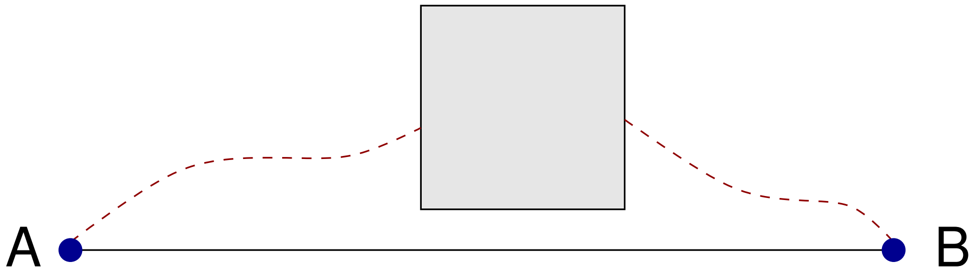

Figure 7.5: Light in a homogeneous medium travels in a straight line from A to B. However, the presence of a solid object clearly blocks some alternative paths, if light truly "explores" all possible paths between A and B. Does this really affect the path of the light as the explanation of Fermat's principle seems to imply?

visualisation to motivate Fermat's principle. By the Huygens-Fresnel principle each wavefront acts as a source of secondary waves that build into subsequent wavefronts. The secondary waves spread in all directions, thus in travelling from A to B, light effectively traverses an infinity of possible paths. The one that wins, and defines what we call "the light ray" is an extremum in terms of overall phase. That is, its phase does not vary to first order under any small perturbation of the path. This works because then there will be a finite sets of neighbouring paths from A to B which all match in phase (to within some tolerance, say a radian) and thus enforce each other. Paths that don't have this property cancel out, and we are not even aware of their existence in geometrical optics.

If all of this is true, consider the case shown in Fig. 7.5? If the idea of alternative nearby paths is correct, blocking some of them as illustrated must have some effect. Given that the object is nowhere near the path of the ray shown, this doesn't seem plausible. In fact there *is* indeed an effect in such cirumstances, it's just that in many cases it would be so small as not to be noticeable. This is the subject of diffraction theory, in particular what is known as "Fresnel diffraction". Within the approximations of Fresnel-Kirchoff diffraction theory, one can calculate such effects. They can also be seen directly, for instance during "lunar occulations". As the Moon occults a star, the light flux from the star varies, becoming successively brighter and then fainter, even before the Moon is directly in the line of sight between the star and the observer.

The change in phase traversing an arbitrary path from A to B can be written

$$\Delta\phi = \int_A^B \boldsymbol{k} \cdot d\boldsymbol{r} = k_0 \int_A^B n \, d\ell, \tag{7.1}$$

with $d\ell$ the length of a line element along the path of a ray and n the refractive index which applies to that element (potentially a function of position). The integral is known as the "optical path length" or "optical path", "O.P." for short. I will sometimes also denote it by τ. That is

$$\text{O.P.} = \tau = \int_A^B n \, d\ell. \tag{7.2}$$

If the path runs through a series of homogeneous media then we may also write

$$\text{O.P.} = \tau = \sum_i n_i \Delta\ell_i, \tag{7.3}$$

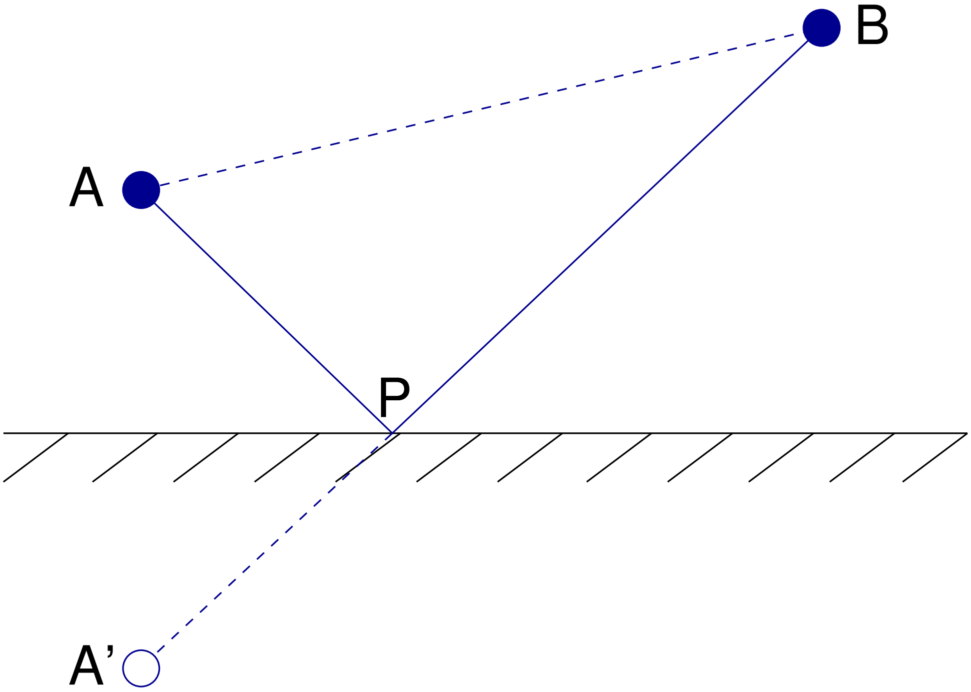

Figure 7.6: Reflection in a mirror. The image of A at A' is located "inside" the mirror, an equal distance from the reflecting surface. This ensures that the line from A' to P to B is straight and therefore an extremum in terms of light path, as therefore is the path from A to P to B. However it is not *the* global minimum as the direct path from A to B is clearly shorter. It is a simple matter of geometry to show that $i = r$ from this figure.

where $\Delta\ell_i$ is the length of a straight segment of path in medium i which has refractive index n_i. This integral is directly related to the time taken (ignoring complications of the group velocity – it is always the wave speed that matters here) which is

$$\Delta t = \int_A^B \frac{d\ell}{v_\phi} = \frac{1}{c}\int_A^B n\, d\ell, \tag{7.4}$$

since the phase velocity $v_\phi = c/n$.

This explanation of Fermat's principle implies that the optical path is an extremum which means that it takes a stationary value with respect to perturbations of the path. It does not show that it needs be "least" and indeed, although very often the time taken is minimal, it is also possible to think of cases where it is maximal, so the expression "least time" is not correct. Thus a modern formulation would be "the time taken by light to travel between two points is an extremum".

Only small perturbations of the path should be considered, which means there can be, and often are, multiple optical paths between the same two points. An obvious case is reflection. Fig. 7.6 shows this for the case of reflection. The path with a reflection is a local minimum, but there is clearly an even shorter direct path, which represents the global minimum time path. Another instance is gravitational lensing in astronomy where multiple images of the same source are sometimes seen.

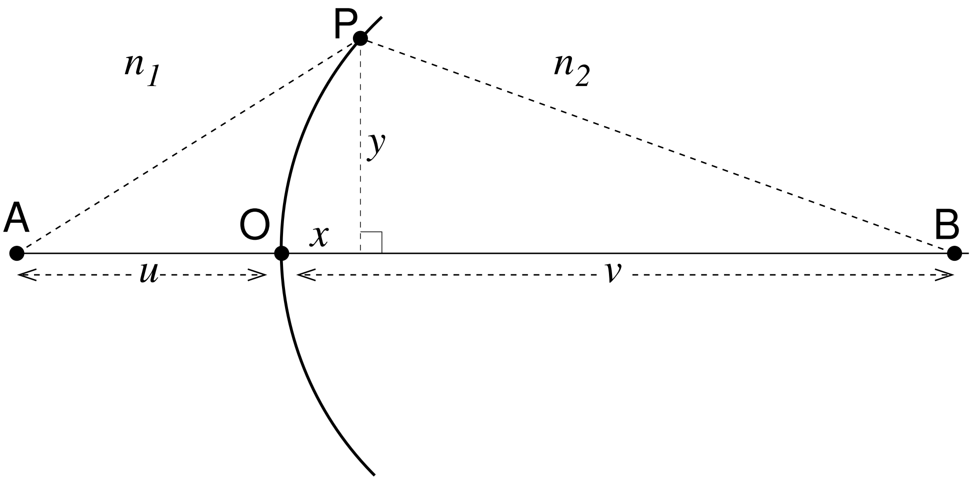

Figure 7.7: An object A is imaged at point B by a spherical interface between two dielectrics. Distances are measured from O, the point on the spherical surface that lies between A and B. The line AOB goes through the centre of the sphere.

7.3 Imaging by a spherical interface

An important aspect of optics is "imaging". We wish to to create an image of an object, perhaps to capture onto a detector, or to look at in real time by eye. This covers cameras, telescopes, binoculars and other optical instruments. An ideal imager takes rays from some point A on the object and focuses them onto a unique point B which becomes the image. By Fermat's principle (and ignoring multiple extrema) all of them should have the same optical path length. We will use the extremum property of the optical path to analyse imaging.

Consider then the situation shown in Fig. 7.7. Rays diverging from an object A are focused at the image B. Only two such rays are shown in the diagram, the *axial* ray, AOB, defining the axis of the system, and the ray APB. We want to find a condition that ensures that the optical path along APB is the same as along AOB to first order. This will ensure something like the invariance of the optical path with choice of ray that we need. This condition will relate the object and image distances, u and v to each other.

The optical path along APB consists of two straight-line segments (homogeneous media), of lengths AP and PB and refractive indices n_1 and n_2. Using Pythagoras' theorem and the definitions from the figure, we find

$$\begin{aligned}\tau &= n_1\left((u+x)^2+y^2\right)^{1/2}+n_2\left((v-x)^2+y^2\right)^{1/2}, &(7.5)\\ &= n_1\left(u^2+2ux+x^2+y^2\right)^{1/2}+n_2\left(v^2-2vx+x^2+y^2\right)^{1/2}. &(7.6)\end{aligned}$$

The point P is on the spherical surface so x and y are related. To obtain the relation between them we can use a theorem from geometry. This intersecting chords theorem states that, when any two chords on a circle cross, then the products of the lengths of the two pieces of each chord are equal (this was first proved by Euclid). Applying this to the chord defined by OB where it crosses the perpendicular dropped from P, which is the other chord, gives $y^2=(2R-x)x$, so $x^2+y^2=2Rx$. The optical path as a function of x is then:

$$\tau(x)=n_1\sqrt{u^2+2(R+u)x}+n_2\sqrt{v^2+2(R-v)x}. \tag{7.7}$$

Fermat's principle requires, $\tau(x)$ in 7.7, to be an extremum.

We require

$$\frac{d\tau(x)}{dx} = \frac{n_1(R+u)}{\sqrt{u^2+2(R+u)x}} + \frac{n_2(R-v)}{\sqrt{v^2+2(R-v)x}} = 0. \tag{7.8}$$

We will limit ourselves to "paraxial rays", that is rays that lie at small angles to the axis so that x is small. This allows us to drop the terms in x inside the square roots. This approximation ignores effects behind what is called "spherical aberration". The approximation is also referred to as "Gaussian optics". We arrive at the condition

$$\frac{n_1(R+u)}{u} + \frac{n_2(R-v)}{v} = 0. \tag{7.9}$$

which after division by the radius of curvature R, can be re-arranged to yield

$$\frac{n_1}{u} + \frac{n_2}{v} = \frac{n_2-n_1}{R}. \tag{7.10}$$

If the spherical surface curved the other way, i.e. the centre of the sphere defining the surface lay to the left rather the right, then the signs in the terms $u+x$ and $v-x$ would be swapped and the end result would be

$$\frac{n_1}{u} + \frac{n_2}{v} = -\frac{n_2-n_1}{R}. \tag{7.11}$$

This is equivalent to setting $1/R \to -1/R$ and we usually refer to this as a negative radius of curvature.

Real-is-positive vs Cartesian conventions

We defined the lengths u and v to be positive in Fig. 7.7, but the relation 7.11 implies that they could also be negative. For instance if u is small enough that

$$\frac{n_1}{u} \geq \frac{n_2-n_1}{R}, \qquad \Rightarrow \qquad \frac{n_2}{v} = \frac{n_2-n_1}{R} - \frac{n_1}{u} \leq 0. \tag{7.12}$$

Rather than converging onto point (B) to the right of the surface, the rays will appear to diverge from a point to the left of the surface. When B is to the right of the surface is called a "real" image – you could put a piece of paper there and see the image projected onto it. When B is to the left, an observer (on the right) sees rays which appear to come from the virtual image at B whereas they come from A but have been refracted. B is then called a "virtual" image. In the real image case $v > 0$, whereas $v < 0$ for the virtual image. The way we have worked is known as the "real-is-positive" convention. It is also possible to have virtual objects (A) if the rays do not diverge from a real point but instead converge on a point that lies to the right of O.

An alternative convention, which is sometimes preferable, is the Cartesian convention. This treats any distance to the left of O as negative and anything to the right positive. Indicating the Cartesian equivalents to u and v by dashes we get instead

$$-\frac{n_1}{u'} + \frac{n_2}{v'} = \frac{n_2-n_1}{R}. \tag{7.13}$$

The Cartesian convention can be easier to use in numerical calculations.

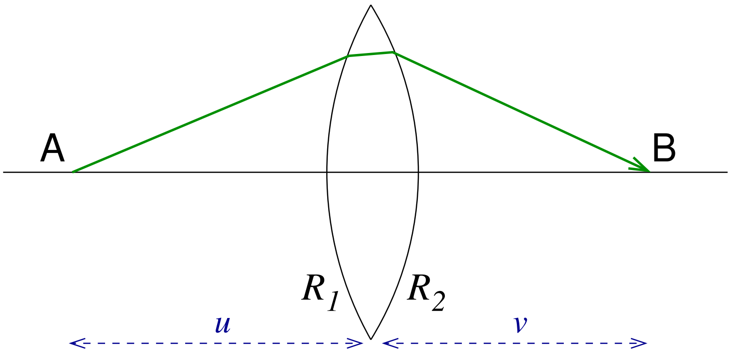

Figure 7.8: Imaging by a lens can be approximated by two successive instances of imaging by a spherical surface. At the first surface the ray goes from air ($n = 1$) to glass ($n = n$); at the second it goes from glass ($n = n$) to air. The radii of curvature can be different and the second surface is reversed in sense compared to the first hence the curvatures enter as $1/R_1$ and $-1/R_2$.

7.4 Thin lenses

An elementary lens can be thought of as two spherical surfaces, one after the other. The image formed by the first surface acts as the object for the next. So let's consider imaging by a biconvex lens in air (Fig. 7.8), with light travelling from air into the glass (refractive index n) at a first surface with radius of curvature $1/R_1$, and then from glass to air at the second surface, radius of curvature $-1/R_2$. Let the object distance be u (we are using the real-is-positive convention) and the image due to the first surface be at v' (not shown in the figure). This acts as the object for the second surface with object distance $u' = -v'$. This is because if $v' > 0$, then as far as the second interface goes, the "object" is virtual and must have negative u'. In writing $u' = -v'$, we have also neglected the offset of the reference point O between the two surfaces: this is the "thin lens" approximation.

Letting the image distance for the second interface be v, the the following two relations are obtained:

$$\frac{1}{u} + \frac{n}{v'} = \frac{n-1}{R_1}, \tag{7.14}$$

$$\frac{n}{-v'} + \frac{1}{v} = -\frac{1-n}{R_2}. \tag{7.15}$$

In Fig. 7.8 the image that would be created by the first spherical surface, with curvature R_1, would be virtual and to the left of the diagram. The 'intermediate' image distance v' would then be negative. We add the equations in 7.15 and the quantity v' drops out:

$$\frac{1}{u} + \frac{1}{v} = (n-1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right). \tag{7.16}$$

If we let $u \to \infty$, then light from infinity is brought together at what is called the focus of the lens at focal length f, and $v = f$. This equation may then be written as a pair of equations. First, one involving nothing but constants:

$$\boxed{\frac{1}{f} = (n-1)(\frac{1}{R_1} + \frac{1}{R_2}).} \tag{7.17}$$

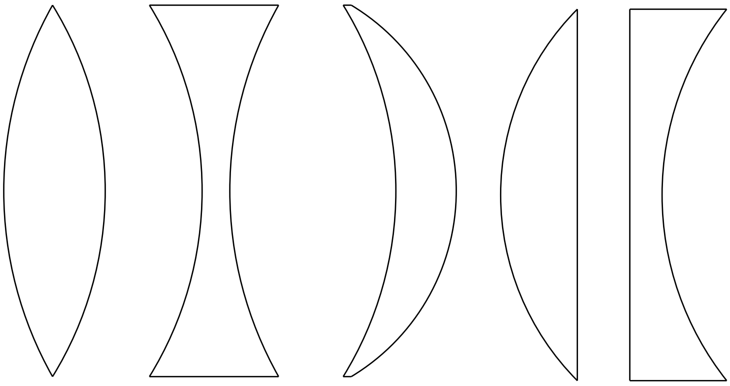

Figure 7.9: Different basic types of lenses. From left-to-right they are converging, diverging, converging, converging and diverging. The central one could also be made to diverge if the curvatures were changed. It is common for some of these to be combined, e.g. the first two can be joined with a transparent glue to cut down reflections. This is used in "achromats" to reduce the effects of chromatic aberration due to the variation of refractive index with wavelength.

This gives the focal length of a lens in terms of the radii of curvature of the two surfaces and is sometimes called the "lensmaker's equation". Once you have your lens with defined focal length, you will tend to use the second equation which is the "thin lens equation":

$$\boxed{\frac{1}{u} + \frac{1}{v} = \frac{1}{f}.} \tag{7.18}$$

This is in real-is-positive form. The Cartesian equivalent is

$$\boxed{-\frac{1}{u'} + \frac{1}{v'} = \frac{1}{f}.} \tag{7.19}$$

Different types of lens

For each surface of a lens one has a choice of positive, negative or zero curvature – see Fig. 7.9. One can have bi-convex lenses with both $1/R_1$, and $1/R_2$ positive in the lens-maker's equation. With $1/f > 0$, this yields a *converging* lens. The opposite is a bi-concave lens giving a *diverging* lens for which $1/f < 0$. The different types all have their uses in optical instruments, but for us it will only be $1/f$ that matters. Opticians use units of "dioptres" (inverse metres) to express the "power" of lens in prescriptions. Someone who is near-sighted would need a lens of negative power to allow them to see more distant objects. Reading glasses with a positive power of $+1.5$ dioptres, prescribed to a long-sighted person, would have a focal length of 0.66 m.

7.4.1 Principal rays for visualising lens imaging

There are three principal rays that allow one to draw diagrams illustrating imaging with lenses. These can provide useful intuition and it is worth drawing out a few such diagrams to get the hang of them. In all cases I assume the ray proceeds from left to right; solid dots mark the focal points which are distance f from the lens on the left and right. This is material best

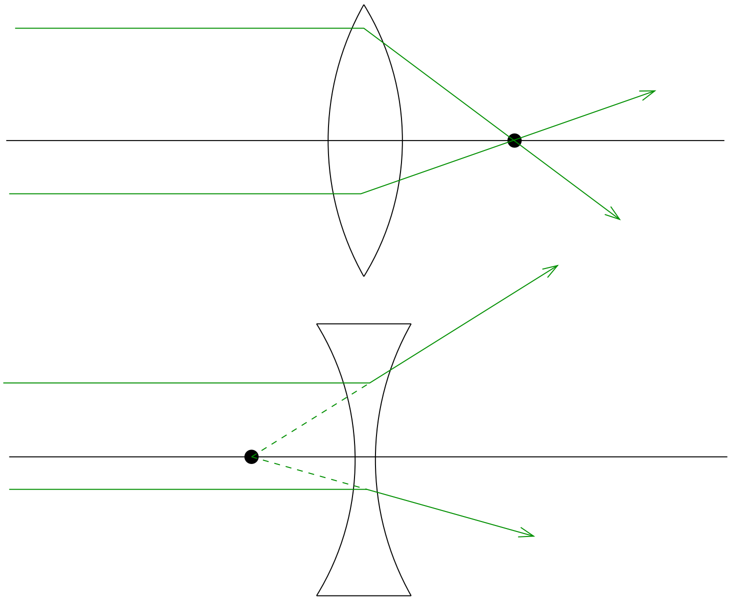

Figure 7.10: Rays parallel to the optic axis prior to hitting the lens are deflected through the right-hand focal point or appear to have come from the left-hand focal point once they have passed through the lens. Remember that any object on the left of the lens puts out rays in all directions so one can always choose to focus on a ray parallely to the optic axis as an example.

seen visually, so study the figures of this section carefully. Each type of ray is numbered 1, 2 or 3 as these will be used later when building composite diagrams.

Rays that start parallel to the optic axis (1)
Any ray that is parallel to the optic axis when left of the lens will either be refracted through the focal point on the right of the lens (converging) or appear to have come from the focal point on the left of the lens (diverging). This is illustrated in Fig. 7.10.

Rays through the centre of the lens (2)
For thin lenses, rays through the centre of a lens are undeviated, so these are the simplest of all – see Fig. 7.11.

Rays that finish parallel to the optic axis (3)
The counterpart to rays that start parallel to the axis are those that finish parallel. Fig. 7.12 shows examples of this.

7.4.2 Ray diagrams for lenses

We can now cover a few example of using these principal rays to work out imaging by a lens.

Cameras
We will start with a converging lens used to make a real image of a real object (Fig. 7.13). This is essentially the situation with a human eye and a camera. The image is projected onto some form of two-dimensional light detector (the retina, a CCD detector, or photographic film). Note that any two of the principal rays would do to establish the position of the image, but all three provides a useful check. The key features of this case are that the lens is converging

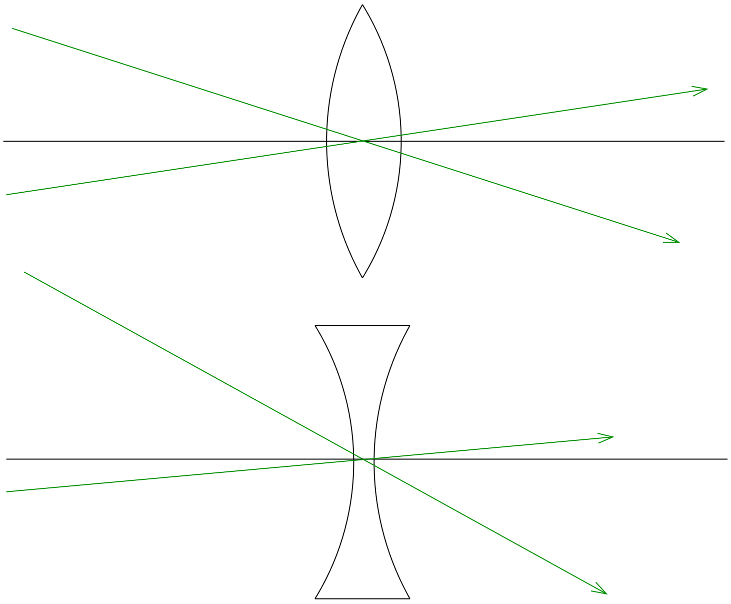

Figure 7.11: Rays that pass through the centre of a lens are undeviated.

