Part II Electrodynamics and Optics

William Royce

September 26, 2024

Part II Physics, The University of Cambridge

Preface

There are, to the best of our knowledge, four forces at play in the Universe. At the very largest scales – those of planets or stars or galaxies – the force of gravity dominates. At the very smallest distances, the two nuclear forces hold sway. For everything in between, it is force of electromagnetism that rules.

At the atomic scale, electromagnetism (admittedly in conjunction with some basic quantum effects) governs the interactions between atoms and molecules. It is the force that underlies the periodic table of elements, giving rise to all of chemistry and, through this, much of biology. It is the force which binds atoms together into solids and liquids. And it is the force which is responsible for the incredible range of properties that different materials exhibit.

At the macroscopic scale, electromagnetism manifests itself in the familiar phenomena that give the force its name. In the case of electricity, this means everything from rubbing a balloon on your head and sticking it on the wall, through to the fact that you can plug any appliance into the wall and be pretty confident that it will work. For magnetism, this means everything from the shopping list stuck to your fridge door, through to trains in Japan which levitate above the rail. Harnessing these powers through the invention of the electric dynamo and motor has transformed the planet and our lives on it.

As if this wasn't enough, there is much more to the force of electromagnetism for it is, quite literally, responsible for everything you've ever seen. It is the force that gives rise to light itself.

Rather remarkably, a full description of the force of electromagnetism is contained in four simple and elegant equations. These are known as the *Maxwell equations*. There are few places in physics, or indeed in any other subject, where such a richly diverse set of phenomena flows from so little. The purpose of this course is to introduce the Maxwell equations and to extract some of the many stories they contain.

However, there is also a second theme that runs through this course. The force of electromagnetism turns out to be a blueprint for all the other forces. There are various mathematical symmetries and structures lurking within the Maxwell equations, structures which Nature then repeats in other contexts. Understanding the mathematical beauty of the equations will allow us to see some of the principles that underly the laws of physics, laying the groundwork for future study of the other forces.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Charge and Current

Each particle in the Universe carries with it a number of properties. These determine how the particle interacts with each of the four forces. For the force of gravity, this property is mass. For the force of electromagnetism, the property is called *electric charge*.

For the purposes of this course, we can think of electric charge as a real number, $q \in \mathbb{R}$. Importantly, charge can be positive or negative. It can also be zero, in which case the particle is unaffected by the force of electromagnetism.

The SI unit of charge is the *Coulomb*, denoted by C. It is, like all SI units, a parochial measure, convenient for human activity rather than informed by the underlying laws of the physics. (We'll learn more about how the Coulomb is defined in Section 3.6). At a fundamental level, Nature provides us with a better unit of charge. This follows from the fact that charge is quantised: the charge of any particle is an integer multiple of the charge carried by the electron which we denoted as $-e$, with

$$e = 1.602176634^{-19}\text{C}. \tag{1.1}$$

A much more natural unit would be to simply count charge as $q = ne$ with $n \in \mathbb{Z}$. Then electrons have charge -1 while protons have charge $+1$ and neutrons have charge 0. Nonetheless, in this course, we will bow to convention and stick with SI units.[1]

One of the key goals of this course is to move beyond the dynamics of point particles and onto the dynamics of continuous objects known as fields. To aid in this, it's useful to consider the *charge density*,

$$\rho(\mathbf{x}, t) \tag{1.2}$$

defined as charge per unit volume. The total charge Q in a given region $\mathcal{V}$ is simply $Q = \int_{\mathcal{V}} \mathrm{d}^3x\, \rho(\mathbf{x}, t)$. In most situations, we will consider smooth charge densities, which can be thought of as arising from averaging over many point-like particles. But, on occasion, we will return to the idea of a single particle of charge q, moving on some trajectory $\mathbf{r}(t)$, by writing $\rho = q\delta(\mathbf{x} - \mathbf{r}(t))$ where the delta-function ensures that all the charge sits at a point.

More generally, we will need to describe the movement of charge from one place to another. This is captured by a quantity known as the *current density* $\mathbf{J}(x, t)$, defined as follows: for every surface $\mathcal{S}$, the integral

$$I = \int_{\mathcal{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S} \tag{1.3}$$

[1] (An aside: the charge of quarks is actually $q = -e/3$ and $q = 2e/3$. This doesn't change the spirit of the above discussion since we could just change the basic unit. But, apart from in extreme circumstances, quarks are confined inside protons and neutrons so we rarely have to worry about this).

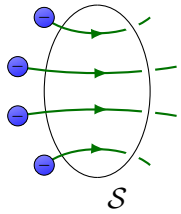


Fig. 1.1: Current flux.

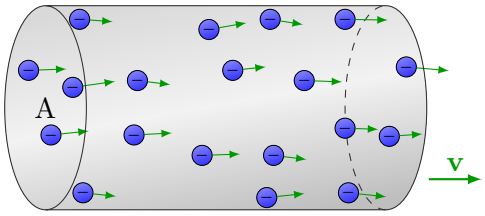


Fig. 1.2: The wire.

counts the charge per unit time passing through $\mathcal{S}$. (Here d$\mathbf{S}$ is the unit normal to $\mathcal{S}$). The quantity I is called the *current*. In this sense, the current density is the current-per-unit-area

The above is a rather indirect definition of the current density. To get a more intuitive picture, consider a continuous charge distribution in which the velocity of a small volume, at point $\mathbf{x}$, is given by $\mathbf{v}(\mathbf{x}, t)$. Then, neglecting relativistic effects, the current density is

$$\mathbf{J} = \rho\mathbf{v}. \tag{1.4}$$

In particular, if a single particle is moving with velocity $\mathbf{v} = \dot{\mathbf{r}}(t)$, the current density will be $\mathbf{J} = q\mathbf{v}\delta^3(\mathbf{x} - \mathbf{r}(t))$. This is illustrated in Fig. 1.1, where the underlying charged particles are shown as blue balls, moving through the surface $\mathcal{S}$.

As a simple example, consider electrons moving along a wire (see Fig. 1.2). We model the wire as a long cylinder of cross-sectional area A as shown below. The electrons move with velocity $\mathbf{v}$, parallel to the axis of the wire. (In reality, the electrons will have some distribution of speeds; we take $\mathbf{v}$ to be their average velocity). If there are n electrons per unit volume, each with charge q, then the charge density is $\rho = nq$ and the current density is $\mathbf{J} = nq\mathbf{v}$. The current itself is $I = |\mathbf{J}|A$..

Throughout this course, the current density $\mathbf{J}$ plays a much more prominent role than the current I. For this reason, we will often refer to $\mathbf{J}$ simply as the "current" although we'll be more careful with the terminology when there is any possibility for confusion.

1.1.1 The Conservation Law

The most important property of electric charge is that it's conserved. This, of course, means that the total charge in a system can't change. But it means much more than that because electric charge is conserved *locally*. An electric charge can't just vanish from one part of the Universe and turn up somewhere else. It can only leave one point in space by moving to a neighbouring point.

The property of local conservation means that ρ can change in time only if there is a compensating current flowing into or out of that region. We express this in the *continuity equation*,

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{J} = 0. \tag{1.5}$$

This is an important equation. It arises in any situation where there is some quantity that is locally conserved.

To see why the continuity equation captures the right physics, it's best to consider the change in the total charge Q contained in some region $\mathcal{V}$.

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_{\mathcal{V}} \mathrm{d}^3x \, \frac{\partial \rho}{\partial t} = -\int_{\mathcal{V}} \mathrm{d}^3x \, \boldsymbol{\nabla} \cdot \mathbf{J} = -\int_{\mathcal{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S} \,. \tag{1.6}$$

From our previous discussion, $\int_{\mathcal{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S}$ is the total current flowing out through the boundary $\mathcal{S}$ of the region $\mathcal{V}$. (It is the total charge flowing *out*, rather than in, because $\mathrm{d}\mathbf{S}$ is the outward normal to the region $\mathcal{V}$). The minus sign is there to ensure that if the net flow of current is outwards, then the total charge decreases.

If there is no current flowing out of the region, then $\mathrm{d}Q/\mathrm{d}t = 0$. This is the statement of (global) conservation of charge. In many applications we will take $\mathcal{V}$ to be all of space, $\mathbb{R}^3$, with both charges and currents localised in some compact region. This ensures that the total charge remains constant.

1.2 Forces and Fields

Any particle that carries electric charge experiences the force of electromagnetism. But the force does not act directly between particles. Instead, Nature chose to introduce intermediaries. These are *fields*.

In physics, a "field" is a dynamical quantity which takes a value at every point in space and time. To describe the force of electromagnetism, we need to introduce two fields, each of which is a three-dimensional vector. They are called the *electric field* $\mathbf{E}$ and the *magnetic field* $\mathbf{B}$,

$$\mathbf{E}(\mathbf{x}, t) \quad \text{and} \quad \mathbf{B}(\mathbf{x}, t). \tag{1.7}$$

When we talk about a "force" in modern physics, we really mean an intricate interplay between particles and fields. There are two aspects to this. First, the charged particles create both electric and magnetic fields. Second, the electric and magnetic fields guide the charged particles, telling them how to move. This motion, in turn, changes the fields that the particles create. We're left with a beautiful dance with the particles and fields as two partners, each dictating the moves of the other.

This dance between particles and fields provides a paradigm which all other forces in Nature follow. It feels like there should be a deep reason that Nature chose to introduce fields associated to all the forces. And, indeed, this approach does provide one overriding advantage: all interactions are local. Any object – whether particle or field – affects things only in its immediate neighbourhood. This influence can then propagate through the field to reach another point in space, but it does not do so instantaneously. It takes time for a

particle in one part of space to influence a particle elsewhere. This lack of instantaneous interaction allows us to introduce forces which are compatible with the theory of special relativity, something that we will explore in more detail in Section 5.

The purpose of this course is to provide a mathematical description of the interplay between particles and electromagnetic fields. In fact, you've already met one side of this dance: the position $\mathbf{r}(t)$ of a particle of charge q is dictated by the electric and magnetic fields through the Lorentz force law,

$$\boxed{\mathbf{F} = q(\mathbf{E} + \dot{\mathbf{r}} \times \mathbf{B}).} \tag{1.8}$$

The motion of the particle can then be determined through Newton's equation $\mathbf{F} = m\ddot{\mathbf{r}}$. Roughly speaking, an electric field accelerates a particle in the direction $\mathbf{E}$, while a magnetic field causes a particle to move in circles in the plane perpendicular to $\mathbf{B}$.

We can also write the Lorentz force law in terms of the charge distribution $\rho(\mathbf{x}, t)$ and the current density $\mathbf{J}(\mathbf{x}, t)$. Now we talk in terms of the *force density* $\mathbf{f}(\mathbf{x}, t)$, which is the force acting on a small volume at point $\mathbf{x}$. Now the Lorentz force law reads

$$\mathbf{f} = \rho\mathbf{E} + \mathbf{J} \times \mathbf{B}. \tag{1.9}$$

1.2.1 The Maxwell Equations

In this course, most of our attention will focus on the other side of the dance: the way in which electric and magnetic fields are created by charged particles. This is described by a set of four equations, known collectively as the *Maxwell equations*. They are:

$$\boxed{\begin{aligned} \boldsymbol{\nabla} \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}, && (1.10)\\ \boldsymbol{\nabla} \cdot \mathbf{B} &= 0, && (1.11)\\ \boldsymbol{\nabla} \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0, && (1.12)\\ \boldsymbol{\nabla} \times \mathbf{B} - \mu_0\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} &= \mu_0 \mathbf{J}. && (1.13) \end{aligned}}$$

The equations involve two constants. The first is the *electric constant* (known also, in slightly old-fashioned terminology, as the *permittivity of free space*),

$$\epsilon_0 \approx 8.85 \times 10^{-12} \mathrm{m}^{-3}\,\mathrm{kg}^{-1}\,\mathrm{s}^2\,\mathrm{C}^2. \tag{1.14}$$

It can be thought of as characterising the strength of the electric interactions. The other is the *magnetic constant* (or *permeability of free space*),

$$\begin{aligned} \mu_0 &= 4\pi \times 10^{-7} \mathrm{m}\,\mathrm{kg}\,\mathrm{C}^{-2} \\ &\approx 1.25 \times 10^{-6} \mathrm{m}\,\mathrm{kg}\,\mathrm{C}^{-2}. \end{aligned} \tag{1.15}$$

The presence of 4π in this formula isn't telling us anything deep about Nature, but simply reflects a rather outdated way in which this constant was first defined. (We will explain this in more detail in Section 3.6). Nonetheless, this can be thought of as characterising the strength of magnetic interactions (in units of Coulombs).

The Maxwell equations (1.10-1.13) will occupy us for the rest of the course. Rather than trying to understand all the equations at once, we'll proceed bit by bit, looking at situations where only some of the equations are important. By the end of the lectures, we will understand the physics captured by each of these equations and how they fit together.

However, equally importantly, we will also explore the mathematical structure of the Maxwell equations. At first glance, they look just like four random equations from vector calculus. Yet this couldn't be further from the truth. The Maxwell equations are special and, when viewed in the right way, are the essentially unique equations that can describe the force of electromagnetism. The full story of why these are the unique equations involves both quantum mechanics and relativity and will only be told in later courses. But we will start that journey here. The goal is that by the end of these lectures you will be convinced of the importance of the Maxwell equations on both experimental and aesthetic grounds.

Chapter 2

Electrostatics

In this section, we will be interested in electric charges at rest. This means that there exists a frame of reference in which there are no currents; only stationary charges. Of course, there will be forces between these charges but we will assume that the charges are pinned in place and cannot move. The question that we want to answer is: what is the electric field generated by these charges?

Since nothing moves, we are looking for time independent solutions to Maxwell's equations with $\mathbf{J} = \mathbf{0}$. This means that we can consistently set $\mathbf{B} = \mathbf{0}$ and we're left with two of Maxwell's equations to solve. They are

$$\boldsymbol{\nabla} \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \tag{2.1}$$

and

$$\boldsymbol{\nabla} \times \mathbf{E} = 0. \tag{2.2}$$

If you fix the charge distribution ρ, Eqs. (2.1) and (2.2) have a unique solution. Our goal in this section is to find it.

2.1 Gauss' Law

Before we proceed, let's first present equation (2.1) in a slightly different form that will shed some light on its meaning. Consider some closed region $\mathcal{V} \subset \mathbb{R}^3$ of space. We'll denote the boundary of $\mathcal{V}$ by $\mathcal{S} = \partial\mathcal{V}$. We now integrate both sides of (2.1) over $\mathcal{V}$. Since the left-hand side is a total derivative, we can use the divergence theorem to convert this to an integral over the surface $\mathcal{S}$. We have

$$\int_{\mathcal{V}} \mathrm{d}^3x\, \boldsymbol{\nabla} \cdot \mathbf{E} = \int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{1}{\epsilon_0} \int_{\mathcal{V}} \mathrm{d}^3x\, \rho. \tag{2.3}$$

The integral of the charge density over $\mathcal{V}$ is simply the total charge contained in the region. We'll call it $Q = \int \mathrm{d}^3x\, \rho$. Meanwhile, the integral of the electric field over $\mathcal{S}$ is called the *flux* through $\mathcal{S}$. We learn that the two are related by

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{Q}{\epsilon_0} \tag{2.4}$$

This is *Gauss's law*. However, because the two are entirely equivalent, we also refer to the original (2.1) as Gauss's law.

Notice that it doesn't matter what shape the surface $\mathcal{S}$ takes. As long as it surrounds a total charge Q, the flux through the surface will always be Q/ϵ_0. This is shown, for example, in Fig 2.1. A fancy way of saying this is that the integral of the flux doesn't depend on the geometry of the surface, but does depend on its topology since it must

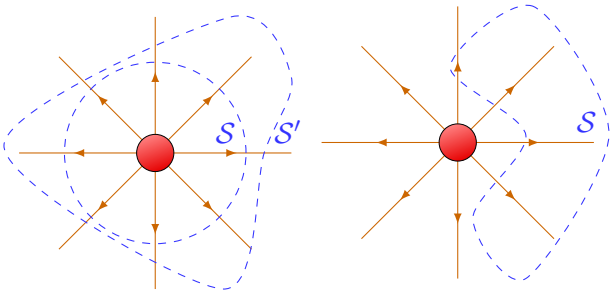


Fig. 2.1: Left: The flux through $\mathcal{S}$ and $\mathcal{S}'$ is the same. Right: The flux through $\mathcal{S}$ vanishes

surround the charge Q. The choice of $\mathcal{S}$ is called the *Gaussian surface*; often there's a smart choice that makes a particular problem simple.

Only charges that lie inside $\mathcal{V}$ contribute to the flux. Any charges that lie outside will produce an electric field that penetrates through $\mathcal{S}$ at some point, giving negative flux, but leaves through the other side of $\mathcal{S}$, depositing positive flux. The total contribution from these charges that lie outside of $\mathcal{V}$ is zero, as illustrated in the right-hand figure above.

For a general charge distribution, we'll need to use both Gauss' law (2.1) and the extra equation (2.2). However, for rather special charge distributions – typically those with lots of symmetry – it turns out to be sufficient to solve the integral form of Gauss' law (2.4) alone, with the symmetry ensuring that (2.2) is automatically satisfied. We start by describing these rather simple solutions. We'll then return to the general case in Section (2.2).