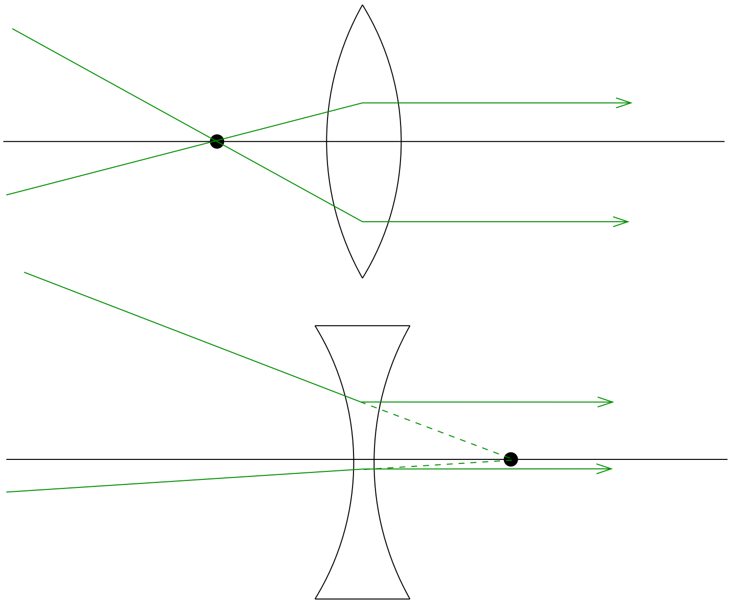

Figure 7.12: Rays parallel to the optic axis after passing through the lens are deflected from those that have come from the left-hand focal point (top) or would have passed through the right-hand focal point in the absence of the lens (bottom). Remember that any object on the left of the lens puts out rays in all directions so one can always choose to focus on rays parallel to the optic axis (either or before passing through the lens) as examples.

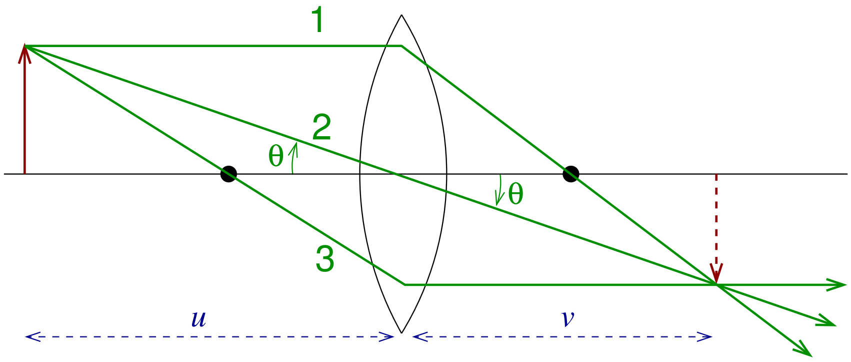

Figure 7.13: A diagram using each of the three principal rays discussed in section 7.4.1 to illustrate imaging. Here the lens forms a real image of an object on the left of the lens.

$(f > 0)$ and that the object is further than f from the lens, i.e. $u > f$. Hence

$$\frac{1}{v} = \frac{1}{f} - \frac{1}{u} > 0, \tag{7.20}$$

and the image is real. The diagram also shows that the image is inverted. Trees are "upside down" on your retina, which is no problem for your brain as it has no sense of a "correct" orientation[1]. The central ray is useful for working out the size of the image L_i relative to the object L_o because the central ray, the optic axis and the object and the central ray, the optic axis and the image form similar triangles from which one can immediately deduce that

$$\frac{L_i}{L_o} = \frac{v}{u}. \tag{7.21}$$

Example 7.1. The focal length of a human eye is $f \approx 22\,\mathrm{mm}$. Calculate the size of image cast on the retina by a 10 m high tree seen from a distance of 150 m.

Answer. Using the thin lens equation

$$v = \frac{1}{1/f - 1/u} = \frac{fu}{u-f} = \frac{0.022 \times 150}{150 - 0.022} = 0.022\,003\,\mathrm{m}. \tag{7.22}$$

This is only a marginally longer than f: a distance of 150 m is much longer than the size of an eye and is effectively almost "at infinity". Therefore the size of the tree on the retina is

$$L_i = \frac{v}{u} L_o = \frac{0.022003 \times 10}{150} = 1.47\,\mathrm{mm}. \tag{7.23}$$

The level of detail one could see in this case shows that the sensors in the retina ("rods" and "cones") must be significantly smaller than this quantity.

Magnifying glasses

A different example of imaging with a converging lens is shown in Fig. 7.14. The key feature

[1] Experiments have been carried out where people wear goggles with mirrors to invert what they see vertically. It takes of order 10 days for people to adapt to this. Makes one feel slightly ill just thinking about it.

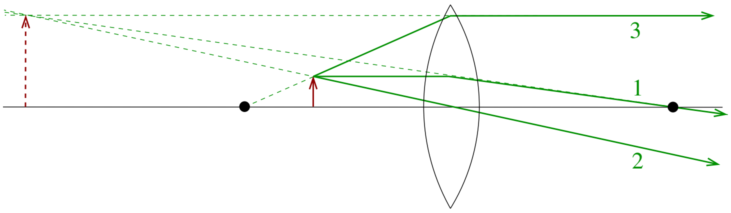

Figure 7.14: Ray diagram of a converging lens being used as to "magnify" an object.

here is that the object is *closer* to the lens that the left-hand focal point $u < f$, hence

$$\frac{1}{v} = \frac{1}{f} - \frac{1}{u} < 0, \tag{7.24}$$

and the image is virtual. A camera-like imaging system (e.g. our eyes) is needed to see or record this image. This is how one uses a magnifying glass.

The sense in which the setup in Figure 7.14 magnifies is a little confusing. You might think that it magnifies as the image is larger than the object. However, the image is also further away from the lens than the object. What matters when it comes to magnification is the *angle subtended* by a given feature in the object. A small feature of size ℓ in an object at distance d from the observer will subtend angle

$$\theta = \frac{\ell}{d} \tag{7.25}$$

radians, making the small angle approximation. The only way (without aid) to make θ larger to see an object is more detail is to make d the distance smaller, but our eyes cannot focus once an object is closer than the "near point", $d_{\rm np}$, which is about 25 cm for a normal-sighted person. The real point about a magnifying glass is that it allows you to reduce d while still managing to focus on the object so long as $|v|$ is larger than one's near point distance.

A slight complication in giving the exact gain is that the eye's power is variable, allowing one to focus (in the ideal case) on anything from infinity to the near point. To be specific, let's assume a "relaxed" eye, focussed on infinity. This is typically a desirable state when using any optical instrument rather than straining to focus on one's near point. In this case one wants $v \rightarrow -\infty$, i.e. one places the object at the focus of the magnifying lens and $d = f$. The magnification of interest is then the ratio of the angle subtended for $d = f$ compared to the case when $d = d_{\rm np}$:

$$M = \frac{d_{\rm np}}{f}. \tag{7.26}$$

To be useful then, a magnifying glass needs $f <$ 25 cm. If you look up "loupes", you can find specifications of magnifiers that are used by jewellers and in electrical work for instance. Standard models have $f =$ 50 mm and 25 mm offering magnifications of 5 and 10 times. It is hard with a single lens to do much better than this because high power lenses start to violate the paraxial approximation and aberrations become significant. Much higher magnifications are possible with microscopes which can be thought of as an initial lens, the "objective" which produces a real image of the object to be examined followed by a magnifying lens with the real image placed at its focus (Fig. 7.15). This gives a magnification of v/u from the first lens times $d_{\rm np}/f_2$ for the second lens. With a factor of 10x from the eyepiece and 100x from the objective, magnifications as high as 1000x are possible.

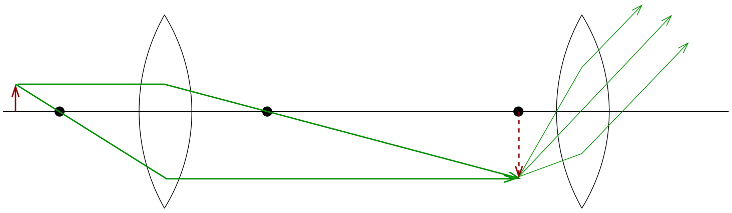

Figure 7.15: Schematic of a microscope. The object on the left is imaged by the first lens to produce a much larger real image. This image acts as the object for a magnifying lens ("eyepiece" or "ocular"). The emergent rays are parallel to give a virtual image at infinity.

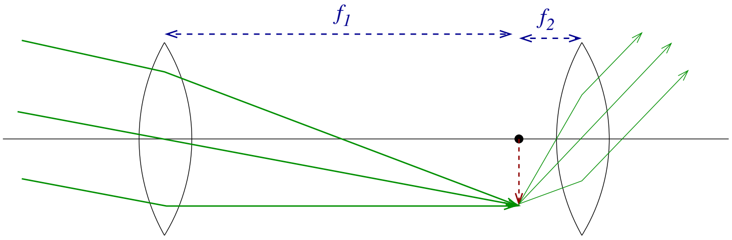

Figure 7.16: Schematic of a telescope. Parallel rays from infinity are imaged at focal length f_1 from the objective lens (left-hand lens). This is examined using the short-focal length eyepiece on the right. Magnification here is seen in the increase in the angle of the rays passing through the centre of the eyepiece compared to the objective.

Telescopes and Binoculars

Telescopes have a similar configuration to microscopes except the object is distant from the objective (Fig. 7.16). In the case of telescopes, rays from infinity form a real image in the focal plane of the objective (a detector there would record the image directly). An eyepiece is used to view this image. The telescope shown has an angular magnification equal to the ratio of the focal lengths, $M = f_1/f_2$.

Fig. 7.16 is shows schematically the angular magnification, but misleading when it comes to the relative sizes of the lenses. If we look only at rays travelling parallel to the axis (Fig. 7.17), we see that the lens diameters are naturally in the same ratio as the focal lengths,

$$\frac{D_1}{D_2} = \frac{f_1}{f_2}, \tag{7.27}$$

otherwise one or other of the lenses will have effectively unused glass. In fact, the eyepiece should be a little larger to collect non-parallel rays, but nevertheless the formula 7.27 gives an idea of the right sizes. To take an example, a common specification of binoculars (essentially a folded telescope) says they are "10x50", meaning a magnification $M = f_1/f_2 = 10$, with

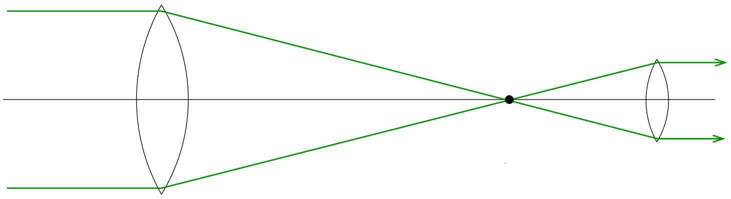

Figure 7.17: Schematic of a telescope with rays travelling parallel to the axis to show the natural relative sizes of the lenses.

an objective lens of diameter 50 mm. This means that the exit beam will have diameter $D_2 = 5$ mm, matching the maximum pupil diameter of the human eye.

When set to match the pupil of the eye, telescopes and binoculars do not change the brightness of an object. By "brightness" we mean the power per square degree on the sky received from the object. This is because the extra amount of light gathered by the telescope is the ratio of the area of the objective to the area of the eye's pupil, $(D_1/D_2)^2$ because we are matching the eyepiece diameter to the eye. However given the angular magnification, this power comes from a solid angle on the sky that is also $(D_1/D_2)^2$ times larger. The two factors cancel and there is no change. This means that you don't get blinded by looking at the Moon through a telescope, even though it can appear very bright at first if one's eye has got used to dark conditions. In practice the brightness will always be less because of losses in the telescope. Be wary when buying telescopes and binoculars where there can sometimes be an overemphasis on the "magnification". 10x50 and 8x40 are good matches to the eye, but 20x40 for example with an exit beam of 2 mm is not (at least in low light level applications), and is likely to be disappointing. One final point to note is that the simple telescope illustrated in Fig. 7.16 inverts the object. This is no problem for astronomical applications, but inverting prisms are added in the case of binoculars intended for terrestrial applications.

Chapter 8

Wave Optics

Even when designing optical instruments purely from the principles of geometric optics, one needs to be aware of the wave nature of light. This imposes an absolute limitation on the resolution that can be achieved, even if all components are manufactured with perfect precision. The wave nature of light is also critical to the widespread use of anti-reflection and related coatings in optical instruments.

8.1 The resolution of optical instruments

If lenses are correctly manufactured, within the approximations of geometric optics, a point-like object should be imaged into a point-like image. Even in this perfect case, the image would not in fact be point-like because of the wave nature of light. Fermat's principle again offers insight. If one considers moving a very small distance from the image of a point object, the phases along the ray paths contributing to the image will begin to change and start to cancel, leading to a drop in intensity. The drop will not be immediate. Typically, you will need to move away by a distance comparable with the wavelength of light for complete cancellation to occur. This is a fundamental limit on the resolution of optical devices.

The limit is most easily expressed in terms of the spread in angle caused by diffraction owing to the limited part of the wavefront accepted by the first lens (objective) in an optical train. For a circular lens (almost universal), the minimum angle between the direction of maximum intensity and zero intensity is given by

$$\Delta\theta = 1.22\frac{\lambda}{D}, \tag{8.1}$$

in radians, where λ is the wavelength of light and D the diameter of the objective lens. The factor 1.22 emerges from Bessel functions associated with the cylindrical symmetry of the problem. The factor λ/D can be understood through the schematic shown in Fig. 8.1 which shows a portion of wavefront of width D travelling from left-to-right.

We are interested in the light intensity at some distant point in a direction slightly off from the direction of travel. If the change in direction is enough that there is a distance offset of λ across the whole width of the wavefront, then, when contributions are added up (via integration) across the wavefront, any two points a distance $D/2$ apart will have opposite

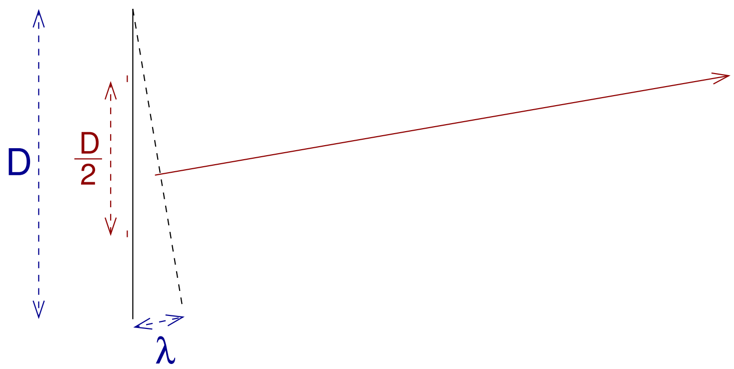

Figure 8.1: Schematic to illustrate the origin of the diffraction limit factor, λ/D. The solid black line shows a wavefront of finite width D which is travelling from left to right. The equivalent ray is a line pointing directly right (not shown). The figure shows a tilted ray (red arrow) with a dashed line drawn perpendicular to it to show what its equivalent wavefront would be. The tilt is just enough to give a shift of one wavelength with respect to the true wavefront. The strength of distant wavefronts in the tilted direction is obtained by summing up all contributions across the wavefront. If the point of interest is far away, then the distance barely varies apart from the 0 to λ tilt across the wavefront. In this case, any two points $D/2$ apart on the wavefront will be π out of phase and will cancel each other. The end result is zero flux in the direction shown, and this will be the smallest tilt away from the left-right direction that leads to zero flux, hence defining the diffraction limit.

phases and cancel in pairs. This will lead to zero flux in that direction and defines the degree of spreading of the light rays. The angle of tilt can be seen to be given by $\tan\theta = \lambda/D$, which for small angles gives $\theta = \lambda/D$. The factor 1.22 emerges from Bessel functions associated with the circular aperture which is effectively narrower on average than a square aperture of the same width. The above value is known as the "diffraction limit" and it is common to hear the expression "diffraction-limited optics" meaning that the above limit is what sets the resolution of a device (rather than, for instance, crudely manufactured lenses).

The way in which the light intensity drops away from a focal point can be addressed quantitatively through application of what is called Kirchoff's diffraction formula, which can be thought of as the mathematical version of Huygen's principle from which Fraunhofer diffraction can be derived as an approximation along with its upgraded counterpart, Fresnel diffraction. An illustration of this is shown in Fig. 8.2. Even this very simple case is complex to investigate analytically, so we don't consider it further, but it is worth knowing that there is a way to quantitatively implement Huygens' principle which is the foundation of all practical applications of wave optics. Appendix J contains the bare bones of Kirchoff's theory.

Example 8.1. Estimate the theoretical angular resolution of the human eye, and thus the size of the light receptors (cones) in the retina.

Answer. The focal length of the human eye is $f \approx 22\,\text{mm}$, and the pupil diameter in reasonable light is $D \approx 2\,\text{mm}$. The wavelength of green light $\lambda = 500\,\text{nm}$. Therefore the diffraction-limited resolution of the eye is

$$\Delta\theta = 1.22\frac{\lambda}{D} = \frac{1.22 \times 5 \times 10^{-7}}{2 \times 10^{-3}} = 3.05 \times 10^{-4}\,\text{rad} = 0.0175°. \tag{8.2}$$

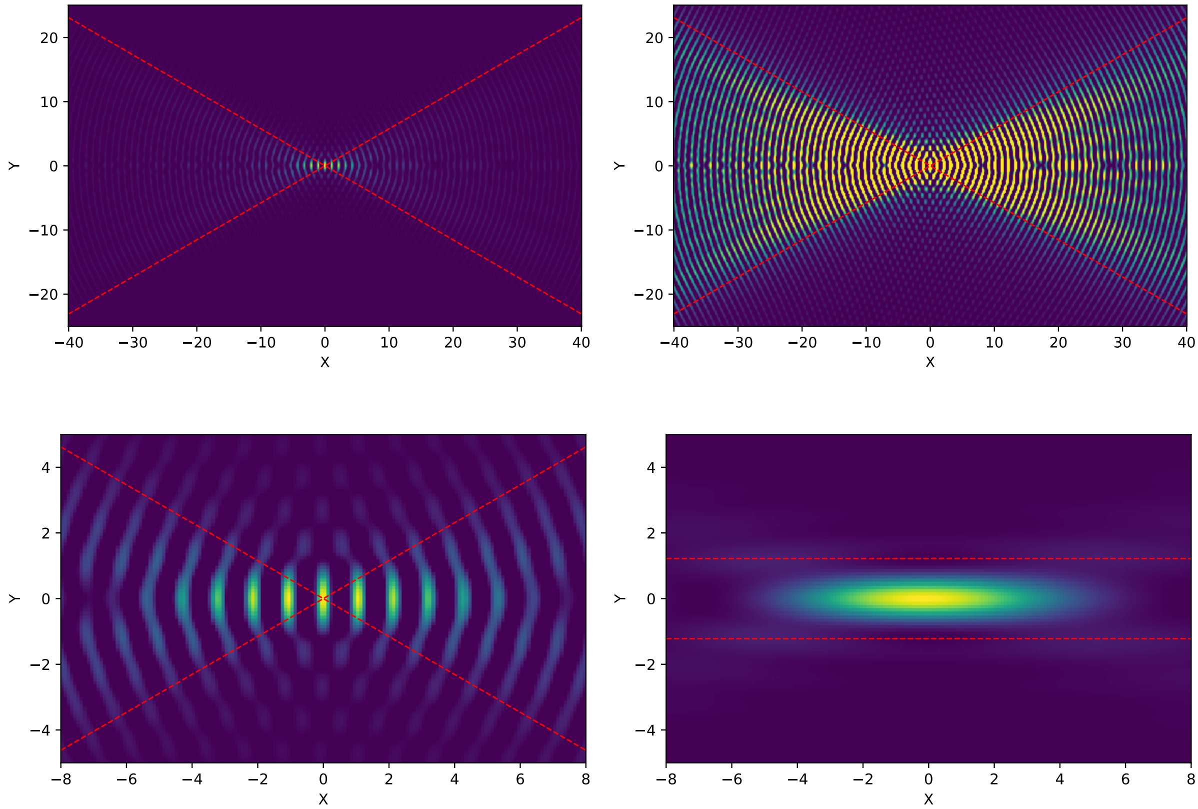

Figure 8.2: *Top row:* The wave field at an instant of time around the focus of a converging spherical wavefront, as would come from a converging lens. All scales are in terms of the wavelength of the light. The wavefronts in this figure start 100 wavelengths to the left, and the light travels from left to right in the figure. Seen from the focal point, the portion of spherical wavefront has half opening-angle of 30°. The dashed lines mark the boundary predicted by geometrical optics. The image on the top-right is the same as the top-left except the upper intensity threshold has been lowered by a factor of 20 to bring up weak features. *Bottom row:* a zoom of the wave field around the focus (left), and the corresponding time-averaged intensity (right). The dashed lines on the right indicate the predicted vertical location of the first zero moving up or down from the focus, using $1.22\lambda/D$. See appendix J for the theory due to Huygens, Fresnel and Kirchoff underlying these images (off syllabus for PX263).

Multiplying by the focal length, this angle projects to 6.7 µm on the retina. Ideally one would hope for at least 2 receptors per "resolution element" of this size, which works out at a density of $\approx$ 90 000 mm^{-2}. Peak densities (cones in the fovea, rods elsewhere) are actually $\approx$ 150 000 mm^{-2}, but we haven't accounted for the multi-colour nature of cones which would require a higher packing density. (For comparison, typical CCD and CMOS detectors have pixel sizes in the range 5 to 15 microns.)

Application to microscopy
The initial stage of a microscope magnifies by $M = v/u$, and with $u > f$, $v = fu/(u - f)$, hence

$$M = \frac{f}{u - f}, \tag{8.3}$$

where f is the focal length of the objective. For high magnification we want to place the object only just beyond the focal point of the objective. The diffraction limit then sets a limit to the smallest features we can see in a specimen of

$$\ell_{min} = 1.22\frac{\lambda}{D}u \approx 1.22\lambda\frac{f}{D}. \tag{8.4}$$

This shows that one wants the objective lens to have as small a value of f/D (known as the "f-number") as possible. A more complete analysis yields a closely-related quantity $\sin\theta$ where θ is the half-opening angle $\tan\theta = D/2f$. This is called the "numerical aperture" in microscopy. Small f-number / large numerical aperture lenses are hard to make, and so the objective lens is critical in high-quality microscopes. At best numerical apertures of order unity are possible, so the spatial resolution of microscopes is of order the wavelength of light (it's always possible to do worse!).

Application to telescopes
The diffraction limit applies directly to telescopes where one critical performance metric is angular resolution. A recent spectacular example was the resolution of the "event horizon" close to the large black-hole in the galaxy M87. This was right at the diffraction limit of a network of radio telescopes combined to effectively give an "objective" comparable in size to the Earth. The observation was made at a wavelength of 1.3 mm, and with the whole diameter of the Earth, the angular resolution was of order 1.3×10^{-10} rad, around 2 million times more acute than the human eye.

8.2 Thin film interference

Colours are generated from reflections off thin films of oil or bubbles. Bubbles are particularly worth attention when they approach bursting. Just before they burst, the colours fade and they become clear. These sorts of behaviour can be explained through Fresnel's relations and study of the reflections from closely-spaced double layers. The physics of such films is of great practical importance for controlling (usually reducing) reflections inside optical devices where a few percent reflection off every one of dozens of interfaces could ruin their operation. Dielectric coatings are used for the most efficient mirrors ($> 99.999\%$ reflectivity can be achieved). They are used, for instance, in the mirrors of gravitational wave detectors.

Anti-reflection Coatings

We will consider thin film interference at normal incidence only; the more general case is left as an exercise. The basic effects are easy to appreciate through consideration of Fresnel's coefficients at each interface. Consider a case with two interfaces with light travelling from medium 1 towards medium 2, and, if transmitted, finally to medium 3. This could be air (n_1) to glass (n_3), with a thin coating on the glass (n_2). At the first interface there is a reflection coefficient of

$$r_{12} = \frac{n_1 - n_2}{n_1 + n_2}, \tag{8.5}$$

while at the second it is

$$r_{23} = \frac{n_2 - n_3}{n_2 + n_3}. \tag{8.6}$$

We will assume that the reflections are weak, so that we can ignore the reduction in amplitude of the wave in medium 2 when computing the amplitude reflected at the 2/3 interface. We will also ignore any subsequent reflections. This is called the weak reflection approximation.

The two reflections considered combine to give the total reflected and transmitted amplitudes. We have to account for the phase shift between the beam reflected at the first interface and beam reflected off the second interface. This phase is picked up when travelling from the first to the second interface and back again. If the layer has thickness d, the phase shift $= 2k_2 d = 2(2\pi/\lambda)d$, where $\lambda = \lambda_0/n_2$ is the wavelength of light in the thin layer. If the layer has thickness $d = \lambda/4$, the phase shift is π, and the second reflection will act to cancel the first resulting (potentially) in zero reflection. For exact cancellation we require that

$$r_{12} - r_{23} = \frac{n_1 - n_2}{n_1 + n_2} - \frac{n_2 - n_3}{n_2 + n_3} = 0, \tag{8.7}$$

with the minus sign appearing because of the π radian phase shift induced by the quarter-wave layer. It follows that

$$(n_1 - n_2)(n_2 + n_3) = (n_2 - n_3)(n_1 + n_2), \tag{8.8}$$

which gives

$$n_2^2 = n_1 n_3. \tag{8.9}$$

This equation gives the refractive index needed for a quarter-wavelength thick coating to lead to zero reflection at the interface between two media with refractive indices n_1 and n_3.

A case of practical importance involves reflections at air/glass interfaces with $n_1 = 1$ and $n_3 = 1.52$ (crown glass). In this case the ideal coating would have refractive index $n_2 = \sqrt{n_1 n_3} = 1.225$. We are restricted by what materials are available, but one can still get useful reductions in reflectivity even with refractive indices quite far off the mark. Magnesium fluoride is a commonly-used material. Although its refractive index is higher than the optimum $(n = 1.38)$, it can nevertheless cut reflection losses from 4% to $\sim 1\%$ (see Fig. 8.3).

The cancellation works best at the precise wavelength for which the coating is a quarter of a wavelength thick. Usually this is designed to be somewhere in the middle of the range of wavelengths the eye is sensitive to, which is green/yellow in colour. This means that coated optics will tend to reflect blue and red light a little more, giving the characteristic purple colour of reflections from camera optics. With multiple layers, it is possible to widen the range of wavelengths over which the reflectivity is reduced.

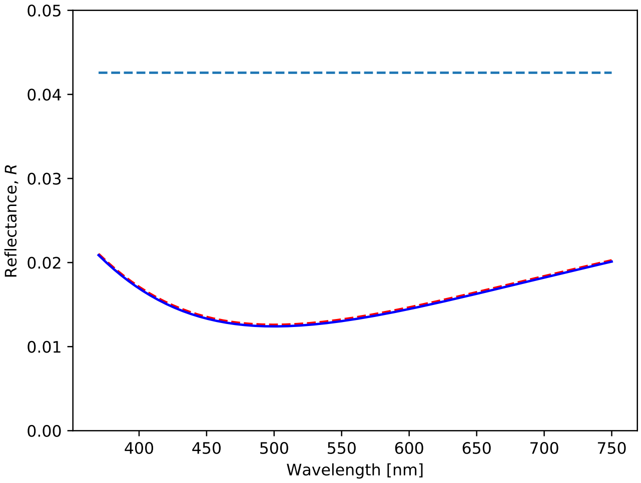

Figure 8.3: The solid line shows the reflectance at normal incidence from a glass block with $n = 1.52$ with a $d = 125$ nm magnesium fluoride $(n = 1.38)$ anti-reflection layer, calculated using the weak reflection approximation (see text). The horizontal dashed line shows the reflectance in the absence of any coating, while the curved dashed line shows the exact value of the reflectance, showing that the weak reflection approximation is valid.

To calculate the wavelength dependence of a given layer we need to account for the general phase shift which introduces a factor of $\exp(i2k_2d)$ into the r_{23} term so that the overall reflection coefficient becomes

$$r = \frac{n_1 - n_2}{n_1 + n_2} + \frac{n_2 - n_3}{n_2 + n_3} e^{i2k_2d}. \tag{8.10}$$

The reflectivity $R = |r|^2 = rr^*$ can be shown to be

$$R = \frac{(n_1 - n_2)^2(n_2 + n_3)^2 + (n_1 + n_2)^2(n_2 - n_3)^2 + 2(n_2^2 - n_1^2)(n_3^2 - n_2^2)\cos(2k_2d)}{(n_1 + n_2)^2(n_2 + n_3)^2}. \tag{8.11}$$

An example of this relation is plotted in Fig. 8.3. The relation 8.11 can be seen to be a good approximation to the exact expression, which is derived in appendix K.

Dielectric Mirrors

Quarter-wave layers, which add a π relative phase shift between the two contributing reflections, reduce reflectivity when $n_1 < n_2 < n_3$ because the two contributing reflections from the 1/2 and 2/3 interfaces are in phase until the depth-dependent phase-shift is added. If $n_1 < n_3 < n_2$ (the dielectric coating has a higher refractive index than air and glass), the 1/2 and 2/3 reflections are intrinsically in anti-phase, and the additional π phase shift from a quarter-wave layer moves them back into phase again. This boosts the reflectivity. Multiple "up/down" layers of this sort can be used to make very high reflectivity mirrors. Fig. 8.4 shows the reflectivity with a single layer

Soap bubbles

Soap bubbles are an easily accessible and interesting example of thin film interference and are well worth careful study. They consist of closely-spaced air/water, water/air interfaces. The

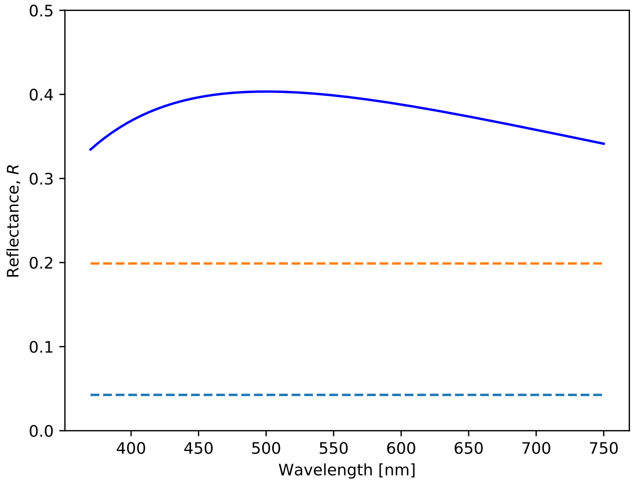

Figure 8.4: The solid line shows the reflectance at normal incidence from a glass block with $n = 1.52$ with a $d =$ 125 nm titanium dioxide coating $(n = 2.61)$ to boost the reflectivity, calculated exactly in appendix K (the weak reflection approximation does not work well for the case of strong reflection considered here). The horizontal dashed lines show the reflectances from the glass alone (lower line) and titanium dioxide alone (upper) showing the significant increase possible with just a single layer.

Fresnel formulae show that

$$r_{12} = \frac{n_1 - n_2}{n_1 + n_2} = -0.14, \tag{8.12}$$

$$r_{23} = \frac{n_2 - n_3}{n_2 + n_3} = +0.14, \tag{8.13}$$

assuming $n_2 = 1.33$. In this case, because of the phase inversion at the first but not the second interface, a quarter-wave thickness leads to *maximum* reflectance of $0.28^2 \approx 8\%$, and this will be strongly coloured as the reflectivity is wavelength dependent and because of the balanced amplitudes of the reflections.

As bubbles age, mass drains under gravity and they get thinner and eventually burst. If you watch this process carefully, you will see the reflection weaken, lose colour and usually the top of bubble will be almost completely clear just before bursting. The two formulae above show why this is the case. As the thickness decreases then the distance-dependent phase shift between the reflections $e^{i2k_2d} \to 1$ (when $d \ll \lambda$). The reflections then cancel for all wavelenths as a result of the built-in π phase-shift resulting from $n_1 < n_2$ at the air/water interface while $n_2 > n_3$ at the water/air interface.

Fig. 8.5 shows the reflectance from different thicknesses of soap film, and the drop of intensity for very thin films should be clear. Just accounting for the two reflections then the overall reflectance R is given by

$$\begin{aligned} R &= \left|r_{12} + r_{23}e^{i2k_2d}\right|^2, \\ &= r_{12}^2\left|1 - e^{i2k_2d}\right|^2, \\ &= r_{12}^2\left(2 - 2\cos 2k_2d\right), \\ &= 4r_{12}^2\sin^2(k_2d) = 0.0784\sin^2(k_2d). \end{aligned} \tag{8.14}$$

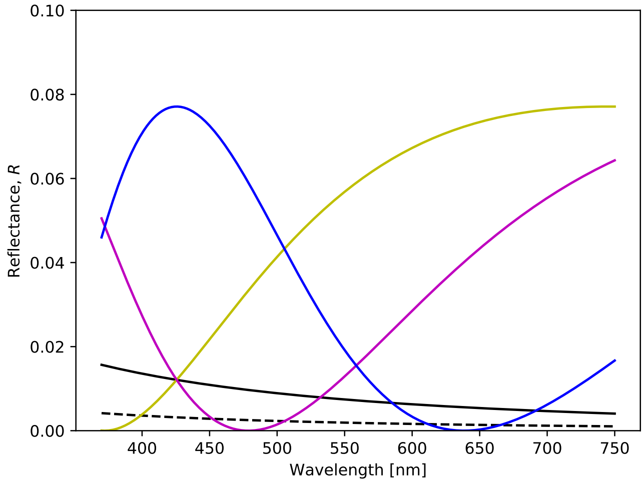

Figure 8.5: The reflectance at normal incidence from soap films of thickness 10 nm (black, dashed), 20 nm (black, solid), 140 nm (yellow), 180 nm (mauve) and 240 nm (blue). The colour coding is an attempt to represent the colour seen in normal light conditions, except for the black lines where the reflection appears white.

A precise version of this relation accounting for all reflections was used to create Fig. 8.5, but the above relation is fairly accurate and can be used to understand all the main features of interest.

Glass windows

At first sight, glass windows might seem equivalent in terms of reflections to soap bubbles, and yet they are not highly coloured. The reason for this is shown in Fig. 8.6. Even for very thin panes 10 000 nm (0.01 mm) the response can be seen to vary too rapidly in wavelength to be discernible to the human eye because of its broad wavelength response, although you may sometimes see soap bubble-like interference colours from thin cracks in glass.

Angle dependence

All the above discussions assumed normal incidence. At non-normal incidence the reflection coefficients off each interface change and become polarisation dependent. The main effect is the result of the depth-related phase factor. The geometry for this is illustrated in Fig. 8.7. The important element involves the relative phases of the two reflected beams that originated from the same incident beam (see figure caption for more explanation). The path length (in terms of phase) from A to B is $k_2 d/\cos t$ where t is the angle of refraction (measured from the surface normal as usual) within the layer. The total path length A to B to C is therefore $2k_2 d/\cos t$. From geometry the distance AD is related to AC by AD $=$ AC $\sin i$ where i is the angle of incidence, while

$$\text{AC} = \frac{2d\sin t}{\cos t}. \tag{8.15}$$

The path length (taken as a phase) corresponding to AD is

$$\frac{2k_1 d\sin t\sin i}{\cos t}. \tag{8.16}$$

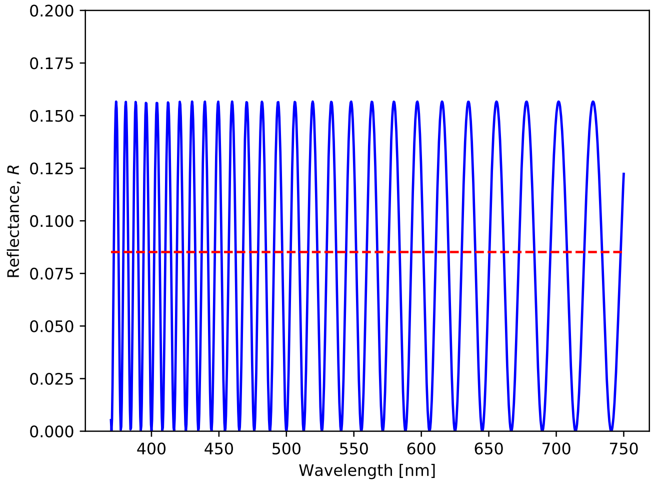

Figure 8.6: The reflectance at normal incidence from a glass window of thickness 10 000 nm. The rapid variation would not be discernible to the eye because of the eye's broad wavelength response. The dashed line shows a level equal to twice the reflectance off a single air/glass interface which matches the average value of the rapidly varying response.

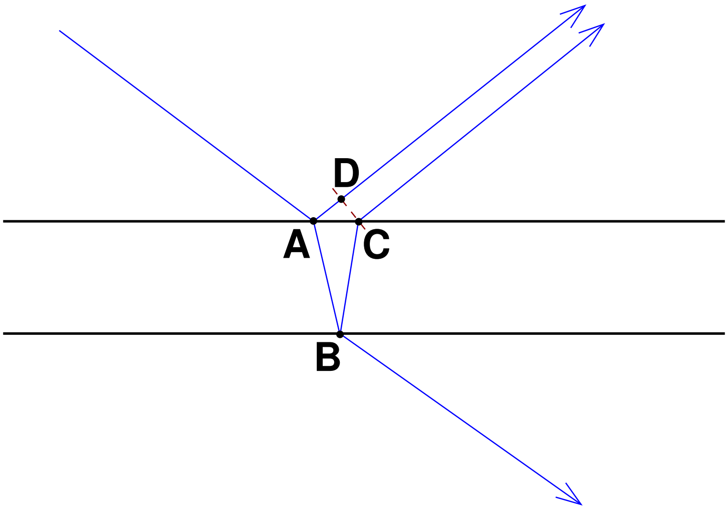

Figure 8.7: An angled reflection from a thin film. The relative phases of the two reflected beams, which are the same at point A, have to be compared at two points (D and C) on the same wave-front. One therefore needs to allow for the distances A to B to C versus A to D.

The path difference between light at D and light at C is

$$\Delta = \frac{2k_2 d - 2k_1 d \sin t \sin i}{\cos t}. \tag{8.17}$$

From the phase condition (Snell's law, Eq. 6.7), $k_1 \sin i = k_2 \sin t$, we have

$$\Delta = \frac{2k_2 d - 2k_2 d \sin^2 t}{\cos t} = 2k_2 d \cos t. \tag{8.18}$$

Now consider a quarter-wave layer for which $2k_2 d \cos t_2 = \pi = 4\pi d n_2 \cos t_2 / \lambda_0$. Thus

$$\lambda_0 = 4 d n_2 \cos t_2 \tag{8.19}$$

maintains the "quarter wave" condition. Here n_2 is the refractive index of the optical coating, t_2 is the transmitted angle within this coating and λ_0 is the vacuum or (approx.) air wavelength.

The dependence on angle in Eq. 8.19 means that, as optical components with dielectric coatings are tilted away from normal incidence, the wavelengths for which they work become shorter. This effect has to be accounted for when making filters to be used in converging or diverging beams of light, and means that the transmission wavelength of interference filters can be affected by any tilts to the filters. As an example consider an interference filter designed to allow light close to 500 nm to pass. If this was tilted so that $i = 1.0°$ (parallel light), then $\sin t_2 = \sin i / n_2$, so $t_2 = i/n_2$ for small angles, and $\cos t_2 \approx 1 - i^2/2n_2^2$. For $i = 1.0°$, the difference from one is -3.8×10^{-5} assuming $n_2 = 2$. This is equivalent to a shift in velocity measured via the Doppler effect in an astronomical observation of ≈ 10 km s^{-1}. This can be relevant to measurements using narrow band filters set up to identify specific spectral lines. Filters can pass a wavelength band a few Angstrom wide and the lines can pass or not depending on small deviations of angles of incidence.

It is worth looking at soap bubbles through polarising filters (polaroids) to see the very strong polarisation that can result from reflections off dielectric layers at non-normal incidence. Rainbows are also surprisingly strongly polarised.

8.3 Fabry-Pérot etalons

Fabry-Pérot etalons (and closely related Fabry-Pérot interferometers) are interference devices using two surfaces of relatively high reflectivity held quite far apart. They have important applications in telecommunications, filters and lasers, amongst other things. Because of the high reflectivity, it is necessary to account for multiple reflections.

The operating principle of a Fabry-Pérot etalon is illustrated in Fig. 8.8. For simplicity, assume that each side is identical. We need to distinguish reflections and transmissions going from air to glass, and glass to air. Let the air-to-glass values be denoted r' and t', while the glass-to-air ones are r and t. These may be complex because of the possible use of metal coatings.

Consider the transmitted ray which will be made up of an infinite series of components which we can associate with the number of times they have been reflected inside the glass by the second interface and then by the first interface. The first component has been transmitted through both interfaces and will end with amplitude (relative to incident ray) $t_1 = t't$ at the

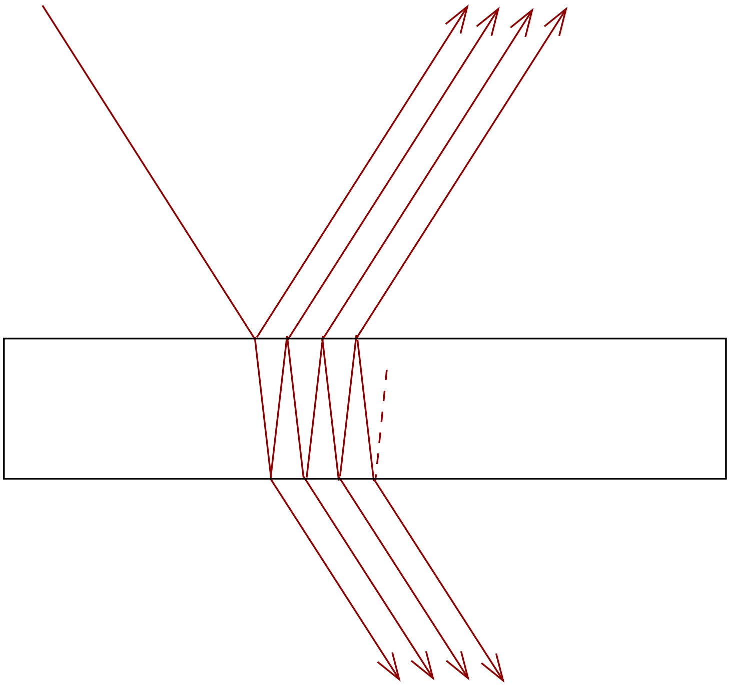

Figure 8.8: Light enters a glass block which has been coated on both sides to increase the reflectivity (could be metal or dielectric). Multiple reflections occur, and the spacing between the layers can be (and usually is) significantly more than one wavelength. The rays pointing down towards the lower-right are the transmitted rays. These are discussed in the text when determining the total transmission.

point where it exits the glass block which we will use as the phase reference point. The second transmitted component passes through the air-to-glass interface (with transmission amplitude t'), is reflected twice internally on glass-to-air interfaces (r^2), before passing through the lower glass-to-air interface (t). There is a phase delay of $2k_g d$ where k_g is the wave vector in the glass and d is the thickness of the block. This is for normal incidence and generalises to $2k_g d\cos\theta_g = 2\phi$ where θ_g is the angle of incidence and transmission inside the glass block and where we have defined $\phi \equiv k_g d\cos\theta_g$. These components form a series of rays with amplitudes:

$$\begin{aligned} t_1 &= t't, \\ t_2 &= t't\left(r^2 e^{i\Delta\phi}\right), \\ t_3 &= t't\left(r^2 e^{i\Delta\phi}\right)^2, \ldots \qquad . \end{aligned}$$

The total transmitted amplitude is

$$\begin{aligned} t_T &= tt'\sum_{n=0}^{\infty}\left(r^2 e^{i2\phi}\right)^n, \\ &= \frac{tt'}{1-r^2 e^{i2\phi}}, \end{aligned}$$

using the usual relation for the sum of an infinite geometric series.

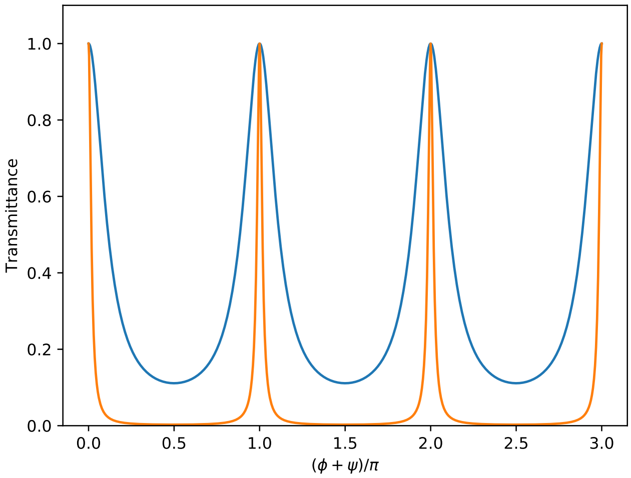

Figure 8.9: The transmittance relative to peak of a Fabry-Pérot etalon as a function of the argument ϕ (see text) divided by π. The blue curve is for a reflectance $R = 0.5$ and corresponding finesse $F = 8$, while the orange curve is for $R = 0.9$, $F = 360$.