2.1.1 The Coulomb Force

We'll start by showing that Gauss' law (2.4) reproduces the more familiar Coulomb force law that we all know and love. To do this, take a spherically symmetric charge distribution, centered at the origin, contained within some radius R. This will be our model for a particle. We won't need to make any assumption about the nature of the distribution other than its symmetry and the fact that the total charge is Q.

We want to know the electric field at some radius $r > R$. We take our Gaussian surface $\mathcal{S}$ to be a sphere of radius r as shown in Fig. 2.2. Gauss' law states

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{Q}{\epsilon_0}. \tag{2.5}$$

At this point we make use of the spherical symmetry of the problem. This tells us that the electric field must point radially outwards: $\mathbf{E}(\mathbf{x}) = E(r)\hat{\mathbf{r}}$. And, since the integral is only over the angular coordinates of the sphere, we can pull the function $E(r)$ outside. We have

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = E(r) \int_{\mathcal{S}} \hat{\mathbf{r}} \cdot \mathrm{d}\mathbf{S} = E(r) 4\pi r^2 = \frac{Q}{\epsilon_0}, \tag{2.6}$$

where the factor of $4\pi r^2$ has arisen simply because it's the area of the Gaussian sphere.

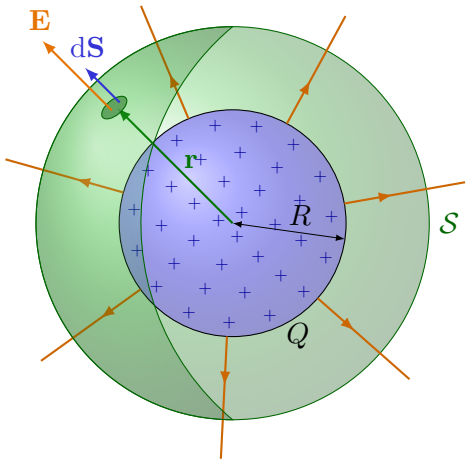


Fig. 2.2: Electric field produced by a spherically symmetric charge distribution, centered at the origin, contained within some radius R.

We learn that the electric field outside a spherically symmetric distribution of charge Q is

$$\mathbf{E}(\mathbf{x}) = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}. \tag{2.7}$$

That's nice. This is the familiar result that we've seen before. The Lorentz force law (1.8) then tells us that a test charge q moving in the region $r > R$ experiences a force

$$\mathbf{F} = \frac{Qq}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}. \tag{2.8}$$

This, of course, is the *Coulomb force* between two static charged particles. Notice that, as promised, $1/\epsilon_0$ characterises the strength of the force. If the two charges have the same sign, so that $Qq > 0$, the force is repulsive, pushing the test charge away from the origin. If the charges have opposite signs, $Qq < 0$, the force is attractive, pointing towards the origin. We see that Gauss's law (2.1) reproduces this simple result that we know about charges.

Finally, note that the assumption of symmetry was crucial in our above analysis. Without it, the electric field $\mathbf{E}(\mathbf{x})$ would have depended on the angular coordinates of the sphere $\mathcal{S}$ and so been stuck inside the integral. In situations without symmetry, Gauss' law alone is not enough to determine the electric field and we need to also use $\boldsymbol{\nabla} \times \mathbf{E} = 0$. We'll see how to do this in Section 2.2. If you're worried, however, it's simple to check that our final expression for the electric field (2.7) does indeed solve $\boldsymbol{\nabla} \times \mathbf{E} = 0$.

2.1.1.1 Coulomb vs Newton

The inverse-square form of the force is common to both electrostatics and gravity. It's worth comparing the relative strengths of the two forces. For example, we can look at the relative strengths of Newtonian attraction and Coulomb repulsion between two electrons. These are point particles with mass me and charge $-e$ given by

$$e \approx 1.6 \times 10^{-19}\text{C} \quad \text{and} \quad m_e \approx 0.1 \times 10^{-31}\text{kg}. \tag{2.9}$$

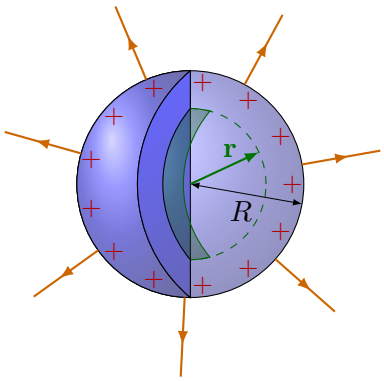


Fig. 2.3: Electric field inside a spherically symmetric charge distribution.

Regardless of the separation, we have

$$\frac{F_{\text{Coulomb}}}{F_{\text{Newton}}} = \frac{e^2}{4\pi\epsilon_0}\frac{1}{Gm_e^2}. \tag{2.10}$$

The strength of gravity is determined by Newton's constant $G \approx 6.7 \times 10^{-11}\text{m}^3\,\text{kg}^{-1}\,\text{s}^2$. Plugging in the numbers reveals something extraordinary:

$$\frac{F_{\text{Coulomb}}}{F_{\text{Newton}}} \approx 10^{42} \tag{2.11}$$

Gravity is puny. Electromagnetism rules. In fact you knew this already. The mere act of lifting up you arm is pitching a few electrical impulses up against the gravitational might of the entire Earth. Yet the electrical impulses win.

However, gravity has a trick up its sleeve. While electric charges come with both positive and negative signs, mass is only positive. It means that by the time we get to macroscopically large objects – stars, planets, cats – the mass accumulates while the charges cancel to good approximation. This compensates the factor of 10^{-42} suppression until, at large distance scales, gravity wins after all.

The fact that the force of gravity is so ridiculously tiny at the level of fundamental particles has consequence. It means that we can neglect gravity whenever we talk about the very small. (And indeed, we shall neglect gravity for the rest of this course). However, it also means that if we would like to understand gravity better on these very tiny distances – for example, to develop a quantum theory of gravity – then it's going to be tricky to get much guidance from experiment.

2.1.2 A Uniform Sphere

The electric field outside a spherically symmetric charge distribution is always given by (2.7). What about inside? This depends on the distribution in question. The simplest is a sphere of radius R with uniform charge distribution ρ (see Fig. 2.3). The total charge is

$$Q = \frac{4\pi}{3}R^3\rho. \tag{2.12}$$

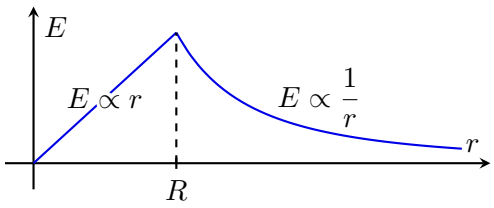


Fig. 2.4: Radial dependence of electric field strength inside and outside a spherically symmetric charge distribution of radius R.

Let's pick our Gaussian surface to be a sphere, centered at the origin, of radius$r < R$. The charge contained within this sphere is $4\pi\rho r^3/3 = Qr^3/R^3$, so Gauss' law gives

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{Qr^3}{\epsilon_0 R^3}. \tag{2.13}$$

Again, using the symmetry argument we can write $\mathbf{E}(\mathbf{r}) = E(r)\hat{\mathbf{r}}$ and compute

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = E(r) \int_{\mathcal{S}} \hat{\mathbf{r}} \cdot \mathrm{d}\mathbf{S} = E(r) 4\pi r^2 = \frac{Qr^3}{\epsilon_0 R^3}. \tag{2.14}$$

This tells us that the electric field grows linearly inside the sphere

$$\mathbf{E}(\mathbf{x}) = \frac{Qr}{4\pi\epsilon_0 R^3}\hat{\mathbf{r}}, \quad r < R. \tag{2.15}$$

Outside the sphere we revert to the inverse-square R r E Figure 7: form (2.7). At the surface of the sphere, $r = R$, the electric field is continuous but the derivative, $\mathrm{d}E/\mathrm{d}r$, is not. This is shown in Fig. 2.4.

2.1.3 Line Charges

Consider, next, a charge smeared out along a line which we'll take to be the z-axis. We'll take uniform charge density η per unit length. (If you like you could consider a solid cylinder with uniform charge density and then send the radius to zero). We want to know the electric field due to this line of charge.