The transmittance $T = |t_T|^2$,

$$\begin{aligned} T &= \frac{t^2 t'^2}{(1 - r^2 e^{i2\phi})(1 - r^2 e^{-i2\phi})}, \\ &= \frac{t^2 t'^2}{1 + r^4 - 2r^2 \cos 2\phi}. \end{aligned}$$

Here we have assumed that t, t' and r are real; this assumption does not alter the important physics. Using $\cos 2\theta = 1 - 2\sin^2\theta$, we obtain

$$\begin{aligned} T &= \frac{t^2 t'^2}{(1 - r^2)^2 + 4r^2 \sin\phi}, \\ &= \frac{t^2 t'^2/(1 - r^2)^2}{1 + F \sin^2\phi}, \end{aligned}$$

where F is known as the “finesse” and is given by

$$F = \frac{4r^2}{(1 - r^2)^2} = \frac{4R}{(1 - R)^2}, \tag{8.20}$$

where R is the reflectance on any one surface. If the system is lossless, it can be shown that the numerator $= t'^2 t^2/(1 - r^2)^2 = 1$, otherwise it will be less than 1, but more importantly, it is a constant.

The important part of 8.20 is the denominator. For quite modest values of R, the finesse F can be quite large, e.g. $R = 0.9$ gives $F = 360$. This then gives a transmission that peaks sharply when ϕ is an integral multiple of π. Fig. 8.9 shows this function for two different values of R. Since $\phi = k_g d \cos\theta_g = 2\pi d n_g \cos\theta_g/\lambda_0$, the etalon can act as a narrow-band filter; two in series could be tuned to coincide at just one wavelength. If either d or θ_g can be continuously varied, a Fabry-Pérot etalon can scan through a range of wavelengths. Laser cavities, which have mirrors at either end, act as Fabry-Pérot etalons, selecting out narrow peaks from the natural wavelength spread of the lasing medium. If an extra etalon is added, just a single one of these peaks can be selected. If an etalon is used to observe an extended source which produces sharp emission lines, then specific values of θ_g will be transmitted and a series of rings results. Images of these are easily looked up online. Fabry-Pérot etalons are in essence optical resonators and of great utility.

8.4 Conclusion

That's it for PX263. Electromagnetism and Optics are both topics of huge practical importance, and are used daily by manufacturers, usually in the form of sophisticated software packages that implement the fundamental physics that we have looked at in the module through techniques such as finite element analysis and ray tracing.

Modules in later years take things further. PX3A3 *Electrodynamics* looks at radiation and the Lorentz invariance of Maxwell's equations. It introduces and works with the vector potential (see Appendix D), which is where Maxwell started. The vector potential is crucial to understanding the quantum effects of em radiation and is the simplest case of a gauge field. Gauge fields are studied in PX454 *Theoretical Particle Physics*.

PX3A4 *Plasma Physics and Fusion* and PX456 *Solar and Space Physics* study plasmas. Plasmas are fluids of charged particles. The motion of these charged particles is controlled by the electromagnetic fields which are imposed from outside and by the fields which the moving charged particles themselves set up.

The astrophysics modules draw heavily upon electromagnetism. Apart from some work looking for neutrinos from astrophysical sources and the recently detected gravitational waves, electromagnetic waves are the evidence we work with when studying the development and structure of the Universe.

Appendix A

Index expressions

A brief explanation of the appendices is in order. The appendices contain material that is for the most part beyond what you are expect to know for this module, although they should be understandable and may answer some questions you might have. Having this material at the end should avoid breaking the flow of the main notes. You may find some of the material in the appendices appearing in a slightly different form in the main notes; in this case you are expected to know it. The main notes therefore define what you are expected to know for this module, while the appendices fill in a few holes.

A.1 An introduction to index expressions

One sometimes has a need for expressions that can't be easily expressed in the forms such as $\nabla \cdot \boldsymbol{A}$ or $\nabla(f\boldsymbol{A})$ encountered in the main notes. For instance, if you try to expand $\nabla(\boldsymbol{A} \cdot \boldsymbol{B})$ you end up with terms like "$\boldsymbol{A}\nabla\boldsymbol{B}$" where there should still be a "dot" product-like association between the $\boldsymbol{A}$ and $\boldsymbol{B}$ but the notation does not allow for it because the ∇ has to apply to $\boldsymbol{B}$ not $\boldsymbol{A}$ (there will be another term where the ∇ applies to $\boldsymbol{A}$ not $\boldsymbol{B}$).

In such cases, it is often easier to revert to components, in a form provided by indices that are taken to represent any of the components. Thus instead of $\boldsymbol{A}$ and $\boldsymbol{B}$, we write A_i and B_j where the indices i and j stand for any of x, y or z (or 1, 2, 3) if you prefer. This may well seem retrogressive compared to the index-free notation, and there are ways around this, but it does quickly yield some progress and in practice "index manipulations" are the bread and butter of calculations using vector and tensor field calculus. Consider two vectors $\boldsymbol{V}$ and $\boldsymbol{W}$. While there might be no obvious meaning for $\boldsymbol{V}\boldsymbol{W}$ (it's neither a scalar nor a cross product), V_iW_j, is perfectly well defined: it's a set of 9 numbers indexed by i and j (these are the components of a tensor, but we don't need to know this).

A.2 Scalar products and the summation convention

Einstein's "summation convention" states that we sum over repeated indices. With this convention, the scalar product can be written as

$$\boldsymbol{V} \cdot \boldsymbol{W} = \sum_{i=1}^{3} V_i W_i = V_i W_i. \tag{A.1}$$

The presence of a repeated index implies that the expression should be interpreted as the sum over all possible values of the index. (Sometimes this applies only to repeated indices where one is a subscript and the other a superscript, this is common in General Relativity for instance, but is not a distinction we need to make here.) Note that we could equally well have written $V_j W_j$ or $V_k W_k$. In each case summation is implied over the index whether it is labelled i, j or k, hence such indices are known as "dummy indices".

It is important to distinguish between indices when they are not to be summed over, so that $V_i W_j \neq V_i W_i$: the left-hand expression represents 9 numbers whereas the right-hand expression is a single number and it would be a grave mistake to have written i instead of j in the left-hand side as it would be an entirely different quantity.

Now let's re-consider $\nabla(\boldsymbol{A} \cdot \boldsymbol{B})$. Using indices this can be written as

$$\nabla(\boldsymbol{A} \cdot \boldsymbol{B}) = \partial_i A_j B_j, \tag{A.2}$$

with summation implied over the dummy index j (i is a free index). But this expression is easily expanded out using the product rule:

$$\nabla(\boldsymbol{A} \cdot \boldsymbol{B}) = \partial_i A_j B_j = A_j \partial_i B_j + B_j \partial_i A_j, \tag{A.3}$$

which we leave in this form since the two right-hand terms can't easily be re-written in non-index form. Similarly

$$\begin{aligned} \nabla \cdot (f\boldsymbol{A}) &= \partial_i f A_i, && \text{(A.4)} \\ &= f \partial_i A_i + A_i \partial_i f, && \text{(A.5)} \\ &= f(\nabla \cdot \boldsymbol{A}) + \boldsymbol{A} \cdot (\nabla f). && \text{(A.6)} \end{aligned}$$

In this case one can re-express the result, but note how trivial the intermediate index-expression steps are.

A.3 Cross products

To write out cross product in index notation we define a set of symbols ϵ_{ijk} dependent upon three indices i, j and k, which take the value $+1$ for $ijk = 123$ and the cyclic permutations thereof (231, 312), -1 for $ijk = 132$ and cyclic permutations thereof, and otherwise zero. With this definition, cross-products can be written

$$\boldsymbol{V} \times \boldsymbol{W} = V_i W_j \epsilon_{ijk}, \tag{A.7}$$

with implied summation over dummy indices i and j, while the index k is not summed over, and can take any of three values, reflecting that we treat the result as a vector. (Best just to write out explicitly with $V_1 = V_x$ etc to understand this expression.)

The numbers ϵ_{ijk} are known as the Levi-Civita or antisymmetric symbols. From their definition it should be clear that $\epsilon_{ijk} = \epsilon_{kij} = \epsilon_{jki}$, while $\epsilon_{ijk} = -\epsilon_{jik}$. Another useful quantity is the Kronecker delta, δ_{ij} defined as $= 1$ for $i = j$ and 0 otherwise, which you will have encountered before. Clearly $\delta_{ij} = \delta_{ji}$, while $\delta_{ii} = 3$ in three dimensions. Here is a simple example of the Kronecker delta in action:

$$V_i W_j \delta_{ij} = V_i W_i = \boldsymbol{V} \cdot \boldsymbol{W}. \tag{A.8}$$

The Levi-Civita symbols and Kronecker delta are related in the following expression which is useful in handling multiple cross-products:

$$\epsilon_{ijk}\epsilon_{lmk} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}. \tag{A.9}$$

(Summation over dummy index k; indices i, j, l and m are free.) Again, this can be shown by writing out a few components, remembering the summation over k. Consider the vector triple product for instance (in full detail):

$$\begin{aligned}
\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) &= a_i(b_j c_k \epsilon_{jkm})\epsilon_{iml}, && \text{(A.10)}\\
&= a_i b_j c_k \epsilon_{jkm}\epsilon_{lim}, && \text{(A.11)}\\
&= a_i b_j c_k \left(\delta_{jl}\delta_{ki} - \delta_{ji}\delta_{kl}\right), && \text{(A.12)}\\
&= a_i b_l c_k \delta_{ki} - a_i b_j c_k \delta_{ji}\delta_{kl}, && \text{(A.13)}\\
&= a_i b_l c_i - a_i b_j c_k \delta_{ji}\delta_{kl}, && \text{(A.14)}\\
&= a_i b_l c_i - a_i b_i c_k \delta_{kl}, && \text{(A.15)}\\
&= a_i b_l c_i - a_i b_i c_l, && \text{(A.16)}\\
&= (\boldsymbol{a} \cdot \boldsymbol{c})\boldsymbol{b} - (\boldsymbol{a} \cdot \boldsymbol{b})\boldsymbol{c}, && \text{(A.17)}
\end{aligned}$$

a standard result. It is the first to the second line that shows the power of the index-based notation because the slightly tricky triple product has been boiled down to simple multiplications and summations. Note the cyclic rotation $\epsilon_{iml} = \epsilon_{lim}$ used to obtain the product of the epsilons into the form needed for application of Eq. A.9. We encounter pretty much this relation in the form

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}, \tag{A.18}$$

although note some sleight-of-hand has been used on the first term on the right-hand side since ∇ can't be left without anything to operate on.

Gauss's and Stokes's theorems can be written

$$\int_V (\partial_i W_i)\, dV = \oint_S W_m\, dS_m, \tag{A.19}$$

$$\int_S (\partial_i W_j \epsilon_{ijk})\, dS_k = \oint_C W_m\, d\ell_m, \tag{A.20}$$

where ∂_i denotes differentiation with respect to coordinate with index i. It now becomes apparent that one could replace W_i in the first expression with for example $W_i V_j$ (simple multiplication by a function we decide to call V_j) and deduce

$$\int_V (\partial_i W_i V_j)\, dV = \oint_S W_m V_j\, dS_m. \tag{A.21}$$

Similarly $W_i \to W_{ijkl}$ yields

$$\int_V \left(\partial_i W_{ijkl}\right) dV = \oint_S W_{mjkl}\, dS_m. \tag{A.22}$$

The W_{ijkl} here could be components of a tensor. This is a simple manner in which to extend these two theorems. As a final example, let's re-express $\nabla \times (f\boldsymbol{W})$:

$$\begin{aligned}
\nabla \times (f\boldsymbol{W}) &= \partial_i f W_j \epsilon_{ijk}, && \text{(A.23)}\\
&= (\partial_i f) W_j \epsilon_{ijk} + f \partial_i W_j \epsilon_{ijk}, && \text{(A.24)}\\
&= (\nabla f) \times \boldsymbol{W} + f \nabla \times \boldsymbol{W}. && \text{(A.25)}
\end{aligned}$$

The main thing to remember is the need for the product rule when moving ∂_i to the other side of a symbol in the index expressions.

Appendix B

Grad, div and curl in non-Cartesian coordinates

B.1 Introduction

The expression $\nabla = \hat{\boldsymbol{x}}\,\partial_x + \hat{\boldsymbol{y}}\,\partial_y + \hat{\boldsymbol{z}}\,\partial_z$ is particular to Cartesian coordinates. Sometimes the geometry is such that it can be preferable to work in other coordinates systems such as cylindrical or spherical polars. Consider the magnetic field around a wire carrying current I. Suppose the wire lies long the z-axis about which the usual azimuthal angle θ of cylindrical polars is measured and that the current flows in the direction of positive z. Then one can write the field at radius r from the wire as

$$\boldsymbol{B} = \frac{\mu_0 I}{2\pi r}\hat{\boldsymbol{\theta}}, \tag{B.1}$$

where $\hat{\boldsymbol{\theta}}$ is a unit vector pointing in the direction of increasing azimuthal angle with r and z fixed. If we want to show that $\nabla \cdot B = 0$ we could first convert to cartesian coordinates, which yields

$$\boldsymbol{B} = \frac{\mu_0 I}{2\pi}\left(-\frac{y}{\sqrt{x^2+y^2}}\,\hat{\boldsymbol{x}} + \frac{x}{\sqrt{x^2+y^2}}\,\hat{\boldsymbol{y}}\right), \tag{B.2}$$

and then apply the Cartesian form of ∇. However, the simpler form of the field in polar coordinates suggests there may be a better way. This appendix looks at this in more detail.

We focus only upon *orthogonal* coordinate systems which are those where the unit vectors defined by each ordinate changing with the others held fixed (e.g. $\hat{\phi}$) are mutually perpendicular to each other at any point. This applies in particular to cylindrical and spherical polar coordinates. This is material that you will have encountered in second year maths courses dealing with vector calculus (e.g. PX275, *Mathematical methods for physicists*). If you are already happy with that, then read no further; I only include it here as a refresher and for completeness. For the purposes of PX263, you only need to know that there are different forms for ∇ and ∇^2 (the Laplacian) according to the coordinate system being used, and how to apply the different expressions (which will be given; don't feel you need to memorise them).

B.2 The gradient

The fundamental relation defining the gradient operator is $df = \nabla f \cdot d\boldsymbol{\ell}$ for a scalar function f. In spherical polars, (r, θ, ϕ), the vector displacement may be written

$$d\boldsymbol{\ell} = (dr)\,\hat{\boldsymbol{r}} + (r\,d\theta)\,\hat{\boldsymbol{\theta}} + (r\sin\theta\,d\phi)\,\hat{\boldsymbol{\phi}}. \tag{B.3}$$

where $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ are mutually orthogonal unit vectors which point in the direction of increasing r, θ and ϕ. Since we know that if f is regarded as function of r, θ and ϕ then

$$df = \frac{\partial f}{\partial r}dr + \frac{\partial f}{\partial \theta}d\theta + \frac{\partial f}{\partial \phi}d\phi, \tag{B.4}$$

and we can write

$$\nabla = \hat{\boldsymbol{r}}\,\frac{\partial}{\partial r} + \hat{\boldsymbol{\theta}}\,\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\boldsymbol{\phi}}\,\frac{1}{r\sin\theta}\frac{\partial}{\partial \phi}, \tag{B.5}$$

in spherical polar coordinates. To see this, combine this expression with the expression for $d\boldsymbol{\ell}$. Note that the "operator" parts ($\partial/\partial r$ etc) have to be written on the right of each term to avoid operating on the other symbols.

An elementary application of this is the Coulomb potential, $\psi = q/4\pi\epsilon_0 r$. Using $\boldsymbol{E} = -\nabla\psi$, the field is seen to be

$$\boldsymbol{E} = \frac{q}{4\pi\epsilon_0 r^2}\,\hat{\boldsymbol{r}}, \tag{B.6}$$

as expected. This is quicker and simpler than performing the same computation in Cartesian coordinates.

B.3 The divergence and Laplacian

A vector field $\boldsymbol{W}$ can be described in spherical polars as

$$\boldsymbol{W} = W_r\hat{\boldsymbol{r}} + W_\theta\hat{\boldsymbol{\theta}} + W_\phi\hat{\boldsymbol{\phi}}. \tag{B.7}$$

Given this and the expression given for ∇ in spherical polars (Eq. B.5), one might be tempted then to write the divergence as

$$\nabla \cdot \boldsymbol{W} = \frac{\partial W_r}{\partial r} + \frac{1}{r}\frac{\partial W_\theta}{\partial \theta} + + \frac{1}{r\sin\theta}\frac{\partial W_\phi}{\partial \phi}, \tag{B.8}$$

but this would be incorrect!. This is where things get a little tricky, because the above expression takes no account of the fact that in non-Cartesian coordinates, the unit vectors in general change with the coordinates, e.g. $\hat{\boldsymbol{\phi}}$ reverses direction if ϕ increases by π. Therefore, when we apply ∇ we need to take derivatives of the unit vectors as well as the components. This does not come up in Cartesian coordinates since the Cartesian unit vectors do not vary with position.

We require a set of derivatives of the form $\partial\hat{\boldsymbol{\phi}}/\partial r$ for all combinations of ordinates (9 in all). Some are easily seen to be zero, (e.g. $\partial\hat{\boldsymbol{r}}/\partial r = 0$), but the following are not (using the compressed notation $\partial_\theta \equiv \partial/\partial\theta$):

$$\begin{aligned}
\partial_\theta \hat{\boldsymbol{r}} &= \hat{\boldsymbol{\theta}}, && \text{(B.9)}\\
\partial_\phi \hat{\boldsymbol{r}} &= \sin\theta\,\hat{\boldsymbol{\phi}}, && \text{(B.10)}\\
\partial_\theta \hat{\boldsymbol{\theta}} &= -\hat{\boldsymbol{r}}, && \text{(B.11)}\\
\partial_\phi \hat{\boldsymbol{\theta}} &= \cos\theta\,\hat{\boldsymbol{\phi}}, && \text{(B.12)}\\
\partial_\phi \hat{\boldsymbol{\phi}} &= -\sin\theta\,\hat{\boldsymbol{r}} - \cos\theta\,\hat{\boldsymbol{\theta}}. && \text{(B.13)}
\end{aligned}$$

If you are wondering where these come from, picture the idea of rotating vectors used to derive $v = \omega r$ in first year mechanics. Alternatively. you can grind them through by converting to and from Cartesian components; this is left as an exercise. Let us calculate $\nabla\cdot\boldsymbol{W}$ properly:

$$\nabla\cdot\boldsymbol{W} = \hat{\boldsymbol{r}}\cdot\partial_r\boldsymbol{W} + \frac{1}{r}\hat{\boldsymbol{\theta}}\cdot\partial_\theta\boldsymbol{W} + \frac{1}{r\sin\theta}\hat{\boldsymbol{\phi}}\cdot\partial_\phi\boldsymbol{W}. \tag{B.14}$$

None of the unit vectors vary as r changes, so the first term simply ends up as $\partial_r W_r$ as before. The same does not apply to the others however, and applying the relations for the derivatives of the unit vectors from above one obtains:

$$\nabla\cdot\boldsymbol{W} = \partial_r W_r + \frac{1}{r}\left(\partial_\theta W_\theta + W_r\right) + \frac{1}{r\sin\theta}\left(\partial W_\phi + W_r\sin\theta + W_\theta\cos\theta\right). \tag{B.15}$$

Grouping terms by the component involved gives

$$\nabla\cdot\boldsymbol{W} = \left(\partial_r W_r + \frac{2W_r}{r}\right) + \frac{1}{r\sin\theta}\left(\sin\theta\partial_\theta W_\theta + W_\theta\cos\theta\right) + \frac{1}{r\sin\theta}\partial W_\phi. \tag{B.16}$$

From this it is a simple matter to obtain the divergence for spherical polars in "traditional" form

$$\nabla\cdot\boldsymbol{W} = \frac{1}{r^2}\frac{\partial\left(r^2 W_r\right)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(W_\theta\sin\theta\right) + \frac{1}{r\sin\theta}\frac{\partial W_\phi}{\partial\phi}. \tag{B.17}$$

Although it was a lot of work getting there, it can be simpler to calculate a divergence from this expression if a given vector field is best expressed using spherical polars. For example, the Coulomb field has a single radial component with a strength that varies as $1/r^2$. It is clear from Eq. B.17 that $\nabla\cdot\boldsymbol{E} = 0$ for $r \neq 0$. It is significantly more work to show the same using Cartesian expressions.

If a vector field W can be derived from a potential ψ, i.e. $\boldsymbol{W} = \nabla\psi$ then (for spherical polars again):

$$\boldsymbol{W} = \hat{\boldsymbol{r}}\frac{\partial\psi}{\partial r} + \hat{\boldsymbol{\theta}}\frac{1}{r}\frac{\partial\psi}{\partial\theta} + \hat{\boldsymbol{\phi}}\frac{1}{r\sin\theta}\frac{\partial\psi}{\partial\phi}. \tag{B.18}$$

and $\nabla\cdot\boldsymbol{W} = \nabla^2\psi$. We can therefore use the expression for the divergence to deduce that

$$\nabla^2\psi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}. \tag{B.19}$$

This is the form of the Laplacian in spherical polars that you will have encountered when studying Schrodinger's equation applied to the hydrogen atom. It is an important equation for electrostatics (since $\boldsymbol{E}$ is derivable from a potential in that case), and much of electrostatics boils down to solving Laplace's equation, $\nabla^2\psi = 0$, subject to various boundary conditions. For instance it could be used to find the electric potential due to a charged, hollow, hemispherical shell, something that cannot be tackled with Gauss's law in integral form because it lacks spherical symmetry.

B.4 Curl

We finish with curl. In spherical polars we have

$$\nabla \times \boldsymbol{W} = \left(\hat{\boldsymbol{r}} \frac{\partial}{\partial r} + \hat{\boldsymbol{\theta}} \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\boldsymbol{\phi}} \frac{1}{r \sin\theta} \frac{\partial}{\partial \phi} \right) \times \left(W_r \hat{\boldsymbol{r}} + W_\theta \hat{\boldsymbol{\theta}} + W_\phi \hat{\boldsymbol{\phi}} \right). \tag{B.20}$$

This expands to 9 terms, each involving the partial derivative with respect to an ordinate of a component times a unit vector, to which we need to apply the product rule since the unit vectors vary with respect to the ordinates in general. (Thus potentially it expands to 18 terms.) We then need to apply the following rule (and permutations thereof) to simplify the resulting cross products:

$$\hat{\boldsymbol{r}} = \hat{\boldsymbol{\theta}} \times \hat{\boldsymbol{\phi}}. \tag{B.21}$$

The gory details now follow:

$$\begin{aligned} \nabla \times \boldsymbol{W} \;\; = \;\; & \hat{\boldsymbol{r}} \times \frac{\partial}{\partial r} \left(W_r \hat{\boldsymbol{r}} + W_\theta \hat{\boldsymbol{\theta}} + W_\phi \hat{\boldsymbol{\phi}} \right) + \\ & \hat{\boldsymbol{\theta}} \times \frac{1}{r} \frac{\partial}{\partial \theta} \left(W_r \hat{\boldsymbol{r}} + W_\theta \hat{\boldsymbol{\theta}} + W_\phi \hat{\boldsymbol{\phi}} \right) + \\ & \hat{\boldsymbol{\phi}} \times \frac{1}{r \sin\theta} \frac{\partial}{\partial \phi} \left(W_r \hat{\boldsymbol{r}} + W_\theta \hat{\boldsymbol{\theta}} + W_\phi \hat{\boldsymbol{\phi}} \right). \end{aligned} \tag{B.22}$$

Hence applying the differential operators to the bracketed vector expansions along with the product rule and the relations for the derivatives of the unit vectors listed earlier we have

$$\begin{aligned} \nabla \times \boldsymbol{W} \;\; = \;\; & \hat{\boldsymbol{r}} \times \left(\hat{\boldsymbol{r}} \partial_r W_r + \hat{\boldsymbol{\theta}} \partial_r W_\theta + \hat{\boldsymbol{\phi}} \partial_r W_\phi \right) + \\ & \frac{1}{r} \hat{\boldsymbol{\theta}} \times \left(\hat{\boldsymbol{r}} \partial_\theta W_r + \hat{\boldsymbol{\theta}} \partial_\theta W_\theta + \hat{\boldsymbol{\phi}} \partial_\theta W_\phi + \hat{\boldsymbol{\theta}} W_r - \hat{\boldsymbol{r}} W_\theta \right) + \\ & \frac{1}{r \sin\theta} \hat{\boldsymbol{\phi}} \times \left(\hat{\boldsymbol{r}} \partial_\phi W_r + \hat{\boldsymbol{\theta}} \partial_\phi W_\theta + \hat{\boldsymbol{\phi}} \partial_\phi W_\phi + \hat{\boldsymbol{\phi}} W_r \sin\theta + \right. \\ & \qquad\qquad \left. \hat{\boldsymbol{\phi}} W_\theta \cos\theta - \hat{\boldsymbol{r}} W_\phi \sin\theta - \hat{\boldsymbol{\theta}} W_\phi \cos\theta \right). \end{aligned} \tag{B.23}$$

Applying the rules for cross-products of the unit vectors:

$$\begin{aligned} \nabla \times \boldsymbol{W} \;\; = \;\; & \hat{\boldsymbol{\phi}} \partial_r W_\theta - \hat{\boldsymbol{\theta}} \partial_r W_\phi + \frac{1}{r} \left(-\hat{\boldsymbol{\phi}} \partial_\theta W_r + \hat{\boldsymbol{r}} \partial_\theta W_\phi + \hat{\boldsymbol{\phi}} W_\theta \right) + \\ & \frac{1}{r \sin\theta} \left(\hat{\boldsymbol{\theta}} \partial_\phi W_r - \hat{\boldsymbol{r}} \partial_\phi W_\theta - \hat{\boldsymbol{\theta}} W_\phi \sin\theta + \hat{\boldsymbol{r}} W_\phi \cos\theta \right). \end{aligned} \tag{B.24}$$

Collecting terms by unit vector:

$$\begin{aligned} \nabla \times \boldsymbol{W} \;\; = \;\; & \left(\frac{\partial_\theta W_\phi}{r} - \frac{\partial_\phi W_\theta}{r \sin\theta} + \frac{W_\phi \cos\theta}{r \sin\theta} \right) \hat{\boldsymbol{r}} + \\ & \left(-\partial_r W_\phi + \frac{\partial_\phi W_r}{r \sin\theta} - \frac{W_\phi}{r} \right) \hat{\boldsymbol{\theta}} + \end{aligned} \tag{B.25}$$

$$\left(\partial_r W_\theta - \frac{\partial_\theta W_r}{r} + \frac{W_\theta}{r} \right) \hat{\boldsymbol{\phi}}. \tag{B.26}$$

The various terms can be slightly simplified

$$\begin{aligned}\nabla\times\boldsymbol{W} &= \left(\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\, W_\phi\right)-\frac{1}{r\sin\theta}\frac{\partial W_\theta}{\partial\phi}\right)\hat{\boldsymbol{r}}+ \\ &\left(\frac{1}{r\sin\theta}\frac{\partial W_r}{\partial\phi}-\frac{1}{r}\frac{\partial}{\partial r}(rW_\phi)\right)\hat{\boldsymbol{\theta}}+ \\ &\left(\frac{1}{r}\frac{\partial}{\partial r}(rW_\theta)-\frac{1}{r}\frac{\partial W_r}{\partial\theta}\right)\hat{\boldsymbol{\phi}}.\end{aligned} \tag{B.27}$$

These can be written in determinant form as

$$\nabla\times\boldsymbol{W}=\frac{1}{r^2\sin\theta}\begin{vmatrix}\hat{\boldsymbol{r}} & r\,\hat{\boldsymbol{\theta}} & r\sin\theta\,\hat{\boldsymbol{\phi}} \\ \partial_r & \partial_\theta & \partial_\phi \\ W_r & rW_\theta & r\sin\theta\, W_\phi\end{vmatrix} \tag{B.28}$$

You absolutely need not try to remember expressions like this – they would be given if needed!

B.5 Cylindrical coordinates

Corresponding results for cylindrical coordinates (r,θ,z), are as follows

$$\nabla f = \frac{\partial f}{\partial r}\hat{\boldsymbol{r}}+\frac{1}{r}\frac{\partial f}{\partial\theta}\hat{\boldsymbol{\theta}}+\frac{\partial f}{\partial z}\hat{\boldsymbol{z}}, \tag{B.29}$$

$$\nabla\cdot\boldsymbol{W} = \frac{1}{r}\frac{\partial}{\partial r}(rW_r)+\frac{1}{r}\frac{\partial W_\theta}{\partial\theta}+\frac{\partial W_z}{\partial z}, \tag{B.30}$$

$$\nabla^2 f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2}+\frac{\partial^2 f}{\partial z^2}, \tag{B.31}$$

$$\nabla\times\boldsymbol{W} = \frac{1}{r}\begin{vmatrix}\hat{\boldsymbol{r}} & r\hat{\boldsymbol{\theta}} & \hat{\boldsymbol{z}} \\ \partial_r & \partial_\theta & \partial_z \\ W_r & rW_\theta & W_z\end{vmatrix}. \tag{B.32}$$

Appendix C

Electric field of a uniformly polarised sphere

Consider a sphere of radius a which has a uniform polarisation $\boldsymbol{P}$ at all points. This has zero volume charge density since $\nabla \cdot \boldsymbol{P} = 0$, but surface charge density $= P \cos\theta$ where θ is the angle subtended by a given point on the surface relative to the direction of $\boldsymbol{P}$ as measured from the centre of the sphere (see Fig. C.1). One could proceed from here to calculate $\boldsymbol{E}$ using a surface integral over the charges, but an easier approach here is to regard the polarised sphere as the superposition of two spheres of uniform volume charge densities $\pm\rho$ displaced from each other by (small) $\boldsymbol{x}$ such that $\boldsymbol{P} = \rho\boldsymbol{x}$. Since the electric field due to a uniformly charged sphere is a standard result (grows linearly with radius inside the sphere, while it is Coulomb-like outside it), we can work out the field due to a uniformly polarised sphere by superposing the fields due to positively and negatively charged uniform spheres. Considering first the field inside the spheres, a well-known application of gaussian surfaces shows that the field in a uniformly charged sphere is radial and linearly increases in strength with radius. The full result can be written

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{\rho}{3\epsilon_0}(\boldsymbol{r} - \boldsymbol{r}'), \text{ for } |\boldsymbol{r} - \boldsymbol{r}'| < a, \tag{C.1}$$

for the field at $\boldsymbol{r}$ due to a sphere of charge density ρ centred at $\boldsymbol{r}'$. Modelling the polarised sphere by two spheres of charge density $\pm\rho$ centred at $\pm\boldsymbol{x}/2$ (with $x \ll a$), the field due to the polarisation charges at $\boldsymbol{r}$ inside the sphere is given by

$$\boldsymbol{E}_P(\boldsymbol{r}) = \frac{\rho}{3\epsilon_0}(\boldsymbol{r} - \boldsymbol{x}/2) - \frac{\rho}{3\epsilon_0}(\boldsymbol{r} + \boldsymbol{x}/2) = -\frac{\rho\boldsymbol{x}}{3\epsilon_0} = -\frac{\boldsymbol{P}}{3\epsilon_0}. \tag{C.2}$$

Since $\boldsymbol{P}$ is constant, so too is $\boldsymbol{E}_P$. This is a very convenient result as it allows us to calculate the polarisation of a sphere placed into a uniform electric field $\boldsymbol{E}_0$ since then the field internal to the sphere, which is the sum of $\boldsymbol{E}_0$ and $\boldsymbol{E}_P$ will be uniform if $\boldsymbol{P}$ is uniform. Assuming the simple constitutive relation $\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E}$ where χ is called the susceptibility or polarisability, we have

$$\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E} = \epsilon_0\chi(\boldsymbol{E}_0 + \boldsymbol{E}_P). \tag{C.3}$$

But using the result for E_P

$$\boldsymbol{P} = \epsilon_0\chi\left(\boldsymbol{E}_0 - \frac{\boldsymbol{P}}{3\epsilon_0}\right), \tag{C.4}$$

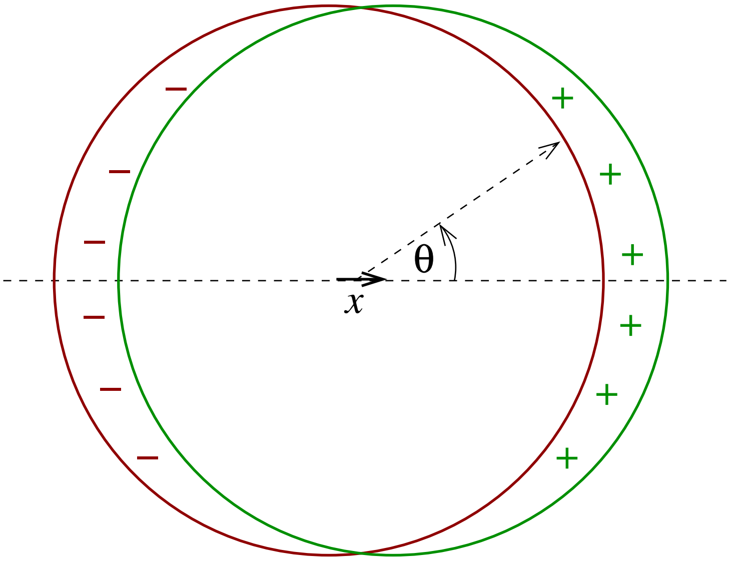

Figure C.1: A uniformly polarised sphere can be imagined as two uniformly-charged spheres of equal but opposite charge, slightly displaced from each other. Polarisation charges appear at the surface of the sphere while its interior is electrically neutral.

hence

$$\boldsymbol{P} = \left(\frac{3\chi}{3+\chi}\right)\epsilon_0 \boldsymbol{E}_0. \tag{C.5}$$

The polarisation in the sphere is thus less than would be obtained by simply applying $\boldsymbol{P} = \epsilon_0\chi\boldsymbol{E}$ with the field $\boldsymbol{E}$ set equal to the applied field $\boldsymbol{E}_0$. With this result, the electric field inside the sphere is given by

$$\boldsymbol{E} = \boldsymbol{E}_0 - \frac{\boldsymbol{P}}{3\epsilon_0} = \frac{3\boldsymbol{E}_0}{3+\chi} = \frac{3}{2+\epsilon_r}\boldsymbol{E}_0. \tag{C.6}$$

Returning to the case of the uniformly polarised sphere, the polarisation charges also generate an electric field *outside* the sphere. Again modelling this as two uniformly but oppositely charged spheres, from the outside each of these acts as a point charge of amount

$$Q = \pm\frac{4\pi}{3}a^3\rho, \tag{C.7}$$

at the centres of the spheres, separated by the small displacement $\boldsymbol{x}$. Therefore outside the sphere, the polarisation charges act as a single dipole at its centre with dipole moment

$$\boldsymbol{p} = Q\boldsymbol{x} = \frac{4\pi}{3}a^3\boldsymbol{P}, \tag{C.8}$$

which equals the volume of the sphere times the polarisation, as might be expected. The electric field outside the sphere is therefore the sum of the applied field $\boldsymbol{E}_0$ and the dipole field from polarisation. It is left as an exercise to show that the field at position x, y relative to the centre of the sphere, with corresponding distance $r > a$ outside the sphere is given by

$$\begin{aligned} E_x &= \left(1 + \frac{a^3\chi(3x^2 - r^2)}{(3+\chi)r^5}\right)E_0, \\ E_y &= \frac{3a^3\chi xy}{(3+\chi)r^5}E_0. \end{aligned}$$

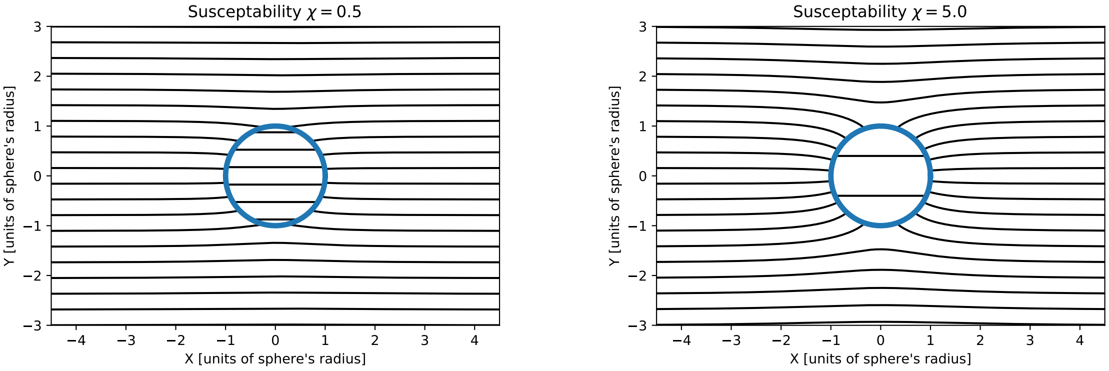

Figure C.2: The electric field of a uniform, isotropic and linear dielectric sphere in a uniform electric field for two values of susceptibility. The number of field lines per unit length along the y-axis is proportional to the field strength. The number of field lines is not conserved on the surface of the sphere owing to polarisation charges, but they are in the bulk medium inside and outside the sphere where there are no polarisation charges.

The field at either "pole" $(x = \pm a,\ y = 0)$ is

$$E_{\mathsf{Pole}} = \frac{3 + 3\chi}{3 + \chi} E_0, \tag{C.9}$$

while at the equator $(x = 0,\ y = \pm a)$

$$E_{\mathsf{Eq}} = \frac{1}{3 + \chi} E_0, \tag{C.10}$$

the field pointing in the x-direction in each case. As $\chi \to \infty$, the field at the equator tends to zero while the field at the poles is $3\times$ stronger than the applied field. Fig. C.2 shows the field patterns near two spheres of different susceptibility placed inside initially uniform electric fields.

Appendix D

Potentials

The material of this appendix would be part of the main text if there were time to cover it as it's an important part of electromagnetism. As remarked in Section 2.3, electrostatic problems are often easier to solve in terms of the electrostatic potential ψ than the electric field $\boldsymbol{E} = -\nabla\psi$. This is because the potential ψ is a scalar field, one number at every point, and also because conductors form equipotentials. Here there are a few more electrostatic potential problems. We also introduce the "vector potential" for magnetic fields.

D.1 Some more electrostatic potential problems

In many cases, charges are confined to conductors, and the solution of electrostatic problems reduces to finding solutions to Laplace's equations compatible with the boundary conditions set by the conductors. The solution for a given set of boundary conditions can be shown to be unique, thus if you find a solution, it is also the solution, and it doesn't matter how you got there. In this section we will go through a series of examples. Some can be solved with more elementary means such as the use of Gaussian surfaces and are useful to know, while others show that we have gained an ability to solve problems lacking the symmetry needed to solve with integral methods.

Potential between two infinite conducting plates
Consider the problem of two parallel conducting plates in a vacuum, parallel to the y–z plane, one located at $x = 0$ and kept at potential $\psi = 0$, the other at $x = d$ and kept at potential $\psi = V$. This is the standard picture of a parallel plate capacitor. Note that conductors allow charges to move and so in electrostatics metals are always at a constant potential otherwise charge would move towards region of low potential until the potential difference had disappeared. In practice, for good conductors such as copper, this happens very quickly. Returning to the problem, we know the answer to be a uniform electric field in the x direction, and thus ψ varies linearly between the plates according to

$$\psi = \frac{V}{d}x. \tag{D.1}$$

This satisfies the boundary conditions since $\psi(x=0)=0$ and $\psi(x=d)=V$. All we need to check to complete the solution is to verify that it also satisfies Laplace's equation

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}=0 \tag{D.2}$$

in between the plates where there are no charges. This is the case for $\psi=(V/d)x$.

Axially symmetric potentials

It is often best to express Laplace's equation and its solutions in terms of a geometry matched to the problem in hand if possible. In this section we look again at the problem of a dielectric sphere in a uniform external field. The sphere suggests using spherical polars. The applied field means that we cannot assume spherical symmetry, but its direction defines an axis which means there will be axial symmetry. The Laplacian in spherical polar coordinates (r,θ,ϕ),

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}=0. \tag{D.3}$$

[NB You do not need to remember this.] Assuming there is axial symmetry around the z-axis, we can drop the ϕ derivative to give

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)=0. \tag{D.4}$$

In the same way that we derived solutions in the 2D Cartesian case, the solutions of this equation can be derived through separation of variables, starting with $\psi(r,\theta)=R(r)\Theta(\theta)$, with the result (which you need not prove or remember as it will be given to you if needed)

$$\psi(r,\theta)=\sum_{n=0}^{\infty}\left(A_n r^n+B_n r^{-(n+1)}\right)P_n(\cos\theta), \tag{D.5}$$

where the P_n are *Legendre polynomials*, which were first introduced by Legendre in the study of gravitational potentials, and A_n and B_n are coefficients to be determined according to the problem in hand. The first few Legendre polynomials are

$$P_0(\cos\theta) = 1, \tag{D.6}$$
$$P_1(\cos\theta) = \cos\theta, \tag{D.7}$$
$$P_2(\cos\theta) = \frac{1}{2}\left(3\cos^2\theta-1\right), \tag{D.8}$$
$$P_3(\cos\theta) = \frac{1}{2}\left(5\cos^3\theta-3\cos\theta\right). \tag{D.9}$$

Returning to the problem of a dielectric sphere in a uniform field, let's orient the field upwards parallel to the z-axis with equivalent potential

$$\psi=-E_0 z=-E_0 r\cos\theta. \tag{D.10}$$

The sphere is assume to be of radius a, centred on the origin and to have dielectric constant (relative permittivity) $\epsilon_r=1+\chi$. This will provide the boundary condition "at infinity" for our problem. The form of this solution exactly matches one of the Legendre polynomials, namely $P_1(\cos\theta)$, hence we can guess that the solution might be of the form

$$\psi(r,\theta)=\left(Ar+Br^{-2}\right)\cos\theta, \tag{D.11}$$

which we now know satisfies Laplace's equation and we have dropped the subscripts on the coefficients. We have to distinguish the regions inside and outside the sphere. In each case the solution will be of the form above, but the coefficients will differ in each case.

Inside the sphere, r can be zero, so we reject the r^{-2} part (the constant $B = 0$), whereas outside the sphere we want the solution to match the imposed external field and hence $A(\text{out}) = -E_0$. Hence our solutions reduce to

$$\begin{aligned} \psi_-(r,\theta) &= Ar\cos\theta,\ \text{for } r < a, && \text{(D.12)} \\ \psi_+(r,\theta) &= \left(-E_0 r + Br^{-2}\right)\cos\theta,\ \text{for } r \geq a, && \text{(D.13)} \end{aligned}$$

with just two undetermined constants, A and B. The potentials must match at the boundary, which is the surface of the sphere $r = a$, hence we must have

$$A = -E_0 + B/a^3. \tag{D.14}$$

We require just one more relation to fully determine the solution. There are no free charges so the perpendicular component of $\boldsymbol{D}$ is the same either side of the boundary, as is the parallel component of $\boldsymbol{E}$. The second condition is automatically satisfied since we have matched potentials and $\boldsymbol{E} = -\nabla\psi$. The first condition involved the dielectric constant and is the extra constraint we need. The perpendicular component in this case corresponds to the radial direction, and together with $\boldsymbol{D} = \epsilon_0\epsilon_r\boldsymbol{E}$, we have

$$\epsilon_r \left.\frac{\partial\psi_-}{\partial r}\right|_{r=a} = \left.\frac{\partial\psi_+}{\partial r}\right|_{r=a}. \tag{D.15}$$

Applying this to the inner and outer solution above leads to

$$\epsilon_r A = -E_0 - 2B/a^3. \tag{D.16}$$

Eqs D.14 and D.16 are a pair of simultaneous equations for A and B which are solved by

$$\begin{aligned} A &= -\frac{3}{\epsilon_r + 2}E_0, && \text{(D.17)} \\ B &= \left(\frac{\epsilon_r - 1}{\epsilon_r + 2}\right)E_0 a^3, && \text{(D.18)} \end{aligned}$$

from which the final solutions for the electric potentials inside and outside the sphere follow

$$\begin{aligned} \psi_-(r,\theta) &= -\frac{3}{\epsilon_r + 2}E_0 r\cos\theta,\ \text{for } r < a, && \text{(D.19)} \\ \psi_+(r,\theta) &= \left(-E_0 r + \left(\frac{\epsilon_r - 1}{\epsilon_r + 2}\right)E_0\frac{a^3}{r^2}\right)\cos\theta,\ \text{for } r \geq a. && \text{(D.20)} \end{aligned}$$

The inner solution is the potential of a uniform field of strength

$$E = \frac{2}{\epsilon_r + 2}E_0, \tag{D.21}$$

the same as we had earlier (Eq. C.6). The outer solution is the superposition of two parts, the potential from the uniform external field and a dipole potential of dipole moment

$$\boldsymbol{p} = 4\pi a^3\left(\frac{\epsilon_r - 1}{\epsilon_r + 2}\right)\epsilon_0 E_0\,\hat{\boldsymbol{z}}, \tag{D.22}$$

from comparing with Eq. 1.30. Given the uniform internal field of the sphere, it must have uniform polarisation of strength

$$\boldsymbol{P} = \frac{\boldsymbol{p}}{4\pi a^3/3} = \left(\frac{3\epsilon_r - 3}{\epsilon_r + 2}\right) \epsilon_0 E_0 \, \hat{\boldsymbol{z}}, \tag{D.23}$$

which is seen to be the same as Eq. C.5.