Our set-up now has cylindrical symmetry. We take the Gaussian surface to be a cylinder of length L and radius r (See Fig. 2.5). We have

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{\eta L}{\epsilon_0}. \tag{2.16}$$

Again, by symmetry, the electric field points in the radial direction, away from the line. We'll denote this vector in cylindrical polar coordinates as $\hat{\mathbf{r}}$ so that $\mathbf{E}(\mathbf{r}) = E(r)\hat{\mathbf{r}}$. The symmetry means that the two end caps of the Gaussian surface, $\mathcal{S}_1$ and $\mathcal{S}_2$ in Fig. 2.5, don't contribute to the integral because their normal points parallel to the $\hat{\mathbf{z}}$ direction and $\hat{\mathbf{z}} \cdot \hat{\mathbf{r}} = 0$. We're left only with a contribution from the curved side of the cylinder, $\mathcal{S}_3$,

$$\int_{\mathcal{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = E(r) 2\pi r L = \frac{\eta L}{\epsilon_0}. \tag{2.17}$$

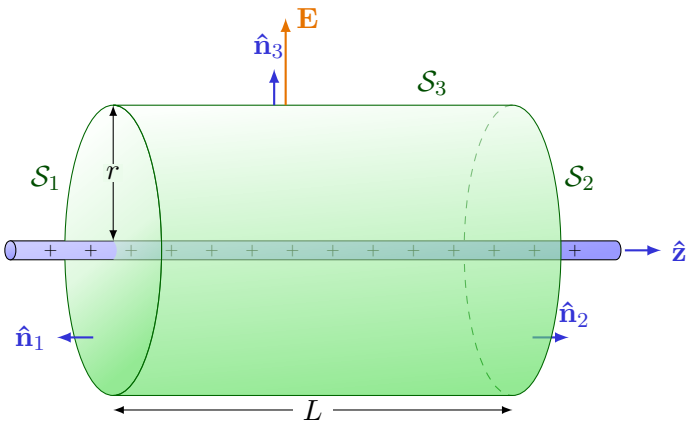


Fig. 2.5: The electric field **E** of an infinite line charge with a uniform linear charge density η. Considering a Gaussian surface in the form of a cylinder at radius r, the electric field has the same magnitude at every point on the surface $\mathcal{S}_3$ and is directed outward, parallel to $\hat{\mathbf{n}}_3$.

So that the electric field is

$$\mathbf{E}(\mathbf{x}) = \frac{\eta}{2\pi\epsilon_0}\hat{\mathbf{r}}. \tag{2.18}$$

Note that, while the electric field for a point charge drops off as $1/r^2$ (with r the radial distance), the electric field for a line charge drops off more slowly as $1/r$. (Of course, the radial distance r means slightly different things in the two cases: it is $r = \sqrt{x^2+y^2+z^2}$ for the point particle, but is $r = \sqrt{x^2+y^2}$ for the line).

2.1.4 Surface Charges and Discontinuities

Now consider an infinite plane, which we take to be $z = 0$, carrying uniform charge per unit area, σ. We again take our Gaussian surface to be a cylinder, this time with its axis perpendicular to the plane as shown in the figure. In this context, the cylinder is sometimes referred to as a Gaussian "pillbox" (on account of Gauss' well known fondness for aspirin). On symmetry grounds, we have

$$\mathbf{E}(\mathbf{r}) = E(z)\hat{\mathbf{z}}. \tag{2.19}$$

Moreover, the electric field in the upper plane, $z > 0$, must point in the opposite direction from the lower plane, $z < 0$, so that $E(z) = -E(-z)$.

The surface integral now vanishes over the curved side of the cylinder and we only get contributions from the end caps, which we take to have area A. This gives

$$\int_{\mathcal{S}} \mathbf{E}\cdot \mathrm{d}\mathbf{S} = E(z)A - E(-z)A = 2E(z) = \frac{\sigma A}{\epsilon_0}. \tag{2.20}$$

The electric field above an infinite plane of charge is therefore

$$\mathbf{E}(\mathbf{x}) = \frac{\sigma}{2\epsilon_0}\hat{\mathbf{z}}. \tag{2.21}$$

Note that the electric field is independent of the distance from the plane! This is because the plane is infinite in extent: the further you move from it, the more comes into view.

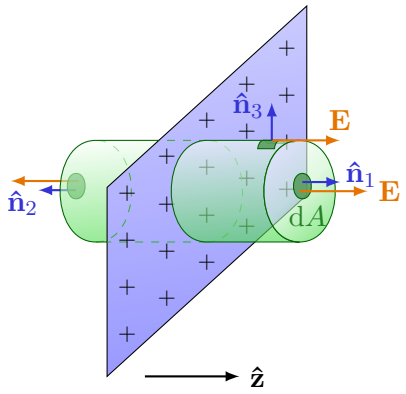


Fig. 2.6: For an infinite sheet of charge with charge density σ, the electric field $\mathbf{E}$ will be perpendicular to the surface. Therefore only the ends of a cylindrical Gaussian surface will contribute to the electric flux. In this case a cylindrical Gaussian surface perpendicular to the charge sheet is used. The resulting field is half that of a conductor at equilibrium with this surface charge density.

There is another important point to take away from this analysis. The electric field is not continuous on either side of a surface of constant charge density. We have

$$E(z \to 0^+) - E(z \to 0^-) = \frac{\sigma}{\epsilon_0}. \tag{2.22}$$

For this to hold, it is not important that the plane stretches to infinity. It's simple to redo the above analysis for any arbitrary surface with charge density σ. There is no need for σ to be uniform and, correspondingly, there is no need for $\mathbf{E}$ at a given point to be parallel to the normal to the surface $\hat{\mathbf{n}}$. At any point of the surface, we can take a Gaussian cylinder, as shown in the left-hand figure above, whose axis is normal to the surface at that point. Its cross-sectional area A can be arbitrarily small (since, as we saw, it drops out of the final answer). If $\mathbf{E}_\pm$ denotes the electric field on either side of the surface, then

$$\hat{\mathbf{n}} \cdot \mathbf{E}_+ - \hat{\mathbf{n}} \cdot \mathbf{E}_- = \frac{\sigma}{\epsilon_0}. \tag{2.23}$$

In contrast, the electric field tangent to the surface is continuous. To see this, we need to do a slightly different calculation. Consider, again, an arbitrary surface with surface charge. Now we consider a loop $\mathcal{C}$ with a length L which lies parallel to the surface and a length a which is perpendicular to the surface. We've drawn this loop in the right-hand figure above, where the surface is now shown side-on. We integrate $\mathbf{E}$ around the loop. Using Stoke's theorem, we have

$$\oint_{\mathcal{C}} \mathbf{E} \cdot \mathrm{d}\mathbf{r} = \int_{\mathcal{S}} \times \mathbf{E} \cdot \mathrm{d}\mathbf{S}\,, \tag{2.24}$$

where $\mathcal{S}$ is the surface bounded by $\mathcal{C}$. In the limit $a \to 0$, the surface $\mathcal{S}$ shrinks to zero size so this integral gives zero. This means that the contribution to line integral must also vanish, leaving us with

$$\hat{\mathbf{n}} \times \mathbf{E}_+ - \hat{\mathbf{n}} \times \mathbf{E}_- = 0. \tag{2.25}$$

This is the statement that the electric field tangential to the surface is continuous.

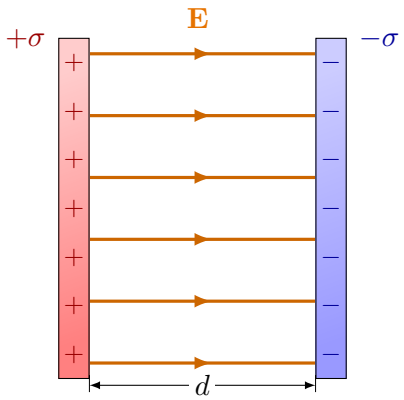


Fig. 2.7: A pair of infinite planes at $z = 0$ and $z = a$, carrying uniform surface charge density $\pm\sigma$

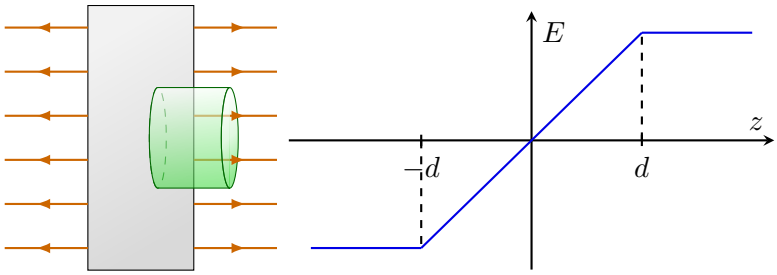


Fig. 2.8: Left: The Gaussian surface for a plane slab. Right: The resulting electric field.