The potential approach hasn't told us anything new in this case, but it is more generalisable than the more intuitive approach adopted in the Appendix C, and it is a good example of the application of boundary conditions to match solutions applicable in different regions.

D.2 The magnetic vector potential

In general there is no scalar field equivalent to the electric potential of electrostatics for magnetism, even for the case of magnetostatics. The magnetostatic equations are

$$\begin{aligned} \nabla \cdot \boldsymbol{B} &= 0, \\ \nabla \times \boldsymbol{B} &= \mu_0 \boldsymbol{J}. \end{aligned}$$

Unless $\boldsymbol{J} = 0$, it is clear that $\nabla \times \boldsymbol{B} \neq \boldsymbol{0}$. Hence $\boldsymbol{B}$ cannot be written as $-\nabla \psi_M$ for any scalar potential ψ_M. Instead, if we want to retain generality, the first relation allows us to write that $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, where $\boldsymbol{A}$ is some vector field known as the "magnetic vector potential". We are making use of the general relation $\nabla \cdot \nabla \times \boldsymbol{A} = 0$ for any vector field $\boldsymbol{A}$, to satisfy Maxwell's equation $\nabla \boldsymbol{B} = 0$ automatically.

Just as the static electric field $\boldsymbol{E}$ is invariant to adding a constant to ψ, so $\boldsymbol{B}$ is invariant to adding the gradient of a scalar field ζ to $\boldsymbol{A}$:

$$\boldsymbol{A} \rightarrow \boldsymbol{A} + \nabla \zeta, \tag{D.24}$$

to the vector potential since

$$\nabla \times (\boldsymbol{A} + \nabla \zeta) = \nabla \times \boldsymbol{A} + \nabla \times \nabla \zeta = \nabla \times \boldsymbol{A}. \tag{D.25}$$

We will now use this "gauge freedom" to simplify the equation relating $\boldsymbol{A}$ to $\boldsymbol{J}$. Substituting $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ into the Maxwell-Ampère relation $\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J}$ gives

$$\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu_0 \boldsymbol{J}, \tag{D.26}$$

where we have used a standard vector derivative relation to obtain the terms in the central expression. (The word "gauge" was introduced by Weyl in another context. He was trying to formulate General Relativity as a theory which was invariant to local changes in gauge or units of length. Einstein was impressed with the idea but soon showed that it led to unphysical results. When the ideas were taken over to quantum theories, Weyl found that invariance under the local changes of phase - rather than length - was equivalent to charge conservation.)

We can set $\nabla \cdot \boldsymbol{A} = 0$ which we are free to do because of the gauge freedom. If some initial vector potential $\boldsymbol{A}'$ did not satisfy this condition then we could choose a ζ such that

$$\nabla \cdot \boldsymbol{A} = \nabla \cdot (\boldsymbol{A}' + \nabla \zeta) = 0, \tag{D.27}$$

i.e. ζ such that $\nabla^2\zeta = -\nabla \cdot \boldsymbol{A}'$, without affecting $\boldsymbol{B}$. With this choice of gauge we deduce

$$\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}, \tag{D.28}$$

which is Poisson's equation for the magnetic vector potential. By analogy with the electrostatic case we can write

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int \frac{\boldsymbol{J}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}dV'. \tag{D.29}$$

This result means that, in the case of a straight wire ($\boldsymbol{J}$ uni-directional), the vector potential field due to runs parallel to the wire.

In practice the vector potential is not as useful for solving particular problems as is the electrostatic potential. First it is a vector, so the above equations have three components, second one has to take the curl rather than the gradient, third there are not many soluble examples of the vector potential, and fourth there are no simple boundary conditions on the vector potential equivalent to the electrostatic case with conductors. Nevertheless, it plays an important role in understanding the radiation from moving charges, in relativistic treatments of electrodynamics, and in incorporating electromagnetism into quantum physics. It is also of interest since Maxwell first expressed his equations using it, and only later was it shown not to be needed directly in classical physics. It is of little use to us in PX263, except for discussing current loops (see below, and in a later appendix where we derive the magnetic vector potential far from a current loop which is later used to examine magnetisation).

The vector potential far from a current loop
Specialising to a current in a wire, a small section of wire of vector length $\boldsymbol{dl}$ and cross-sectional area α carrying current I has mean current density $J = I/\alpha$ and volume $dV = \alpha\, d\ell$. The vector $d\boldsymbol{\ell}$ is parallel to $\boldsymbol{J}$ and the element $\boldsymbol{J}\, dV'$ in the vector potential integral Eq. D.29 can be written as $(I/\alpha)\alpha\, d\boldsymbol{\ell} = I\, d\boldsymbol{\ell}$. The vector potential at $\boldsymbol{r}$ due to a (closed) current loop is then given by

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0 I}{4\pi}\oint \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{\ell}. \tag{D.30}$$

Here $\boldsymbol{r}'$ is the position of the line element $d\boldsymbol{\ell}$.

We can reference the line elements in D.30 from a point close to the loop $\boldsymbol{r}_0$ so that $\boldsymbol{r}' = \boldsymbol{r}_0 + \boldsymbol{\epsilon}$. If we consider a position $\boldsymbol{r}$ far from the loop so that $|\boldsymbol{r}-\boldsymbol{r}_0| \gg |\boldsymbol{\epsilon}|$ for all $\boldsymbol{\epsilon}$ around the loop, we can approximate the integrand which can be considered to be a scalar function of $\boldsymbol{r}'$, $|\boldsymbol{r}-\boldsymbol{r}'|^{-1} = f(\boldsymbol{r}')$, using

$$f(\boldsymbol{r}') = f(\boldsymbol{r}_0 + \boldsymbol{\epsilon}) = f(\boldsymbol{r}_0) + \nabla f(\boldsymbol{r}_0)\cdot\boldsymbol{\epsilon} + \ldots. \tag{D.31}$$

We won't need any higher order terms.

The first term in D.31 is a constant which drops out when substituted into the line integral for the vector potential since one ends with $\oint d\boldsymbol{\ell}$ which must be zero. We are left with

$$\boldsymbol{A}(\boldsymbol{r}) \approx \frac{\mu_0 I}{4\pi}\oint \nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)_{\boldsymbol{r}_0}\cdot\boldsymbol{\epsilon}\, d\boldsymbol{\epsilon} = \frac{\mu_0 I}{4\pi}\oint \frac{(\boldsymbol{r}-\boldsymbol{r}_0)}{|\boldsymbol{r}-\boldsymbol{r}_0|^3}\cdot\boldsymbol{\epsilon}\, d\boldsymbol{\epsilon} \tag{D.32}$$

$$\equiv \oint(\boldsymbol{g}\cdot\boldsymbol{\epsilon})\, d\boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{g} = \frac{\mu_0 I}{4\pi|\boldsymbol{r}-\boldsymbol{r}_0|^3}(\boldsymbol{r}-\boldsymbol{r}_0). \tag{D.33}$$

In the integral in D.33, the vector, $\boldsymbol{g}$, is constant. (The dash on the gradient operator signifies that it acts on $\boldsymbol{r}'$, not $\boldsymbol{r}$, and the subscript shows that it takes whatever value it has at $\boldsymbol{r}' = \boldsymbol{r}_0$, i.e. it is constant.) The approximation holds when far from the loop.

Using the relation for the vector triple-product (Eq. 1.16), we can write

$$\boldsymbol{g} \times (\boldsymbol{\epsilon} \times d\boldsymbol{\epsilon}) = (\boldsymbol{g} \cdot d\boldsymbol{\epsilon})\boldsymbol{\epsilon} - (\boldsymbol{g} \cdot \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon}, \tag{D.34}$$

and so

$$(\boldsymbol{g} \cdot \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon} = (\boldsymbol{g} \cdot d\boldsymbol{\epsilon})\boldsymbol{\epsilon} - \boldsymbol{g} \times (\boldsymbol{\epsilon} \times d\boldsymbol{\epsilon}). \tag{D.35}$$

The first term on the right-hand side can be in turn transformed using

$$d\,[(\boldsymbol{g} \cdot \boldsymbol{\epsilon})\boldsymbol{\epsilon}] = (\boldsymbol{g} \cdot d\boldsymbol{\epsilon})\boldsymbol{\epsilon} + (\boldsymbol{g} \cdot \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon}, \tag{D.36}$$

and therefore

$$2(\boldsymbol{g} \cdot \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon} = d\,[(\boldsymbol{g} \cdot \boldsymbol{\epsilon})\boldsymbol{\epsilon}] - \boldsymbol{g} \times (\boldsymbol{\epsilon} \times d\boldsymbol{\epsilon}). \tag{D.37}$$

Integrating around the loop we deduce that

$$2 \oint (\boldsymbol{g} \cdot \boldsymbol{\epsilon})\, d\boldsymbol{\epsilon} = -\boldsymbol{g} \times \oint \boldsymbol{\epsilon} \times d\boldsymbol{\epsilon}. \tag{D.38}$$

For a flat loop

$$\frac{1}{2} \oint \boldsymbol{\epsilon} \times d\boldsymbol{\epsilon} \tag{D.39}$$

gives the (vector) area of the loop and multiplying by I, we recognise the magnetic dipole moment of the loop. In general we define the magnetic moment of a current loop by

$$\boldsymbol{m} = \frac{I}{2} \oint \boldsymbol{\epsilon} \times d\boldsymbol{\epsilon}, \tag{D.40}$$

a constant vector. Swapping the order of the vector product to absorb the minus sign, inserting the expression for $\boldsymbol{g}$ from D.33, and identifying $\boldsymbol{m}$, gives

$$\boldsymbol{A}(\boldsymbol{r}) \approx \frac{\mu_0}{4\pi} \frac{\boldsymbol{m} \times (\boldsymbol{r} - \boldsymbol{r}_0)}{|\boldsymbol{r} - \boldsymbol{r}_0|^3}. \tag{D.41}$$

This is the magnetic vector potential at $\boldsymbol{r}$ at large distance from a current loop at $\boldsymbol{r}_0$. If you compare this expression with the Biot-Savart law for current elements (Eq. 1.37), you will see that, like the magnetic field lines due to a small current element, the vector potential due to a small loop forms circles in planes that lie perpendicular to $\boldsymbol{m}$.They are parallel to the loop if the latter is flat. This result is used in appendix F to analyse magnetisation currents.

Appendix E

EM and Hamiltonian mechanics

This part is way off syllabus and is just to give a sense of how EM is included in quantum mechanics. You will need to have studied some Hamiltonian mechanics to understand it.

In classical theory, the potentials are helpful but not absolutely necessary: you can use just the fields. They really come to the fore in quantum theory. To see roughly why, let's first set $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ in Faraday's law:

$$\nabla \times \boldsymbol{E} = -\frac{\partial}{\partial t} \nabla \times \boldsymbol{A}, \tag{E.1}$$

then reversing the ∂_t and ∇ on the right-hand side and rearranging shows that

$$\nabla \times \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} \right) = \boldsymbol{0}. \tag{E.2}$$

Recognising a conservative field, we can say that there is a ψ such that

$$\boldsymbol{E} = -\nabla \psi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{E.3}$$

This is the time-variable generalisation of the usual electrostatic relation $\boldsymbol{E} = -\nabla \psi$. With this we can write the Lorentz force in terms of potentials:

$$\boldsymbol{F} = q \left(-\nabla \psi - \frac{\partial \boldsymbol{A}}{\partial t} - \boldsymbol{v} \times (\nabla \times \boldsymbol{A}) \right). \tag{E.4}$$

The final term can be rewritten using an identity:

$$\boldsymbol{v} \times (\nabla \times \boldsymbol{A}) = \nabla(\boldsymbol{A} \cdot \boldsymbol{v}) - (\boldsymbol{v} \cdot \nabla)\boldsymbol{A}. \tag{E.5}$$

This gives

$$\boldsymbol{F} = q \left(-\nabla \psi + \nabla(\boldsymbol{A} \cdot \boldsymbol{v}) - \left[\frac{\partial \boldsymbol{A}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{A} \right] \right). \tag{E.6}$$

The terms in square brackets in E.6 are the *total* derivative of $\boldsymbol{A}$ along the track of the particle. There are two contributions. One comes from the changing of $\boldsymbol{A}$ at a particular location, the other from the amount it changes owing to the particle's motion. We can therefore write

$$\boldsymbol{F} = q \left(-\nabla(\psi - \boldsymbol{A} \cdot \boldsymbol{v}) - \frac{d\boldsymbol{A}}{dt} \right). \tag{E.7}$$

Remembering that $\boldsymbol{F} = d\boldsymbol{p}/dt$ where p is the momentum,

$$\frac{d}{dt}(\boldsymbol{p} + q\boldsymbol{A}) = -q\nabla(\psi - \boldsymbol{A}\cdot\boldsymbol{v}). \tag{E.8}$$

These equations of motion can be derived using Euler-Lagrange equations from the Lagrangian

$$L = \frac{p^2}{2m} + q\boldsymbol{A}\cdot\boldsymbol{v} - q\psi. \tag{E.9}$$

The quantity $\boldsymbol{p'} = \boldsymbol{p} + q\boldsymbol{A}$ emerges from this as a momentum conjugate to the position vector. This is known as the "canonical" momentum and appears in the equivalent Hamiltonian

$$H = \frac{1}{2m}\left(\boldsymbol{p'} - q\boldsymbol{A}\right)^2 + q\psi, \tag{E.10}$$

which is the starting point for including EM in (non-relativistic) QM.

Appendix F

Polarisation and magnetisation again

In chapter 4 we gave pictorial justifications for treating a distribution of electric polarisation as equivalent to a set of charges of surface density $\sigma = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$ and volume density $\rho = -\nabla\cdot\boldsymbol{P}$ and a distribution of magnetisation as equivalent to currents of volume density $\nabla\times\boldsymbol{M}$ and surface density $\boldsymbol{M}\times\hat{\boldsymbol{n}}$. There is an alternative justification based upon manipulation of integrals which some may be happier with. We look at this here, beginning with electric polarisation. As usual, it's beyond what you need to know for PX263 and is really for those who want a bit more background.

Polarisation
We know that the electrostatic potential at position $\boldsymbol{r}$ due to a dipole $\boldsymbol{p}$ located at $\boldsymbol{r}'$ is given by Eq. 1.29:

$$\psi(\boldsymbol{r}) = \frac{\boldsymbol{p}\cdot(\boldsymbol{r}-\boldsymbol{r}')}{4\pi\epsilon_0|\boldsymbol{r}-\boldsymbol{r}'|^3}, \tag{F.1}$$

and that the dipole moment of a small volume of polarised material is (by definition) $\boldsymbol{P}(\boldsymbol{r}')\,dV'$. Hence the electrostatic potential due to a polarised object is

$$\psi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\int_V \frac{\boldsymbol{P}(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3}\,dV'. \tag{F.2}$$

The prime on dV' indicates that the element of volume is at position $\boldsymbol{r}'$. The position vector dependent term in the integrand can be written using

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

with the prime on the "del" now indicates that we are taking partial derivatives with respect to the x', y' and z' of $\boldsymbol{r}'$ rather than the x, y and z of $\boldsymbol{r}$.

Consider the following general product-rule for an arbitrary vector field $\boldsymbol{A}$ and scalar field f:

$$\nabla\cdot(f\boldsymbol{A}) = f\nabla\cdot\boldsymbol{A} + \boldsymbol{A}\cdot\nabla f. \tag{F.4}$$

Setting $f = 1/|\boldsymbol{r}-\boldsymbol{r}'|$ and $\boldsymbol{A} = \boldsymbol{P}$, and remembering that here $\boldsymbol{r}'$ is the position of interest, we can re-write

$$\frac{\boldsymbol{P}(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3} = \boldsymbol{P}(\boldsymbol{r}')\cdot\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) = \nabla'\cdot\left(\frac{\boldsymbol{P}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\right) - \frac{\nabla'\cdot\boldsymbol{P}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}. \tag{F.5}$$

Substituting this into Eq. F.2, and transforming the first of the resulting two integrals from a volume integral into a surface integral using Gauss' theorem, gives

$$\psi(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\oint_S \frac{\boldsymbol{P}(\boldsymbol{r}')\cdot d\boldsymbol{S}'}{|\boldsymbol{r}-\boldsymbol{r}'|} - \frac{1}{4\pi\epsilon_0}\int_V \frac{\nabla'\cdot\boldsymbol{P}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,dV'. \tag{F.6}$$

This is the final result. Each term in this expression contains an integral of a Coulomb $1/r$-potential where $r = |\boldsymbol{r}-\boldsymbol{r}'|$ is the distance from the point of integration $\boldsymbol{r}'$ to the point $\boldsymbol{r}$ where the potential is calculated. Playing the role of "q" in each case is $\boldsymbol{P}(\boldsymbol{r}')\cdot d\boldsymbol{S}'$ in the first term and $-\nabla'\cdot\boldsymbol{P}(\boldsymbol{r}')\,dV'$ in the second term. We deduce that the electric potential due to a polarised object can be modelled with surface and volume charge distributions with densities $\sigma = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$ and $\rho = -\nabla\cdot P$ (see 4.2 and 4.4).

The magnetisation current formulae

In Section 4.2 formulae were quoted for the current densities equivalent to distributions of magnetic dipoles as measured by a magnetic dipole moment per unit volume, called the magnetisation. We now work through a more formal derivation of these relations, analogous to the approach used to derive the polarisation charge formulae above. We use the magnetic vector potential described in appendix D. Read that appendix first before tackling this section.

Imagine breaking up the material into small volume elements which at position $\boldsymbol{r}'$ will have dipole moment $d\boldsymbol{m} = \boldsymbol{M}(\boldsymbol{r}')\,dV'$, then using Eq. D.41 and integrating over the volume of the magnetised object in question, the vector potential at $\boldsymbol{r}$ can be written

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{M}(\boldsymbol{r}')\times(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3}dV' = \frac{\mu_0}{4\pi}\int_V \boldsymbol{M}(\boldsymbol{r}')\times\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)dV'. \tag{F.7}$$

In the second form, the position vector term has been re-written as a gradient of a scalar field. The integrand can then be re-written using the product-rule in the form of Eq. 1.19, and one finds

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int_V \frac{\nabla'\times\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}dV' - \frac{\mu_0}{4\pi}\int_V \nabla'\times\left(\frac{\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)dV'. \tag{F.8}$$

Comparing with Eq. D.29, the first term can be interpreted as the vector potential due to currents of volume current density $\boldsymbol{J} = \nabla\times\boldsymbol{M}$, just as outlined in section 4.2. The second term can be transformed into a surface integral with a generalised version of Gauss's theorem which shows that

$$\int_V \nabla'\times\left(\frac{\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)dV' = \oint_S \hat{\boldsymbol{n}}\times\left(\frac{\boldsymbol{M}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)dS = -\oint_S \frac{\boldsymbol{M}(\boldsymbol{r}')\times\hat{\boldsymbol{n}}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,dS. \tag{F.9}$$

Here we have used $d\boldsymbol{S} = dS\hat{\boldsymbol{n}}$.

The final term in the vector potential then becomes

$$\frac{\mu_0}{4\pi}\oint \frac{\boldsymbol{M}(\boldsymbol{r}')\times\hat{\boldsymbol{n}}}{|\boldsymbol{r}-\boldsymbol{r}'|}\,dS, \tag{F.10}$$

which can be interpreted as the vector potential due to surface currents of density $\boldsymbol{j} = \boldsymbol{M}\times\hat{\boldsymbol{n}}$ at any point on the surface where the magnetisation is $\boldsymbol{M}$ and the outward-pointing normal is $\hat{\boldsymbol{n}}$. This completes the more formal derivation of the results of section 4.2, although you may regard it more as a useful exercise in the manipulation of the integrals typical of electromagnetism than offering greater insight than the presentation of section 4.2.

Appendix G

Non-isotropic media, birefringence

Material here is for interest only.

There are materials that exhibit more than one refractive index. A beam of light hitting a block of such material may split into two beams, which are perpendicularly polarised with respect to each other, one of which does not satisfy Snell's law and is known as the "extraordinary ray". This is the interesting phenomenon of *birefringence*.

The key to understanding birefringence is an underlying anisotropy in the substance. Although the relationship between $\boldsymbol{D}$ and $\boldsymbol{E}$ may remain linear to a good approximation, it does not always have the simple form $\boldsymbol{D} = \epsilon\boldsymbol{E}$ with ϵ a single number. Instead in general it is a *tensor* relation which can be expressed as

$$\boldsymbol{D} = \boldsymbol{\epsilon}\boldsymbol{E}, \tag{G.1}$$

where ϵ is now a 3x3 matrix (in more detail it contains the components of a second order tensor that can conveniently be represented by a matrix). The matrix can be shown to be symmetric, and it is always possible to find Cartesian axes which reduce it to diagonal form (the "principal axes"). Using such axes, the matrix reduces to

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} \tag{G.2}$$

and the case of interest is when the diagonal elements are not all equal. In this case $\boldsymbol{D}$ is not in general parallel to $\boldsymbol{E}$ unless it lies along one of the principal axes. Materials then split into three classes:

1. *Isotropic*: $\epsilon_1 = \epsilon_2 = \epsilon_3$. This is just the case studied in the main text where we could assume that $\boldsymbol{D}$ and $\boldsymbol{E}$ were parallel. Crystals of sufficient symmetry (e.g. salt, NaCl) and non-crystalline materials such as glasses fall into this category.

2. *Uniaxial*: $\epsilon_1 = \epsilon_2 \neq \epsilon_3$. These exhibit a polarisation dependent-effect called *birefringence* (see below), but they have a specific direction or axis along which they do not show this effect.

3. *Biaxial*: $\epsilon_1 \neq \epsilon_2 \neq \epsilon_3$. These again show birefringence but have two axes along which they show no signs of it.

The relations derived from Maxwell's equations are unchanged

$$\boldsymbol{k}\cdot\boldsymbol{D}=0, \tag{G.3}$$
$$\boldsymbol{k}\cdot\boldsymbol{B}=0, \tag{G.4}$$
$$\boldsymbol{k}\times\boldsymbol{E}=\omega\boldsymbol{B}, \tag{G.5}$$
$$\boldsymbol{k}\times\boldsymbol{H}=-\omega\boldsymbol{D}, \tag{G.6}$$

but we now cannot assume that $\boldsymbol{E}$ and $\boldsymbol{D}$ are parallel, so that, for example, $\boldsymbol{k}\cdot\boldsymbol{D}=0$ does not imply that $\boldsymbol{k}\cdot\boldsymbol{E}=0$.

Taking the cross-product of the third relation with $\boldsymbol{k}$ and combining it with the fourth gives

$$k^2\boldsymbol{E}-(\boldsymbol{k}\cdot\boldsymbol{E})\boldsymbol{k}=\mu\omega^2\boldsymbol{\epsilon}\boldsymbol{E}. \tag{G.7}$$

This can be written in the following form

$$\left(k^2\mathbf{I}-\mathbf{K}-\mu\omega^2\boldsymbol{\epsilon}\right)\boldsymbol{E}=0, \tag{G.8}$$

where $\mathbf{I}$ is the identity matrix and the $\mathbf{K}$ stands for the matrix whose ij element is given by $K_{ij}=k_ik_j$ (this is sometimes termed a "dyad", which is a second order tensor generated from the dyadic product of two vectors). For this to have non-zero solutions for $\boldsymbol{E}$, the matrix in brackets must be singular (have zero determinant) which leads to:

$$\begin{vmatrix} (k_0^2n_x^2-k_y^2-k_z^2) & k_xk_y & k_xk_z \\ k_yk_x & (k_0^2n_y^2-k_z^2-k_x^2) & k_yk_z \\ k_zk_x & k_zk_y & (k_0^2n_z^2-k_x^2-k_y^2) \end{vmatrix}=0, \tag{G.9}$$

where $k_0=\omega/c$ (equal to the wavenumber in a vacuum), and the material has been assumed to be non-magnetic so that ϵ can be replaced by the square of the refractive index.

Expanding the determinant and re-arranging terms according to powers of k_0 one finds

$$k_0^4-\left(\frac{k_x^2+k_y^2}{n_z^2}+\frac{k_y^2+k_z^2}{n_y^2}+\frac{k_z^2+k_x^2}{n_x^2}\right)k_0^2+\left(\frac{k_x^2}{n_y^2n_z^2}+\frac{k_y^2}{n_z^2n_x^2}+\frac{k_z^2}{n_x^2n_y^2}\right)k^2=0, \tag{G.10}$$

with $k^2=k_x^2+k_y^2+k_z^2$. Remembering that $k_0=\omega/c$, this is an anisotropic dispersion relation for the EM waves in an anisotropic medium. In a uniaxial material (we won't look at the more complex case of biaxial crystals), if we set $n_x=n_y=n_o$ and $n_z=n_e$ (z becomes what is called the "optic axis"), this expression can be factored:

$$\left(\frac{k_x^2}{n_o^2}+\frac{k_y^2}{n_o^2}+\frac{k_z^2}{n_o^2}-\frac{\omega^2}{c^2}\right)\left(\frac{k_x^2}{n_e^2}+\frac{k_y^2}{n_e^2}+\frac{k_z^2}{n_o^2}-\frac{\omega^2}{c^2}\right)=0. \tag{G.11}$$

The first factor defines a spherical surface of allowed values of $\boldsymbol{k}$ for a given ω of radius $k=n_o\omega/c$. The medium then appears simply to have refractive index n_o in all directions. This corresponds to what is known as the "ordinary ray". The other factor defines an ellipsoid squashed or stretched in the z-direction relative to x and y. This corresponds to the "extraordinary ray" for which the effective refractive index depends upon the direction of $\boldsymbol{k}$. The condition of Eq. G.8 constrains the possible electric field vector as well as requiring singularity of the matrix. If the condition for the ordinary ray is satisfied, i.e.

$$k_0^2n_o^2=k_x^2+k_y^2+k_z^2, \tag{G.12}$$

then the left-hand matrix becomes

$$\begin{pmatrix} k_x^2 & k_xk_y & k_xk_z \\ k_yk_x & k_y^2 & k_yk_z \\ k_zk_x & k_zk_y & k_0^2n_e^2 - k_x^2 - k_y^2 \end{pmatrix}. \tag{G.13}$$

Multiplying this into the electric vector gives written as

$$k_x(k_xE_x + k_yE_y) + k_xk_zE_z = 0, \tag{G.14}$$

$$k_y(k_xE_x + k_yE_y) + k_yk_zE_z = 0, \tag{G.15}$$

$$k_z(k_xE_x + k_yE_y) + (k_0^2n_e^2 - k_x^2 - k_y^2)E_z = 0. \tag{G.16}$$

Dividing the first of these by k_x, the second by k_y and the third by k_z

$$(k_xE_x + k_yE_y) + k_zE_z = 0, \tag{G.17}$$

$$(k_xE_x + k_yE_y) + k_zE_z = 0, \tag{G.18}$$

$$(k_xE_x + k_yE_y) + \frac{k_0^2n_e^2 - k_x^2 - k_y^2}{k_z}E_z = 0. \tag{G.19}$$

Since in the last of these relations the fraction does not in general equal k_z while the first term has the same value in each case, we need $E_z = 0$ and $k_xE_x + k_yE_y = 0$ (which together imply that $\boldsymbol{k} \cdot \boldsymbol{E} = 0$ for the ordinary ray). The electric vector is perpendicular to z and must lie in the x–y plane. Since the x- and y-directions are equivalent in uniaxial materials, we are free to rotate them, and thus define the x-axis by the electric field vector. With this definition, $k_x = 0$ while $\boldsymbol{E} = (E, 0, 0)$, which we will use in the next section to simplify the relations for the extraordinary ray.

If the case of the extraordinary ray is satisfied, setting $k_x = 0$, the left-hand matrix becomes

$$\begin{pmatrix} (k_0^2n_o^2 - k_y^2 - k_z^2) & 0 & 0 \\ 0 & (k_0^2n_o^2 - k_z^2) & k_yk_z \\ 0 & k_zk_y & (n_e/n_o)^2k_z^2 \end{pmatrix}. \tag{G.20}$$

The top-left term is non-zero, so $E_x = 0$, and therefore the electric vector of the extraordinary ray is perpendicular to the electric vector for the ordinary ray (but note that $\boldsymbol{k} \cdot \boldsymbol{E} \neq 0$ for the extraordinary ray). It follow that the ordinary and extraordinary rays have orthogonal polarisations, an important property of birefringence.

These considerations explain several of the basic properties of birefringence described at the start of this appendix, in particular in relation to the orthogonal polarisations of the ordinary and extraordinary rays. The most well-known example of this is calcite, CaCO_3 (aka "Iceland spar"), a uniaxial crystal with $n_0 = 1.658$ and $n_e = 1.486$, an unusually large difference. If you place a piece of calcite over a page of text you will see two images of the letters. There have been suggestions that the Vikings may have used calcite to determine the location of the Sun from the polarisation of the sky even after sunset or after it had gone behind clouds on the horizon. Experiments have shown that it is possible to locate the Sun to within a few degrees with this method (Ropars et al, Proc. Royal Soc. A (2012), 468, 671). Calcite plates and prisms are often used in optical instruments where it is important to separate the two components of polarisation.

One of the most remarkable features of birefringence is that the energy of the extraordinary ray is not propagated perpendicular to its wavefronts. As a consequence, when light enters calcite perpendicular to its face, and thus the wavefronts are parallel to the face, the extraordinary ray is deflected, i.e. it does not appear to obey Snell's law. This leads to the doubling of images referred to above.

The anisotropy that underlies birefringence can be induced in otherwise non-birefringent materials through stress. Viewing transparent materials through pairs of crossed polaroids can reveal this through coloured bands. Transparent plastic rulers and protractors are good test objects to try out. It is also sometimes visible from car windscreens on cloudless days owing to the strong polarisation of blue sky, and the stresses frozen into them during manufacture.

Appendix H

Momentum in EM fields

We justified the existence of radiation pressure by dividing the Poynting vector by c using the relation $E = pc$ from relativity, see Section 5.3. There is a way to get there from Maxwell's equations. We will stick to the vacuum equations in this case because the question of momentum in EM in the presence of matter is complex. It's a little involved even in the vacuum case, hence its appearance in an appendix. You don't need to know this.

We begin from the Lorentz force $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$, from which the total EM force acting upon a distribution of charges and currents is

$$\boldsymbol{F} = \int_V (\rho \boldsymbol{E} + \boldsymbol{J} \times \boldsymbol{B})\, dV. \tag{H.1}$$

Using Maxwell's equations to replace the charge and current terms, the integrand on the right becomes

$$\epsilon_0 (\nabla \cdot \boldsymbol{E})\boldsymbol{E} + \frac{1}{\mu_0}(\nabla \times \boldsymbol{B}) \times \boldsymbol{B} - \epsilon_o (\partial_t \boldsymbol{E}) \times \boldsymbol{B} \tag{H.2}$$

Writing

$$-(\partial_t \boldsymbol{E}) \times \boldsymbol{B} = \boldsymbol{B} \times (\partial_t \boldsymbol{E}) = \partial_t (\boldsymbol{B} \times \boldsymbol{E}) - (\partial_t \boldsymbol{B}) \times \boldsymbol{E}, \tag{H.3}$$

using the product rule, and Faraday's law we obtain

$$\epsilon_0 \boldsymbol{E}(\nabla \cdot \boldsymbol{E}) + \frac{1}{\mu_0}\boldsymbol{B}(\nabla \cdot \boldsymbol{B}) + \frac{1}{\mu_0}(\nabla \times \boldsymbol{B}) \times \boldsymbol{B} + \epsilon_o \left[\partial_t (\boldsymbol{B} \times \boldsymbol{E}) + (\nabla \times \boldsymbol{E}) \times \boldsymbol{E}\right]. \tag{H.4}$$

A term $\boldsymbol{B}(\nabla \cdot \boldsymbol{B})/\mu_0$ has been added for symmetry reasons (this is possible because $\nabla \cdot \boldsymbol{B} = 0$).

Returning to the volume integral over a static region, recognising the force as the rate of change of mechanical momentum $\boldsymbol{p}_{\text{mech}}$, and taking the time derivative term over to the right-hand side, we may write

$$\begin{aligned} \frac{d\boldsymbol{p}_{\text{mech}}}{dt} + \epsilon_0 \frac{d}{dt}\int_V \boldsymbol{E} \times \boldsymbol{B}\, dV &= \int_V \Bigg[\epsilon_0 \boldsymbol{E}(\nabla \cdot \boldsymbol{E}) + \frac{1}{\mu_0}\boldsymbol{B}(\nabla \cdot \boldsymbol{B}) + \\ &\qquad \frac{1}{\mu_0}(\nabla \times \boldsymbol{B}) \times \boldsymbol{B} + \epsilon_o (\nabla \times \boldsymbol{E}) \times \boldsymbol{E}\Bigg]\, dV. \end{aligned} \tag{H.5}$$

Writing $\boldsymbol{B} = \mu_0 \boldsymbol{H}$, and recognising $\mu_0 \epsilon_0 = 1/c^2$, the second term on the left takes the form

$$\frac{d}{dt}\int_V \frac{\boldsymbol{E} \times \boldsymbol{H}}{c^2}\, dV. \tag{H.6}$$

If we suppose that the momentum in the EM field is represented by

$$\boldsymbol{p}_{\mathrm{EM}} = \int_V \frac{\boldsymbol{E} \times \boldsymbol{H}}{c^2}\, dV, \tag{H.7}$$

then the quantity in the integrand is the momentum per unit volume or momentum density.

This is a motivational derviation rather than a proof; the final steps are to show that the term on the right hand side can be expressed as a divergence because then by Gauss' theorem it can be transformed into a surface integral representing the stress (force per unit area) acting upon the volume in question. The quantity involved is a tensor, the Maxwell stress tensor, the ij-th component of which can be written

$$T_{ij} = D_i E_j + B_i H_j - \frac{1}{2} \left(\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{B} \cdot \boldsymbol{H} \right) \delta_{ij}. \tag{H.8}$$

The main point about this is the appearance of terms closely related to the electric and magnetic energy densities that we had before. Here they are appearing as pressures acting upon the volume in question. Angular momentum can be handled similarly. Look at chapter 6 of Jackson, "Classical Electrodynamics" to see some of this worked out in more detail (but beware of the non-SI units used).

Appendix I

Electromagnetism to geometrical optics

In the main notes and lectures we use a pictorial way to introduce wavefronts and rays. This is a good way to think to about them, but there is a more formal way to make the transition to geometrical optics, which brings out better where the approximations are hiding. See chapter 3 of Born & Wolf "Principle of Optics" for full details, but note, as is often the case with classic old textbooks, that they do not use the SI versions of Maxwell's equations. Below is a summary of their derivation translated into the familiar forms of these. The derivation itself stems from a paper by Somerfeld and Runge in 1911.

I.1 From EM to geometrical optics

The starting point is to assume oscillating fields (not necessarily just plane waves!) of the form

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= \boldsymbol{E}_0(\boldsymbol{r})e^{-i\omega t}, & \text{(I.1)} \\ \boldsymbol{H}(\boldsymbol{r},t) &= \boldsymbol{H}_0(\boldsymbol{r})e^{-i\omega t}. & \text{(I.2)} \end{aligned}$$

The material forms of Maxwell's equations with no charge or current (we will always be assuming dielectrics) become, after setting $\boldsymbol{D} = \epsilon\boldsymbol{E}$ and $\boldsymbol{B} = \mu\boldsymbol{H}$,

$$\begin{aligned} \nabla\epsilon\boldsymbol{E}_0 &= 0, & \text{(I.3)} \\ \nabla\mu\boldsymbol{H}_0 &= 0, & \text{(I.4)} \\ \nabla\times\boldsymbol{E}_0 &= i\mu\omega\boldsymbol{H}_0, & \text{(I.5)} \\ \nabla\times\boldsymbol{H}_0 &= -i\epsilon\omega\boldsymbol{E}_0. & \text{(I.6)} \end{aligned}$$

We then suppose spatial variations of the form

$$\boldsymbol{E}_0(\boldsymbol{r}) = \boldsymbol{e}(\boldsymbol{r})e^{ik_0\tau(\boldsymbol{r})}, \qquad \text{(I.7)}$$

$$\boldsymbol{H}_0(\boldsymbol{r}) = \boldsymbol{h}(\boldsymbol{r})e^{ik_0\tau(\boldsymbol{r})}, \qquad \text{(I.8)}$$

where $\tau(\boldsymbol{r})$ is a real function of position called the "optical path" and $k_0 = \omega/c$ is the wave number in the vacuum.

The idea here is that we extract the rapid part of the variation into the exponential factor so that the field amplitude factors $\boldsymbol{e}$ and $\boldsymbol{h}$ are slowly varying.

To give a specific example of a familiar case, consider a plane wave travelling in direction $\hat{\boldsymbol{s}}$ in a material of refractive index n. The wave vector $\boldsymbol{k} = (\omega/v_\phi)\hat{\boldsymbol{s}} = k_0 n\hat{\boldsymbol{s}}$, so $\boldsymbol{k}\cdot\boldsymbol{r} = k_0 n\boldsymbol{r}\cdot\hat{\boldsymbol{s}}$. The optical path in this case is given by $\tau(\boldsymbol{r}) = n\boldsymbol{r}\cdot\hat{\boldsymbol{s}}$ and the amplitude vectors $\boldsymbol{e}$ and $\boldsymbol{h}$ would be constant.

We substitute for $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ in the forms involving τ into Maxwell's equations, where we make use of standard vector calculus identities (which we won't specify to avoid clutter). For instance

$$\begin{aligned}\nabla\times\boldsymbol{H}_0 &= \nabla\times\boldsymbol{h}e^{ik_0\tau}, && \text{(I.9)}\\ &= [\nabla\times\boldsymbol{h} + ik_0(\nabla\tau)\times\boldsymbol{h}]e^{ik_0\tau}. && \text{(I.10)}\end{aligned}$$

Doing the same for all four of Maxwell's equations we arrive at

$$\begin{aligned}(\nabla\epsilon)\cdot\boldsymbol{e} + \epsilon\nabla\cdot\boldsymbol{e} + ik_0\epsilon\boldsymbol{e}\cdot\nabla\tau &= 0, && \text{(I.11)}\\ (\nabla\mu)\cdot\boldsymbol{h} + \mu\nabla\cdot\boldsymbol{h} + ik_0\mu\boldsymbol{h}\cdot\nabla\tau &= 0, && \text{(I.12)}\\ ik_0\nabla\tau\times\boldsymbol{e} + \nabla\times\boldsymbol{e} &= i\mu\omega\boldsymbol{h}, && \text{(I.13)}\\ ik_0\nabla\tau\times\boldsymbol{h} + \nabla\times\boldsymbol{h} &= -i\epsilon\omega\boldsymbol{e}, && \text{(I.14)}\end{aligned}$$

Up to this point we have been exact, but now we are in a position to make approximations suitable for geometrical optics. The feature of geomtrical optics is the neglect of the wavelength compared to other scales of interest, i.e. both ω and k_0 can be assumed to be large. Neglecting all other terms in the above equations, and using $\omega/k_o = c$, we arrive at a mercifully simpler set of relations:

$$\begin{aligned}\boldsymbol{e}\cdot\nabla\tau &= 0, && \text{(I.15)}\\ \boldsymbol{h}\cdot\nabla\tau &= 0, && \text{(I.16)}\\ \nabla\tau\times\boldsymbol{e} &= \mu c\boldsymbol{h}, && \text{(I.17)}\\ \nabla\tau\times\boldsymbol{h} &= -\epsilon c\boldsymbol{e}. && \text{(I.18)}\end{aligned}$$

The first two equations can be derived from the last two by taking the scalar product with $\nabla\tau = \mathrm{grad}(\tau)$, so we need only focus of the second two. Multiplying the fourth equation by μc and using the third equation to substitute for $\mu c\boldsymbol{h}$ gives

$$\nabla\tau\times(\nabla\tau\times\boldsymbol{e}) = -\mu\epsilon c^2\boldsymbol{e}. \tag{I.19}$$

Using the usual relation for vector triple products gives:

$$(\nabla\tau\cdot\boldsymbol{e})\nabla\tau - (\nabla\tau)^2\boldsymbol{e} = -n^2\boldsymbol{e}. \tag{I.20}$$

We have used the relation $v_\phi^2 = 1/\mu\epsilon = c^2/n^2$, thus $n^2 = c^2\mu\epsilon$. The first term drops out and we are left, for non-trivial $\boldsymbol{e}$ with the rather beautiful relation

$$(\nabla\tau)^2 = n^2. \tag{I.21}$$

Equation I.21 is a differential equation for τ in terms of the refractive index n, which can be a function of position. It is the fundamental equation of geometrical optics. The function τ which I call the optical path is sometimes called the "eikonal" and the above equation the "eikonal equation". Be careful to avoid the temptation when looking at the eikonal equation to think that it means $\nabla\tau = \pm n$: this is completely wrong. $\nabla\tau$ is a vector. Fully expanded in Cartesian coordinates the eikonal equation looks like:

$$\left(\frac{\partial\tau}{\partial x}\right)^2 + \left(\frac{\partial\tau}{\partial y}\right)^2 + \left(\frac{\partial\tau}{\partial z}\right)^2 = n^2. \tag{I.22}$$

Surfaces of constant τ:

$$\tau(\boldsymbol{r}) = \text{constant}, \tag{I.23}$$

are surfaces of constant phase, i.e. the wave-fronts that I showed pictorially in the main section. Rays of light run perpendicular to these wavefronts, i.e. parallel to $\nabla\tau$. Given this, $\boldsymbol{e}$ and $\boldsymbol{h}$ are perpendicular to the ray direction, as one might expect.

Energy densities and flow

The energy densities in the fields are given (as we know) by $\epsilon E^2/2$ and $\mu H^2/2$. Usually we are only interested in their time averages because of their very rapid variation at optical frequency. One must take care when calculating energies and powers to use the real parts (otherwise one tends to end up with a final value of zero). Here we have

$$\boldsymbol{E} = \boldsymbol{e}(\boldsymbol{r})e^{ik_0\tau - i\omega t}, \tag{I.24}$$

$$\boldsymbol{H} = \boldsymbol{h}(\boldsymbol{r})e^{ik_0\tau - i\omega t}. \tag{I.25}$$

The real part of $\boldsymbol{E}$ is given by

$$\text{Re}(\boldsymbol{E}) = \frac{1}{2}\left[\boldsymbol{E} + \boldsymbol{E}^*\right], \tag{I.26}$$

$$= \frac{1}{2}\left[\boldsymbol{e}e^{i(k_0\tau - \omega t)} + \boldsymbol{e}^* e^{-i(k_0\tau - \omega t)}\right], \tag{I.27}$$

with asterisks denoting comple conjugates. Hence the real part squared gives

$$\text{Re}(\boldsymbol{E})^2 = \frac{1}{4}\left[\boldsymbol{e}\cdot\boldsymbol{e}e^{2i(k_0\tau - \omega t)} + 2\boldsymbol{e}\cdot\boldsymbol{e}^* + \boldsymbol{e}^*\cdot\boldsymbol{e}^* e^{-2i(k_0\tau - \omega t)}\right]. \tag{I.28}$$

The rotating phasor terms (first and last on the right-hand side) drop out on taking the time-average, leaving

$$\left\langle \text{Re}(\boldsymbol{E})^2 \right\rangle = \frac{1}{2}\boldsymbol{e}\cdot\boldsymbol{e}^*. \tag{I.29}$$

With a similar expression involving $\boldsymbol{h}$, the mean energy density in the EM field in terms of $\boldsymbol{e}$ and $\boldsymbol{h}$ is

$$\langle u \rangle = \langle u_E \rangle + \langle u_H \rangle = \frac{1}{4}\epsilon\,\boldsymbol{e}\cdot\boldsymbol{e}^* + \frac{1}{4}\mu\,\boldsymbol{h}\cdot\boldsymbol{h}^*. \tag{I.30}$$

Substituting for $\boldsymbol{e}^*$ from Eq. I.18 into I.30, the first term can be transformed as follows

$$\epsilon\,\boldsymbol{e}\cdot\boldsymbol{e}^* = \frac{1}{c}\boldsymbol{e}\cdot\boldsymbol{h}^*\times\nabla\tau. \tag{I.31}$$

The second term in the energy density can be similarly manipulated by substitution of $\boldsymbol{h}$ using Eq. I.17 to end up with the same final expression:

$$\mu\,\boldsymbol{h}\cdot\boldsymbol{h}^* = \frac{1}{c}\boldsymbol{e}\cdot\boldsymbol{h}^*\times\nabla\tau. \tag{I.32}$$

To the level of approximation used in geometric optics, the time-averaged electric and magnetic energy densities are equal and we can write

$$\langle u\rangle = \frac{1}{2}\epsilon\, \boldsymbol{e}\cdot\boldsymbol{e}^*, \tag{I.33}$$

a relation we will use below.