2.1.4.1 A Pair of Planes

As a simple generalisation, consider a pair of infinite planes at $z = 0$ and $z = a$, carrying uniform surface charge density $\pm\sigma$ respectively as shown in Fig. 2.7. To compute the electric field we need only add the fields for arising from two planes, each of which takes the form (2.21). We find that the electric field between the two planes is

$$\mathbf{E} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{z}}, \qquad 0 < z < a, \tag{2.26}$$

while $\mathbf{E} = 0$ outside the planes.

2.1.4.2 A Plane Slab

We can rederive the discontinuity (2.23) in the electric field by considering an infinite slab of thickness $2d$ and charge density per unit volume ρ. When our Gaussian pillbox lies inside the slab, with $z < d$, we have

$$2AE(z) = \frac{2zA\rho}{\epsilon_0} \qquad \Longrightarrow \qquad E(z) = \frac{\rho z}{\epsilon_0}. \tag{2.27}$$

Meanwhile, for $z > d$ we get our earlier result (2.21). The electric field is now continuous as shown in the figure. Taking the limit $d \to 0$ and $\rho \to \infty$ such that the surface charge $\sigma = \rho d$ remains constant reproduces the discontinuity (2.22).

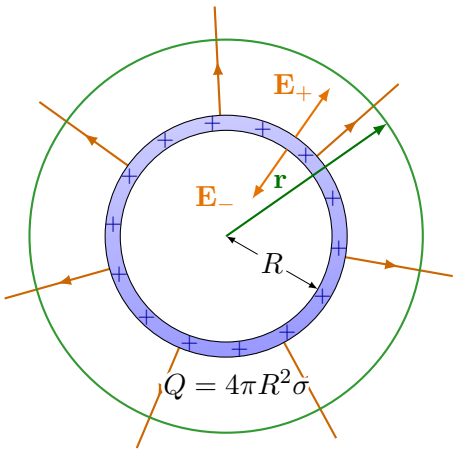


Fig. 2.9: A spherical shell of radius R, centered at the origin, with uniform surface charge density σ.

2.1.4.3 A Spherical Shell

Let's give one last example that involves surface charge and the associated discontinuity of the electric field. We'll consider a spherical shell of radius R, centered at the origin, with uniform surface charge density σ. The total charge is

$$Q = 4\pi R^2\sigma \tag{2.28}$$

We already know that outside the shell, $r > R$, the electric field takes the standard inverse-square form (2.7). What about inside? Well, since any surface with $r < R$ doesn't surround a charge, Gauss' law tells us that we necessarily have $E = 0$ inside. That means that there is a discontinuity at the surface $r = R$,

$$\mathbf{E}\cdot\hat{\mathbf{r}}_+ - \mathbf{E}\cdot\hat{\mathbf{r}}_- = \frac{Q}{4\pi R^2\epsilon_0} = \frac{\sigma}{\epsilon_0}, \tag{2.29}$$

in accord with the expectation (2.23).

2.2 The Electrostatic Potential

For all the examples in the last section, symmetry considerations meant that we only needed to consider Gauss' law. However, for general charge distributions Gauss' law is not sufficient. We also need to invoke the second equation, $\boldsymbol{\nabla}\times\mathbf{E} = 0$.

In fact, this second equation is easily dispatched since $\boldsymbol{\nabla}\times\mathbf{E} = 0$ implies that the electric field can be written as the gradient of some function,

$$\mathbf{E} = -\boldsymbol{\nabla}\phi. \tag{2.30}$$

The scalar ϕ is called the *electrostatic potential* or *scalar potential* (or, sometimes, just the *potential*). To proceed, we revert to the original differential form of Gauss' law (2.1). This now takes the form of the *Poisson equation*

$$\boldsymbol{\nabla}\cdot\mathbf{E} = \frac{\rho}{\epsilon_0} \quad\Longrightarrow\quad \nabla^2\phi = -\frac{\rho}{\epsilon_0}. \tag{2.31}$$

In regions of space where the charge density vanishes, we're left solving the Laplace equation

$$\nabla^2\phi = 0. \tag{2.32}$$

Solutions to the Laplace equation are said to be *harmonic* functions.

A few comments:

- The potential ϕ is only defined up to the addition of some constant. This seemingly trivial point is actually the beginning of a long and deep story in theoretical physics known as *gauge invariance*. We'll come back to it in Subsection 5.0.1. For now, we'll eliminate this redundancy by requiring that $\phi(r) \to 0$ as $r \to \infty$.
- We know from our study of Newtonian mechanics that the electrostatic potential is proportional to the potential energy experienced by a test particle. Specifically, a test particle of mass m, position $\mathbf{r}(t)$ and charge q moving in a background electric field has conserved energy
$$E = \frac{1}{2}m\dot{\mathbf{r}}\cdot\dot{\mathbf{r}} + q\phi(\mathbf{r}). \tag{2.33}$$
- The Poisson equation is linear in both ϕ and ρ. This means that if we know the potential ϕ_1 for some charge distribution ρ_1 and the potential ϕ_1 for another charge distribution ρ_2, then the potential for $\rho_1 + \rho_2$ is simply $\phi_1 + \rho_2$. What this really means is that the electric field for a bunch of charges is just the sum of the fields generated by each charge. This is called the *principle of superposition* for charges. This linearity of the equations is what makes electromagnetism easy compared to other forces of Nature.
- We stated above that $\boldsymbol{\nabla}\times\mathbf{E} = 0$ is equivalent to writing $\mathbf{E} = -\boldsymbol{\nabla}\phi$. This is true when space is $\mathbb{R}^3$ or, in fact, if we take space to be any open ball in $\mathbb{R}^3$. But if our background space has a suitably complicated topology then there are solutions to $\boldsymbol{\nabla}\times\mathbf{E} = 0$ which cannot be written in the form $\mathbf{E} = -\boldsymbol{\nabla}\phi$. This is tied ultimately to the beautiful mathematical theory of de Rham cohomology. Needless to say, in this starter course we're not going to worry about these issues. We'll always take spacetime to have topology $\mathbb{R}^4$ and, correspondingly, any spatial hypersurface to be $\mathbb{R}^3$.

2.2.1 The Point Charge

Let's start by deriving the Coulomb force law yet again. We'll take a particle of charge Q and place it at the origin. This time, however, we'll assume that the particle really is a point charge. This means that the charge density takes the form of a delta-function, $\rho(\mathbf{x}) = Q\delta^3(\mathbf{x})$. We need to solve the equation

$$\nabla^2\phi = -\frac{Q}{\epsilon_0}\delta^3(\mathbf{x}). \tag{2.34}$$

The solution is essentially the Green's function (See Appendix A.1) for the Laplacian ∇^2, an interpretation that we'll return to in Subsection 2.2.3. Let's recall how we find this solution. We first look away from the origin, $r \neq 0$, where there's no funny business going

on with delta-function. Here, we're looking for the spherically symmetric solution to the Laplace equation. This is

$$\phi = \frac{\alpha}{r}, \tag{2.35}$$

or some constant α. To see why this solves the Laplace equation, we need to use the result

$$\boldsymbol{\nabla} r = \hat{\mathbf{r}} \tag{2.36}$$

where $\hat{\mathbf{r}}$ is the unit radial vector in spherical polar coordinates, so $\mathbf{x} = r\hat{\mathbf{r}}$. Using the chain rule, this means that $\boldsymbol{\nabla}(1/r) = -\hat{\mathbf{r}}/r^2 = -\mathbf{x}/r^3$. This gives us

$$\boldsymbol{\nabla}\phi = -\frac{\alpha}{r^3}\mathbf{x} \quad \Longrightarrow \quad \nabla^2\phi = -\alpha\left(\frac{\boldsymbol{\nabla}\cdot\mathbf{x}}{r^3} - \frac{3\mathbf{x}\cdot\mathbf{x}}{r^5}\right). \tag{2.37}$$

But $\boldsymbol{\nabla}\cdot\mathbf{x} = 3$ and we find that $\nabla^2\phi = 0$ as required.

It remains to figure out what to do at the origin where the delta-function lives. This is what determines the overall normalisation α of the solution. At this point, it's simplest to use the integral form of Gauss' law to transfer the problem from the origin to the far flung reaches of space. To do this, we integrate (2.34) over some region $\mathcal{V}$ which includes the origin. Integrating the charge density gives

$$\rho(\mathbf{x}) = Q\delta^3(\mathbf{x}) \quad \Longrightarrow \quad \int_{\mathcal{V}} \mathrm{d}^3x\, \rho = Q. \tag{2.38}$$

So, using Gauss' law (2.4), we require

$$\int_{\mathcal{S}} \boldsymbol{\nabla}\phi \cdot \mathrm{d}\mathbf{S} = -\frac{Q}{\epsilon_0}. \tag{2.39}$$

But this is exactly the kind of surface integral that we were doing in the last section. Substituting $\phi = \alpha/r$ into the above equation, and choosing $\mathcal{S}$ to be a sphere of radius r, tells us that we must have $\alpha = Q/4\pi\epsilon_0$, or

$$\phi = \frac{Q}{4\pi\epsilon_0 r}. \tag{2.40}$$

Taking the gradient of this using (2.36) gives us Coulomb's law

$$\mathbf{E}(\mathbf{x}) = -\boldsymbol{\nabla}\phi = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}. \tag{2.41}$$

The derivation of Coulomb's law using the potential was somewhat more involved than the technique using Gauss' law alone that we saw in the last section. However, as we'll now see, introducing the potential allows us to write down the solution to essentially any problem.

A Note on Notation Throughout these lectures, we will use $\mathbf{x}$ and $\mathbf{r}$ interchangeably to denote position in space. For example, sometimes we'll write integration over a volume as $\int \mathrm{d}^3x$ and sometimes as $\int \mathrm{d}^3r$. The advantage of the r notation is that it looks more natural when working in spherical polar coordinates. For example, we have $|\mathbf{r}| = r$ which is nice. The disadvantage is that it can lead to confusion when working in other coordinate systems, in particular cylindrical polar. For this reason, we'll alternate between the two notations, adopting the attitude that clarity is more important than consistency.

2.2.2 The Dipole

A *dipole* consists of two point charges, Q and $-Q$, a distance d apart. We place the first charge at the origin and the second at $\mathbf{r} = -\mathbf{D}$. The potential is simply the sum of the potential for each charge,

$$\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{r} - \frac{Q}{|\mathbf{r}+\mathbf{D}|}\right). \tag{2.42}$$

Similarly, the electric field is just the sum of the electric fields made by the two point charges. This follows from the linearity of the equations and is a simple application of the principle of superposition that we mentioned earlier.

It will prove fruitful to ask what the dipole looks like far from the two point charges, at a distance $r \gg |\mathbf{D}|$. We need to Taylor expand the second term above. The vector version of the Taylor expansion for a general function $f(\mathbf{r})$ is given by

$$f(\mathbf{r}+\mathbf{D}) \approx f(\mathbf{r}) + \mathbf{D}\cdot\boldsymbol{\nabla} f(\mathbf{r}) + \frac{1}{2}(\mathbf{D}\cdot\boldsymbol{\nabla})^2 f(\mathbf{r}) + \cdots . \tag{2.43}$$

Applying this to the function $1/|\mathbf{r}+\mathbf{D}|$ gives

$$\begin{aligned}\frac{1}{|\mathbf{r}+\mathbf{D}|} &\approx \frac{1}{r} + \mathbf{D}\cdot\boldsymbol{\nabla}\frac{1}{r} + \frac{1}{2}(\mathbf{D}\cdot\boldsymbol{\nabla})^2\frac{1}{r} + \cdots \\ &= \frac{1}{r} - \frac{\mathbf{D}\cdot\mathbf{r}}{r^3} - \frac{1}{2}\left(\frac{\mathbf{D}\cdot\mathbf{D}}{r^3} - \frac{3(\mathbf{D}\cdot\mathbf{r})^2}{r^5}\right) + \cdots\end{aligned} \tag{2.44}$$

(To derive the last term, it might be easiest to use index notation for $\mathbf{D}\cdot\boldsymbol{\nabla} = d_i\partial_i$). For our dipole, we'll only need the first two terms in this expansion. They give the potential

$$\phi \approx \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{r} - \frac{1}{r} + \mathbf{D}\cdot\boldsymbol{\nabla}\frac{1}{r} + \cdots\right) = \frac{Q}{4\pi\epsilon_0}\frac{\mathbf{D}\cdot\mathbf{r}}{r^3} + \cdots . \tag{2.45}$$

We see that the potential for a dipole falls off as $1/r^2$. Correspondingly, the electric field drops off as $1/r^3$; both are one power higher than the fields for a point charge.

The electric field is not spherically symmetric. The leading order contribution is governed by the combination

$$\mathbf{P} = Q\mathbf{D}. \tag{2.46}$$

This is called the electric *dipole moment*. By convention, it points from the negative charge to the positive. The dipole electric field is

$$\mathbf{E} = -\boldsymbol{\nabla}\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{3(\mathbf{P}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{P}}{r^3}\right) + \cdots . \tag{2.47}$$

Notice that the sign of the electric field depends on where you sit in space. In some parts, the force will be attractive; in other parts repulsive.

It's sometimes useful to consider the limit $d \to 0$ and $Q \to \infty$ such that $\mathbf{P} = Q\mathbf{D}$ remains fixed. In this limit, all the $\cdots$ terms in (2.45) and (2.47) disappear since they contain higher powers of d. Often when people talk about the "dipole", they implicitly mean taking this limit.

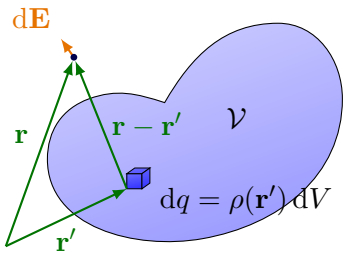


Fig. 2.10: Electric field at $\mathbf{r}$ with $|\mathbf{r}| \gg |\mathbf{r}'|$.

2.2.3 General Charge Distributions

Our derivation of the potential due to a point charge (2.40), together with the principle of superposition, is actually enough to solve – at least formally – the potential due to any charge distribution. This is because the solution for a point charge is nothing other than the Green's function for the Laplacian. The Green's function is defined to be the solution to the equation

$$\phi(\mathbf{r}) = -\frac{1}{\epsilon_0}\int_{\mathcal{V}} \mathrm{d}^3r'\, G(\mathbf{r};\mathbf{r}')\rho(\mathbf{r}') = \frac{1}{4\pi\epsilon_0}\int_{\mathcal{V}} \mathrm{d}^3r'\, \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \tag{2.48}$$

(To check this, you just have to keep your head and remember whether the operators are hitting $\mathbf{r}$ or $\mathbf{r}'$. The Laplacian acts on $\mathbf{r}$ so, if we compute $\nabla^2\phi$, it passes through the integral in the above expression and hits $G(\mathbf{r};\mathbf{r}')$, leaving behind a delta-function which subsequently kills the integral).

Similarly, the electric field arising from a general charge distribution is

$$\begin{aligned}\mathbf{E}(\mathbf{r}) = -\boldsymbol{\nabla}\phi(\mathbf{r}) &= -\frac{1}{4\pi\epsilon_0}\int_{\mathcal{V}} \mathrm{d}^3r'\, \rho(\mathbf{r}')\boldsymbol{\nabla}\frac{1}{|\mathbf{r}-\mathbf{r}'|} \\ &= -\frac{1}{4\pi\epsilon_0}\int_{\mathcal{V}} \mathrm{d}^3r'\, \rho(\mathbf{r}')\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}.\end{aligned} \tag{2.49}$$

Given a very complicated charge distribution $\rho(\mathbf{r})$, this equation will give back an equally complicated electric field $\mathbf{E}(\mathbf{r})$. But if we sit a long way from the charge distribution, there's a rather nice simplification that happens.

2.2.3.1 Long Distance Behaviour

Suppose now that you want to know what the electric field looks like far from the region $\mathcal{V}$. This means that we're interested in the electric field at $\mathbf{r}$ with $|\mathbf{r}| \gg |\mathbf{r}'|$ for all $\mathbf{r}' \in \mathcal{V}$. We can apply the same Taylor expansion (2.43), now replacing $\mathbf{D}$ with $-\mathbf{r}'$ for each $\mathbf{r}'$ in the charged region. This means we can write

$$\begin{aligned}frac1|\mathbf{r}-\mathbf{r}'| &= \frac{1}{r} - \mathbf{r}'\cdot\boldsymbol{\nabla}\frac{1}{r} + \frac{1}{2}(\mathbf{r}\cdot\boldsymbol{\nabla})^2\frac{1}{r} + \cdots \\ &= \frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \frac{1}{2}\left(\frac{3(\mathbf{r}\cdot\mathbf{r}')^2}{r^5} - \frac{\mathbf{r}'\cdot\mathbf{r}'}{r^3}\right) + \cdots,\end{aligned} \tag{2.50}$$

and our potential becomes

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_{\mathcal{V}} \mathrm{d}^3r'\,\rho(\mathbf{r}')\left(\frac{1}{r} + \frac{\mathbf{r}'\cdot\mathbf{r}'}{r^3} + \cdots\right). \tag{2.51}$$

The leading term is just

$$\phi(\mathbf{r}) = \frac{Q}{4\pi\epsilon_0} + \cdots, \tag{2.52}$$

where $Q = \int_{\mathcal{V}} \mathrm{d}^3r'\,rho(\mathbf{r}')$ is the total charge contained within $\mathcal{V}$. So, to leading order, if you're far enough away then you can't distinguish a general charge distribution from a point charge localised at the origin. But if you're careful with experiments, you can tell the difference. The first correction takes the form of a dipole,

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{r} + \frac{\mathbf{P}\cdot\hat{\mathbf{r}}}{r^2} + \cdots\right), \tag{2.53}$$

where

$$\mathbf{P} = \int_{\mathcal{V}} \mathrm{d}^3r'\,\mathbf{r}'\rho(\mathbf{r}') \tag{2.54}$$

is the dipole moment of the distribution. One particularly important situation is when we have a neutral object with $Q = 0$. In this case, the dipole is the dominant contribution to the potential.