We can carry out a similar procedure to time-average the Poynting vector to find that

$$\langle \boldsymbol{S}\rangle = \frac{1}{4}\left[\boldsymbol{e}\times\boldsymbol{h}^* + \boldsymbol{e}^*\times\boldsymbol{h}\right] = \frac{1}{2}\mathsf{Re}(\boldsymbol{e}\times\boldsymbol{h}^*). \tag{I.34}$$

Again substituting for $\boldsymbol{h}^*$ we obtain

$$\begin{aligned}
\langle \boldsymbol{S}\rangle &= \frac{1}{2\mu c}\mathsf{Re}(\boldsymbol{e}\times(\nabla\tau\times\boldsymbol{e}^*), && \text{(I.35)}\\
&= \frac{1}{2\mu c}\mathsf{Re}((\boldsymbol{e}\cdot\boldsymbol{e}^*)\nabla\tau - (\boldsymbol{e}\cdot\nabla\tau)\boldsymbol{e}^*), && \text{(I.36)}\\
&= \frac{1}{2\mu c}(\boldsymbol{e}\cdot\boldsymbol{e}^*)\nabla\tau, && \text{(I.37)}
\end{aligned}$$

showing that the energy flow is along the ray direction, perpendicular to the wavefronts, confirming what we found earlier. Taking into account the expression for $\langle u\rangle$ we have

$$\langle \boldsymbol{S}\rangle = \frac{\langle u\rangle}{\mu\epsilon c}\nabla\tau, \tag{I.38}$$

Setting $\mu\epsilon = n^2/c^2$, leads to

$$\langle \boldsymbol{S}\rangle = \frac{c\,\langle u\rangle}{n^2}\nabla\tau. \tag{I.39}$$

We know from the eikonal equation, $n = |\nabla\tau|$, while $c/n = v_\phi$, the wavespeed or phase velocity in the material. Defining a unit vector along the ray as $\boldsymbol{s} = \nabla\tau/|\nabla\tau|$, we find:

$$\langle \boldsymbol{S}\rangle = v_\phi\,\langle u\rangle\,\boldsymbol{s}. \tag{I.40}$$

This equation has the simple interpretation that in geometrical optics, the mean energy flux along a ray is equal to the mean energy density times the wave speed.

If the position a length s along a ray is denoted by $\boldsymbol{r}(s)$, then clearly

$$\frac{d\boldsymbol{r}(s)}{ds} = \boldsymbol{s} = \frac{\nabla\tau}{|\nabla\tau|}, \tag{I.41}$$

or

$$n\frac{d\boldsymbol{r}}{ds} = \nabla\tau. \tag{I.42}$$

Differentiating this with respect to s

$$\begin{aligned}
\frac{d}{ds}\left(n\frac{d\boldsymbol{r}}{ds}\right) &= \frac{d}{ds}(\nabla\tau), && \text{(I.43)}\\
&= \frac{d\boldsymbol{r}}{ds}\cdot\nabla(\nabla\tau), && \text{(I.44)}\\
&= \frac{1}{n}\nabla\tau\cdot\nabla(\nabla\tau), && \text{(I.45)}\\
&= \frac{1}{2n}\nabla(\nabla\tau)^2, && \text{(I.46)}\\
&= \frac{1}{2n}\nabla n^2, && \text{(I.47)}
\end{aligned}$$

using the eikonal equation to obtain the last line. We a differential equation for the ray

$$\frac{d}{ds}\left(n\frac{d\boldsymbol{r}}{ds}\right) = \nabla n. \tag{I.48}$$

If n is constant, this reduces to

$$\frac{d^2\boldsymbol{r}}{ds^2} = 0, \tag{I.49}$$

for which the solution is a straight line. Otherwise, if $\nabla n \neq \mathbf{0}$, we have the possibility of curved ray paths as with mirages where temperature stratification means there is a vertical gradient of refractive index causing light rays to be bent both upwards (hot ground surface) and downwards (cold ground surface).

You can refer to Chapter 3 of Born and Wolf for further details where several other basic theorems of geometrical optics are established along these lines.

Appendix J

Elements of Huygens-Fresnel-Kirchoff diffraction theory

J.1 Introduction

The next approximation beyond geometrical optics is the scalar wave theory of Huygens and Fresnel, which was put onto a more mathematical basis by Kirchoff in the 1880s. This quantifies the idea of secondary wavelets and provides an extremely useful approximation to many problems in optics. It contains as special approximations Fraunhofer far-field and Fresnel near-field diffraction theory. The first leads to integrals across wavefronts involve phase that changes linearly with position which leads to Fourier integrals. The second allows for quadratic phase variations and much more difficult integrals. The exposition presented here follows that of Born & Wolf's "Principles of Optics", chapter 8.

J.2 Huygens-Fresnel principle

According to Huygens' principle each part of a wavefront acts as the source of secondary waves whose envelope forms the next wavefront. Fresnel generalised this by allowing the wavelets to interfere and introducing an angle-dependent factor to prevent back-propagating wavefronts. We assume monochromatic waves and focus on the spatial variation of the wave. Fig. J.1 shows the key elements of the Huygens-Fresnel principle. A spherical wave emitted from a point P_0 leads to the highlighted wavefront. The wave amplitude at a point P is the sum across the wavefront of secondary waves emitted from the wavefront. The amplitude of a spherical wave drops with distance from its source, so the amplitude at Q and all other points on the wavefront can be written as Ae^{ikr_0}/r_0, thus we guess that the amplitude U at point P should be something like

$$U(P) = \int_S \frac{Ae^{ikr_0}}{r_0} \frac{e^{ikr}}{r} \, dS, \tag{J.1}$$

that is a surface integral across the wavefront accounting for the distance r between the area element (Q in the diagram) and P and, particularly important, the phase change kr from that element to P.

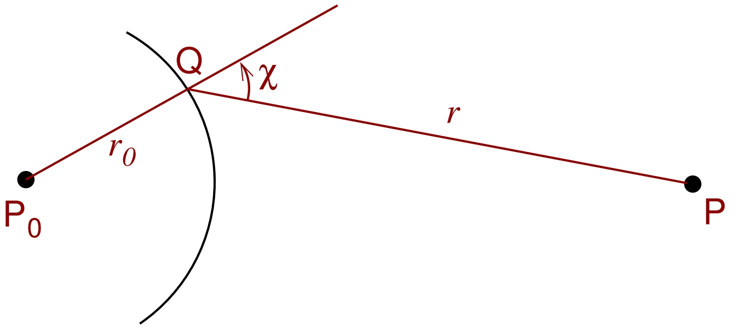

Figure J.1: Construction to show key concepts of the Huygens-Fresnel principle. The wave amplitude t P due to a spherical wavefront emitted from P_0 is constructed by adding contributions at all points from the wavefront. Q is one such point. The angle χ is between the direction of travel of the wavefront and the line from Q to P.

Equation J.1 is not quite right because we know that the wavefront travels in the direction $\chi = 0$, whereas the expression is symmetric with respect to setting $\chi = \pi$; this is the problem of backwards-propagating waves implicit in Huygens' construction. Fresnel supposed that there was in addition some function $K(\chi)$ which favoured $\chi = 0$; he assumed that it peaked at $\chi = 0$ and had dropped to 0 by $\chi = 0$. Adding this, and removing the constant amplitude across the wavefront we deduce

$$U(P) = \frac{Ae^{ikr_0}}{r_0} \int_S \frac{e^{ikr}}{r} K(\chi)\, dS. \tag{J.2}$$

Fresnel split the integral into "zones" across the wavefront with alternating phases. His theory was important in establishing the wave theory of light. He submitted it in 1818 in response to a prize competition on the nature of diffraction at a time when Newton's corpuscular theory of light was still believed by many to be correct. The judging panel consisted of Laplace, Biot, Poisson, Gay-Lussac and Arago, a veritable who's-who of famous French physicists, several of whom were in the corpuscular camp. Poisson soon made the surprising deduction that if Fresnel was right, there should be a spot of light at the centre of the geometrical shadow cast by a circular disc where the intensity of light is the same as if the disc were not there. Poisson assumed that this was not the case so Fresnel's theory was wrong, but it was subsequently confirmed by Arago. Apparently even this did not persuade the committee of the wave theory, but Fresnel was awarded the prize in any case. See Born & Wolf for more on Fresnel's work.

The integral above seems plausible, but lacks justification, and has no explicit form for $K(\chi)$. Although in many cases of interest, χ can be assumed to be close to zero, so the exact functional form does not matter, this is not very satisfactory. Kirchoff provided an answer in the 1880s.

J.3 Kirchoff's diffraction theory

Like the Huygens-Fresnel theory, Kirchoff's work concerns scalar waves that satisfy the wave equation

$$\nabla^2 V = \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2}, \tag{J.3}$$

so polarisation phenomena are not accounted for from the start. We specialise to monochromatic waves

$$V(\boldsymbol{r},t) = U(\boldsymbol{r})e^{-i\omega t}, \tag{J.4}$$

which on substitution in the wave equation shows that U satisfies

$$\nabla^2 U + k^2 U = 0, \tag{J.5}$$

where $k = \omega/c$ as usual. This is known as *Helmholz's equation*.

The path we will now follow is to work out expressions for U at some arbitrary point P in terms of its value (and derivatives) on some surface surrounding P. The surface will later be identified with a wavefront, and we will obtain an expression that can be compared with the Huygens-Fresnel version of the previous section. To get there we need a vector calculus identity due to Green (of Green's theorem fame). Consider the two scalar fields ϕ and ψ, and from them form the vector field $\boldsymbol{W} = \phi\nabla\psi$. Gauss' theorem applied to this leads to

$$\int_V \nabla\cdot(\phi\nabla\psi)\, dV = \oint_S (\phi\nabla\psi)\cdot d\boldsymbol{S}. \tag{J.6}$$

The divergence on the left is expanded

$$\nabla\cdot(\phi\nabla\psi) = \nabla\phi\cdot\nabla\psi + \phi\nabla^2\psi. \tag{J.7}$$

We can subtract a similar equation with ϕ and ψ swapped to deduce that

$$\int_V \left(\phi\nabla^2\psi - \psi\nabla^2\phi\right)\, dV = \oint_S (\phi\nabla\psi - \psi\nabla\phi)\cdot d\boldsymbol{S}. \tag{J.8}$$

This is the identity that we want which we apply to two functions U and U', that both satisfy Helmholz's equation which allows us to replace the action of applying the Laplacian, ∇^2, by mutliplying by $-k^2$.

We apply the identity J.8 to a surface surrounding the point P at which we want to determine the value of the wave amplitude U given its value on the surrounding surface. Making these substitutions into Green's identity gives

$$-k^2 \int_V (UU' - U'U)\, dV = -\oint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right) dS. \tag{J.9}$$

Here we have set the $\nabla U\cdot d\boldsymbol{S} = -\partial U/\partial n$ where n is the distance along the surface element formals but defined to point *inwards* (this will correspond to the direction of travel of the wavefront later on). We deduce that

$$\oint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right) dS = 0. \tag{J.10}$$

Now let $U' = e^{ikr}/r$ where r is the distance from P. This is a spherical wave and a solution of Helmholz's equation, i.e. $\nabla^2 U' + k^2 U' = 0$, as required. It diverges at P, so we must exclude a small region around P from the original volume integral. Let this small region be a sphere of radius ϵ with surface S' then, using a hopefully obvious notation

$$\oint_{S'} + \oint_S \left(U\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial U}{\partial n}\right) dS = 0. \tag{J.11}$$

The normals to the small sphere point inwards to the main region of integration, i.e. radially away from P, hence the derivative $\partial/\partial n$ on the sphere is equivalent to $\partial/\partial r$. This allows us to write

$$\begin{aligned}\oint_S \left(U \frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial U}{\partial n}\right) dS &= -\oint_{S'} \left(U \frac{e^{ikr}}{r}\left(ik - \frac{1}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial U}{\partial n}\right) dS', \\ &= -\oint_{S'} \left(U \frac{e^{ik\epsilon}}{\epsilon}\left(ik - \frac{1}{\epsilon}\right) - \frac{e^{ik\epsilon}}{\epsilon}\frac{\partial U}{\partial n}\right) dS'.\end{aligned}$$

The element of area on the small sphere can be written $dS' = \epsilon^2 \, d\Omega$, where $d\Omega$ is an element of solid angle. The integral on the left is independent of ϵ, so on the right we can take the limit $\epsilon \to 0$ whereby only the second of the three terms in the integrand on the right survives and one finds

$$U(P) = \frac{1}{4\pi} \oint_S \left(U \frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial U}{\partial n}\right) dS, \tag{J.12}$$

with the 4π coming from $\oint d\Omega$.

Equation J.12 is the integral theorem of Helmholz and Kirchoff. It gives the value of the scalar wave function U at P in terms of an integral over u and its gradient on an arbitrary surface enclosing P. This is not the same as the Huygens-Fresnel integral, but Kirchoff argued that in many cases one could restrict the integral to the portion of wavefront defined by the entrance aperture (e.g. objective lens) to an instrument, i.e. one could assume $U = 0$ and $\partial U/\partial n = 0$ on the rest of the surface. See Born & Wolf for further details. Taking the case of Fig. J.1, U at the wavefront is given by

$$U = \frac{A e^{ikr_0}}{r_0}, \tag{J.13}$$

while $\partial/\partial n \equiv \partial/\partial r_0$ when applied to this function, hence

$$\frac{\partial U}{\partial n} = \frac{A e^{ikr_0}}{r_0}\left(ik - \frac{1}{r_0}\right). \tag{J.14}$$

When applied to the term e^{ikr}/r, $\partial/\partial n$ is equivalent to travelling to towards small r and at an angle χ compared to the radial direction towards P (see Fig. J.1), and therefore $\partial/\partial n \equiv -\cos\chi \partial/\partial r$. This leaves

$$U(P) \approx \frac{1}{4\pi}\frac{A e^{ikr_0}}{r_0} \int_A \left[-\frac{e^{ikr}}{r}\left(ik - \frac{1}{r}\right)\cos\chi - \frac{e^{ikr}}{r}\left(ik - \frac{1}{r_0}\right)\right] dS, \tag{J.15}$$

where the A on the integral indicates an integral over the wavefront defined by the entrance aperture.

If we do not assume that the wavelength is small compared to the distance so that $k \gg r$ and $k \gg r_0$, and remember $k = 2\pi/\lambda$, then we can drop the $1/r$ and $1/r_0$ terms in J.15 to obtain

$$U(P) \approx -\frac{i}{2\lambda}\frac{A e^{ikr_0}}{r_0} \int_A \frac{e^{ikr}}{r}(1 + \cos\chi)\, dS. \tag{J.16}$$

This is Kirchoff's diffraction equation. Comparing with the Huygens-Fresnel equation, Eq. J.2, we see that

$$K(\chi) = -\frac{i}{2\lambda}(1 + \cos\chi). \tag{J.17}$$

This indeed peaks in the direction $\chi = 0$, and drops to zero at $\chi = \pi$, killing the backwards wave, but interestingly is non-zero for $\chi = \pi/2$, contrary to Fresnel's presumption. You should refer once more to chapter 8 of Born & Wolf if you want to know further details of this.

Kirchoff's theory is still only an approximation to reality, but it is a good one when the wavelength is small compared to other length scales of the problem. We used it to compute the wave field shown in Fig. 8.2. It might not be obvious, but it is the e^{ikr} factor in the integral that is most important, as in many cases the $1/r$ and $1 + \cos\chi$ terms are relatively constant over the region of integration for points P of interest.

Kirchoff's theory provides the justification for Fraunhofer and Fresnel diffraction, and is a fair approximation to the few cases where the diffraction of EM waves can be calculated rigorously, as long as the wavelength remains small compared to other length scales; see Born & Wolf chapter 11 for more on this.

Appendix K

Exact analysis of thin film interference

When discussing thin film interference in section 8.2, weak reflections were assumed so that multiple reflections could be ignored. One could include these using the sums of infinite series, but here a different approach is considered, one that can be easily extended to multiple layers, which is not practical with the multiple reflection approach.

Consider Fig. K.1 which shows schematically a wave travelling from left to right towards an interface between medium 1 and 2, and then 2 and 3, with 4 reference points identified, before and after the 1-to-2 interface and before and after the 2-to-3 interface. Rather than consider a series of reflections, we will model this in terms of right- and left-travelling waves in each medium and work out relations between their amplitudes at each of the reference points indexed 1, 2, 3 and 4. The fields of right- and left-travelling waves will be distinguished by dashed on the right-travelling ones. Matching the parallel components of $\boldsymbol{E}$ and $\boldsymbol{H}$ at the 1/2 interface gives

$$\begin{aligned} E_1 + E_1' &= E_2 + E_2', && \text{(K.1)} \\ H_1 - H_1' &= H_2 - H_2'. && \text{(K.2)} \end{aligned}$$

For non-magnetic media $H \propto nE$, we obtain

$$\begin{aligned} E_1 + E_1' &= E_2 + E_2', && \text{(K.3)} \\ n_1 E_1 - n_1 E_1' &= n_2 E_2 - n_2 E_2'. && \text{(K.4)} \end{aligned}$$

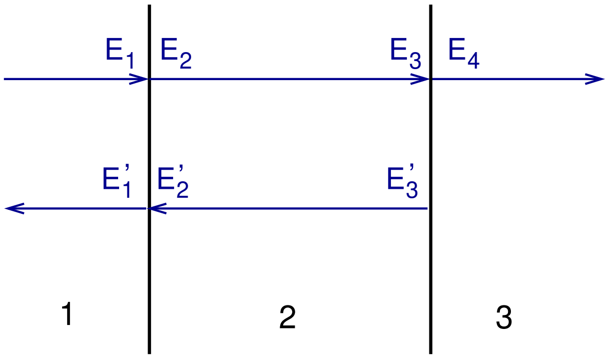

Figure K.1: The reflection of a wave travelling rightwards in medium 1 is modelled with a series of rightwards and leftwards waves in each medium except the last since no wave comes from the far right. The labels define the field amplitudes at each interface.

In matrix form we have

$$\begin{pmatrix} 1 & 1 \\ n_1 & -n_1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ n_2 & -n_2 \end{pmatrix} \begin{pmatrix} E_2 \\ E_2' \end{pmatrix} . \tag{K.5}$$

Reference points 2 and 3 are both in medium 2 (the coating), but their respective fields differ by phase factors due to propagation over distance d. This leads to

$$\begin{pmatrix} E_3 \\ E_3' \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{ik_2 d} & 0 \\ 0 & \mathrm{e}^{-ik_2 d} \end{pmatrix} \begin{pmatrix} E_2 \\ E_2' \end{pmatrix} , \tag{K.6}$$

which follows from the form of travelling waves, $\exp i(\pm kx - \omega t)$, with the $+$ and $-$ signs applying to waves travelling to the right and left respectively, and k_2 since it is the wave number in medium 2 that matters. Finally we have a relation similar to the first relating the fields on either side of the 2/3 interface:

$$\begin{pmatrix} 1 & 1 \\ n_2 & -n_2 \end{pmatrix} \begin{pmatrix} E_3 \\ E_3' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ n_3 & -n_3 \end{pmatrix} \begin{pmatrix} E_4 \\ 0 \end{pmatrix} , \tag{K.7}$$

where the left-wards travelling wave in medium 3 is set to zero because there is none.

Multiplying the final equation by the inverse of the right-hand matrix

$$\begin{pmatrix} E_4 \\ 0 \end{pmatrix} = \frac{1}{2n_3} \begin{pmatrix} n_3 & 1 \\ n_3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ n_2 & -n_2 \end{pmatrix} \begin{pmatrix} E_3 \\ E_3' \end{pmatrix} . \tag{K.8}$$

Using the other two relations similarly then leads to

$$\begin{pmatrix} E_4 \\ 0 \end{pmatrix} = \frac{1}{4n_2 n_3} \begin{pmatrix} n_3 & 1 \\ n_3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ n_2 & -n_2 \end{pmatrix} \begin{pmatrix} z & 0 \\ 0 & 1/z \end{pmatrix} \begin{pmatrix} n_2 & 1 \\ n_2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ n_1 & -n_1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix} , \tag{K.9}$$

where $z = \exp(ik_2 d)$. Combining the first and last pairs of matrices

$$\begin{pmatrix} E_4 \\ 0 \end{pmatrix} = \frac{1}{4n_2 n_3} \begin{pmatrix} n_3 + n_2 & n_3 - n_2 \\ n_3 - n_2 & n_3 + n_2 \end{pmatrix} \begin{pmatrix} z & 0 \\ 0 & 1/z \end{pmatrix} \begin{pmatrix} n_2 + n_1 & n_2 - n_1 \\ n_2 - n_1 & n_2 + n_1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix} , \tag{K.10}$$

so

$$\begin{pmatrix} E_4 \\ 0 \end{pmatrix} = \frac{1}{4n_2 n_3} \begin{pmatrix} (n_3 + n_2)z & (n_3 - n_2)/z \\ (n_3 - n_2)z & (n_3 + n_2)/z \end{pmatrix} \begin{pmatrix} n_2 + n_1 & n_2 - n_1 \\ n_2 - n_1 & n_2 + n_1 \end{pmatrix} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix} . \tag{K.11}$$

The bottom row of this equation leads to

$$\frac{E_1'}{E_1} = r = -\frac{(n_3 - n_2)(n_2 + n_1)z + (n_3 + n_2)(n_2 - n_1)/z}{(n_3 - n_2)(n_2 - n_1)z + (n_3 + n_2)(n_2 + n_1)/z} , \tag{K.12}$$

and the reflectance $R = rr^*$ can be shown to be

$$R = \frac{(n_3 - n_2)^2(n_2 + n_1)^2 + (n_3 + n_2)^2(n_2 - n_1)^2 + 2(n_3^2 - n_2^2)(n_2^2 - n_1^2)\cos(2k_2 d)}{(n_3 - n_2)^2(n_2 - n_1)^2 + (n_3 + n_2)^2(n_2 + n_1)^2 + 2(n_3^2 - n_2^2)(n_2^2 - n_1^2)\cos(2k_2 d)} . \tag{K.13}$$

where use has been made of $zz^* = 1$.

Comparing the result for R in K.13 with the approximate expression of Eq. 8.11, we see two extra terms in the denominator. If $n_1 < n_2 < n_3$, then each of the refractive index dependent terms in the numerator is positive, and to reduce the value of R we require $\cos(2k_2 d) < 0$, ideally -1. Assuming this, then $2k_2 d = \pi + 2\pi m$ for integer m. For $m = 0$, this is a quarter-wave layer, and zero reflectance is obtained for

$$(n_3 - n_2)^2(n_2 + n_1)^2 + (n_3 + n_2)^2(n_2 - n_1)^2 - 2(n_3^2 - n_2^2)(n_2^2 - n_1^2) = 0, \tag{K.14}$$

which is equivalent to

$$(n_3 - n_2)(n_2 + n_1) = (n_3 + n_2)(n_2 - n_1), \tag{K.15}$$

which leads to the same result as obtained before in 8.9, $n_2 = \sqrt{n_1 n_3}$. The more sophisticated approach leads to the same main results as before, apart from the difference in the denominator (see 8.11 and Fig. 8.3).

Eq. K.13 is periodic in $2k_2 d$. Thus the reflectance value for $d = 0$, which in effect means that there is no dielectric coating, repeats every time $2k_2 d = 2\pi m$ or $d/\lambda_2 = m\pi/2$, with m an integer. Thus half-wave layers act as if they are not there at all. The matrix approach can be extended for multiple layers and lends itself to numerical solution.