We see that an arbitrarily complicated, localised charge distribution can be characterised by a few simple quantities, of decreasing importance. First comes the total charge Q. Next the dipole moment $\mathbf{P}$ which contains some basic information about how the charges are distributed. But we can keep going. The next correction is called the quadrupole and is given by

$$\Delta\phi = \frac{1}{2}\frac{1}{4\pi\epsilon_0}\frac{r_i r_j \mathbb{Q}_{ij}}{r^5}, \tag{2.55}$$

where $\mathbb{Q}_{ij}$ is a symmetric traceless tensor known as the quadrupole moment, given by

$$\mathbb{Q}_{ij} = \int_{\mathcal{V}} \mathrm{d}^3r'\,\rho(\mathbf{r}')\left(3r'_i r'_j - \delta_{ij}r'^2\right) \tag{2.56}$$

It contains some more refined information about how the charges are distributed. After this comes the octopole and so on. The general name given to this approach is the *multipole expansion*. It involves expanding the function ϕ in terms of spherical harmonics. A systematic treatment can be found, for example, in the book by Jackson.

2.2.3.2 A Comment on Infinite Charge Distribution

In the above, we assumed for simplicity that the charge distribution was restricted to some compact region of space, $\mathcal{V}$. The Green's function approach still works if the charge distribution stretches to infinity. However, for such distributions it's not always possible to pick $\phi(\mathbf{r}) \to 0$ as $r \to \infty$. In fact, we saw an example of this earlier. For an infinite line charge of density η, we computed the electric field in (2.18). It goes as

$$E(\mathbf{x}) = \frac{\eta}{2\pi\epsilon_0}\hat{\mathbf{r}}, \tag{2.57}$$

where now $r^2 = x^2 + y^2$ is the cylindrical radial coordinate perpendicular to the line. The potential ϕ which gives rise to this is

$$\phi(r) = -\frac{\eta}{2\pi\epsilon_0}\log\left(\frac{r}{r_0}\right). \tag{2.58}$$

Because of the log function, we necessarily have $\phi(r) \to 0$ as $r \to \infty$. Instead, we need to pick an arbitrary, but finite distance, r_0 at which the potential vanishes.

2.2.4 Field Lines

The usual way of depicting a vector is to draw an arrow whose length is proportional to the magnitude. For the electric field, there's a slightly different, more useful way to show what's going on. We draw continuous lines, tangent to the electric field $\mathbf{E}$, with the density of lines proportional to the magnitude of $\mathbf{E}$. This innovation, due to Faraday, is called the *field line*. (They are what we have been secretly drawing throughout these notes).

Field lines are continuous. They begin and end only at charges. They can never cross.

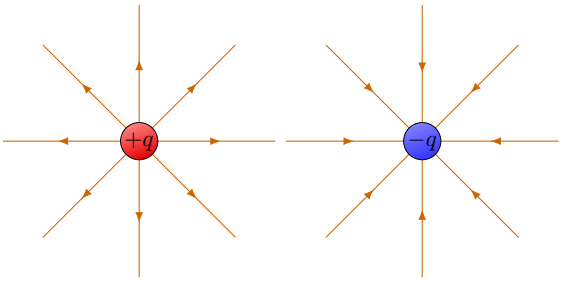


Fig. 2.11: Field lines for positive and negative point charges

By convention, the positive charges act as sources for the lines, with the arrows emerging. The negative charges act as sinks, with the arrows approaching.

It's also easy to draw the equipotentials – surfaces of constant ϕ – on this same figure. These are the surfaces along which you can move a charge without doing any work. The relationship $\mathbf{E} = -\boldsymbol{\nabla}\phi$ ensures that the equipotentials cut the field lines at right angles. We usually draw them as dotted lines.

2.2.5 Electrostatic Equilibrium

Here's a simple question: can you trap an electric charge using only other charges? In other words, can you find some arrangements of charges such that a test charge sits in stable equilibrium, trapped by the fields of the others.

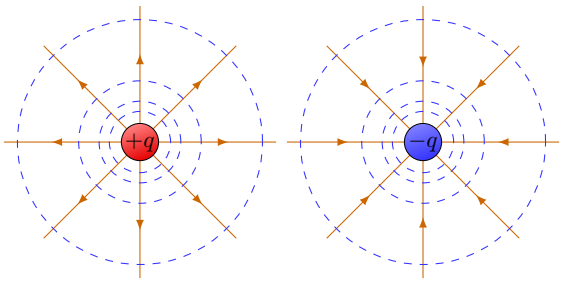


Fig. 2.12: Equipotentials for positive and negative point charges

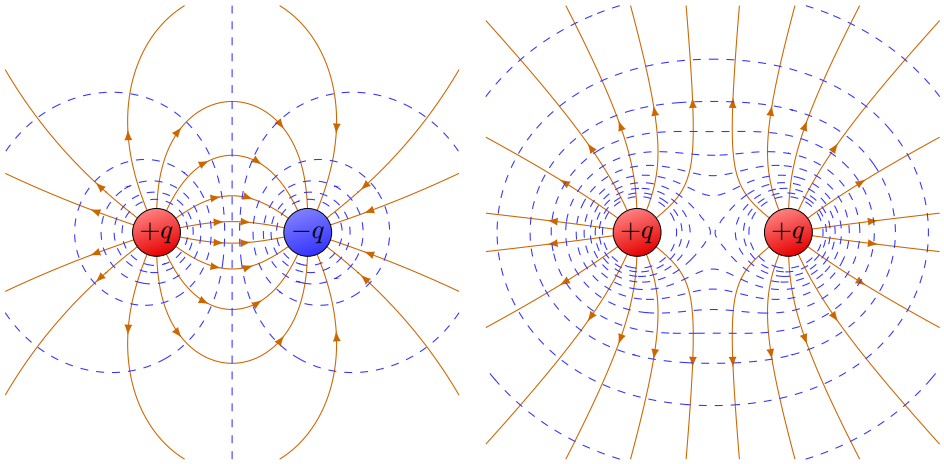


Fig. 2.13: Field lines and equipotentials for the dipole (left) and for a pair of charges of the same sign (on the right).

There's a trivial way to do this: just allow a negative charge to sit directly on top of a positive charge. But let's throw out this possibility. We'll ask that the equilibrium point lies away from all the other charges.

There are some simple set-ups that spring to mind that might achieve this. Maybe you could place four positive charges at the vertices of a pyramid; or perhaps 8 positive charges at the corners of a cube. Is it possible that a test positive charge trapped in the middle will be stable? It's certainly repelled from all the corners, so it might seem plausible.

The answer, however, is no. There is no electrostatic equilibrium. You cannot trap an electric charge using only other stationary electric charges, at least not in a stable manner. Since the potential energy of the particle is proportional to ϕ, mathematically, this is the statement that a harmonic function, obeying $\nabla^2\phi = 0$, can have no minimum or maximum.

To prove that there can be no electrostatic equilibrium, let's suppose the opposite: that there is some point in empty space $\mathbf{r}_*$ that is stable for a particle of charge $q > 0$. By "empty space", we mean that $\rho(\mathbf{r}) = 0$ in a neighbourhood of $\mathbf{r}_*$. Because the point is stable, if the particle moves away from this point then it must always be pushed back. This, in turn, means that the electric field must always point inwards towards the point

$\mathbf{r}_*$; never away. We could then surround $\mathbf{r}_*$ by a small surface $\mathcal{S}$ and compute

$$\int_{\mathcal{S}} \mathbf{E} \cdot d\mathbf{S} < 0. \tag{2.59}$$

But, by Gauss' law, the right-hand side must be the charge contained within $\mathcal{S}$ which, by assumption, is zero. This is our contradiction: electrostatic equilibrium does not exist.

Of course, if you're willing to use something other than electrostatic forces then you can construct equilibrium situations. For example, if you restrict the test particle to lie on a plane then it's simple to check that equal charges placed at the corners of a polygon will result in a stable equilibrium point in the middle. But to do this you need to use other forces to keep the particle in the plane in the first place.

2.3 Electrostatic Energy

There is energy stored in the electric field. In this section, we calculate how much.

Suppose we have some test charge q moving in a background electrostatic potential ϕ. We'll denote the potential energy of the particle as $U(\mathbf{r})$. The potential $U(\mathbf{r})$ of the particle can be thought of as the work done bringing the particle in from infinity;

$$U(\mathbf{r}) = -\int_{\infty}^{r} \mathbf{F} \cdot d\mathbf{r} = +q\int_{\infty}^{r} \boldsymbol{\nabla}\phi \cdot d\mathbf{r} = q\phi(\mathbf{r}), \tag{2.60}$$

where we've assumed our standard normalisation of $\phi(\mathbf{r}) \to 0$ as $r \to \infty$.

Consider a distribution of charges which, for now, we'll take to be made of point charges q_i at positions $\mathbf{r}_i$. The electrostatic potential energy stored in this configuration is the same as the work required to assemble the configuration in the first place. (This is because if you let the charges go, this is how much kinetic energy they will pick up). So how much work does it take to assemble a collection of charges?

Well, the first charge is free. In the absence of any electric field, you can just put it where you like - say, $\mathbf{r}_1$. The work required is $W_1 = 0$.

To place the second charge at $\mathbf{r}_2$ takes work

$$W_2 = \frac{q_1 q_2}{4\pi\epsilon_0} \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}. \tag{2.61}$$

Note that if the two charges have the same sign, so $q_1 q_2 > 0$, then $W_2 > 0$ which is telling us that we need to put work in to make them approach. If $q_1 q_2 < 0$ then $W_2 < 0$ where the negative work means that the particles wanted to be drawn closer by their mutual attraction.

The third charge has to battle against the electric field due to both q_1 and q_2. The work required is

$$W_3 = \frac{q_3}{4\pi\epsilon_0}\left(\frac{q_2}{|\mathbf{r}_2 - \mathbf{r}_2|} + \frac{q_1}{|\mathbf{r}_1 - \mathbf{r}_3|}\right), \tag{2.62}$$

and so on. The total work needed to assemble all the charges is the potential energy stored in the configuration,

$$U = \sum_{i=1}^{N} W_i = \frac{1}{4\pi\epsilon_0} \sum_{i<j} \frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{2.63}$$

where $\sum_{i<j}$ means that we sum over each pair of particles once. In fact, you probably could have just written down (2.63) as the potential energy stored in the configuration. The whole purpose of the above argument was really just to nail down a factor of 1/2: do we sum over all pairs of particles $\sum_{i<j}$ or all particles $\sum_{i\neq j}$? The answer, as we have seen, is all pairs.

We can make that factor of 1/2 even more explicit by writing

$$U = \frac{1}{2}\frac{1}{4\pi\epsilon_0} \sum_i \sum_{j\neq i} \frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{2.64}$$

where now we sum over each pair twice.

There is a slicker way of writing (2.64). The potential at $\mathbf{r}_i$ due to all the other charges q_j, $j \neq i$ is

$$\phi(\mathbf{r}_i) = \frac{1}{4\pi\epsilon_0} \sum_{i\neq j} \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.65}$$

which means that we can write the potential energy as

$$U = \frac{1}{2} \sum_{i=1}^{N} q_i \phi(\mathbf{r}_i). \tag{2.66}$$

This is the potential energy for a set of point charges. But there is an obvious generalisation to charge distributions $\rho(\mathbf{r})$. We'll again assume that $\rho(\mathbf{r})$ has compact support so that the charge is localised in some region of space. The potential energy associated to such a charge distribution should be

$$U = \frac{1}{2} \int \mathrm{d}^3 r\, \rho(\mathbf{r})\phi(\mathbf{r}), \tag{2.67}$$

where we can quite happily take the integral over all of $\mathbb{R}^3$, safe in the knowledge that anywhere that doesn't contain charge has $\rho(\mathbf{r}) = 0$ and so won't contribute.

Now this is in a form that we can start to play with. We use Gauss' law to rewrite it as

$$U = \frac{\epsilon_0}{2} \int \mathrm{d}^3 r\, (\boldsymbol{\nabla}\cdot\mathbf{E})\phi = \frac{\epsilon_0}{2} \int \mathrm{d}^3 r\, [\boldsymbol{\nabla}\cdot(\mathbf{E}\phi) - \mathbf{E}\cdot\boldsymbol{\nabla}\phi]. \tag{2.68}$$

But the first term is a total derivative. And since we're taking the integral over all of space and $\phi(\mathbf{r}) \to \infty$ as $r \to \infty$, this term just vanishes. In the second term we can replace $\boldsymbol{\nabla}\phi = -\mathbf{E}$. We find that the potential energy stored in a charge distribution has an elegant expression solely in terms of the electric field that it creates,

$$U = \frac{\epsilon_0}{2} \int \mathrm{d}^3 r\, \mathbf{E}\cdot\mathbf{E}. \tag{2.69}$$

Isn't that nice!

2.3.1 The Energy of a Point Particle

There is a subtlety in the above derivation. In fact, I totally tried to pull the wool over your eyes. Here it's time to own up.

First, let me say that the final result (2.69) is right: this is the energy stored in the electric field. But the derivation above was dodgy. One reason to be dissatisfied is that we computed the energy in the electric field by equating it to the potential energy stored in a charge distribution that creates this electric field. But the end result doesn't depend on the charge distribution. This suggests that there should be a more direct way to arrive at (2.69) that only talks about fields and doesn't need charges. And there is. We will see it later.

But there is also another, more worrying problem with the derivation above. To illustrate this, let's just look at the simplest situation of a point particle. This has electric field

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}. \tag{2.70}$$

So, by (2.69), the associated electric field should carry energy. But we started our derivation above by assuming that a single particle didn't carry any energy since it didn't take any work to put the particle there in the first place. What's going on?

Well, there was something of a sleight of hand in the derivation above. This occurs when we went from the expression $q\phi$ in (2.66) to $\rho\phi$ in (2.67). The former omits the "self-energy" terms; there is no contribution arising from $q_i\phi(\mathbf{r}_i)$. However, the latter includes them. The two expressions are not quite the same. This is also the reason that our final expression for the energy (2.69) is manifestly positive, while $q\phi$ can be positive or negative.

So which is right? Well, which form of the energy you use rather depends on the context. It is true that (2.69) is the correct expression for the energy stored in the electric field. But it is also true that you don't have to do any work to put the first charge in place since we're obviously not fighting against anything. Instead, the "self-energy" contribution coming from $\mathbf{E}\cdot\mathbf{E}$ in (2.70) should simply be thought of – using $E = mc^2$ – as a contribution to the mass of the particle.

We can easily compute this contribution for, say, an electron with charge $q = -e$. Let's call the radius of the electron a. Then the energy stored in its electric field is

$$\text{Energy} = \frac{\epsilon_0}{2}\int \mathrm{d}^3r\, \mathbf{E}\cdot\mathbf{E} = \frac{e^2}{32\pi\epsilon_0}\int_a^\infty \mathrm{d}r\, \frac{4\pi r^2}{r^4} = \frac{e^2}{8\pi\epsilon_0}\frac{1}{a}. \tag{2.71}$$

We see that, at least as far as the energy is concerned, we'd better not treat the electron as a point particle with $a \to 0$ or it will end up having infinite mass. And that will make it really hard to move.

So what is the radius of an electron? For the above calculation to be consistent, the energy in the electric field can't be greater than the observed mass of the electron m_e. In other words, we'd better have

$$m_e c^2 > \frac{e^2}{8\pi\epsilon_0}\frac{1}{a} \quad\Longrightarrow\quad a > \frac{e^2}{8\pi\epsilon_0}\frac{1}{m_e c^2}. \tag{2.72}$$

That, at least, puts a bound on the radius of the electron, which is the best we can do using classical physics alone. To give a more precise statement of the radius of the electron, we need to turn to quantum mechanics.

2.3.1.1 A Quick Foray into Quantum Electrodynamics

To assign a meaning of "radius" to seemingly point-like particles, we really need the machinery of quantum field theory. In that context, the size of the electron is called its *Compton wavelength*. This is the distance scale at which the electron gets surrounded by a swarm of electron-positron pairs which, roughly speaking, smears out the charge distribution. This distance scale is

$$a = \frac{\hbar}{m_e c}. \tag{2.73}$$

We see that the inequality (2.72) translates into an inequality on a bunch of fundamental constants. For the whole story to hang together, we require

$$\frac{e^2}{8\pi\epsilon_0\hbar c} < 1. \tag{2.74}$$

This is an almost famous combination of constants. It's more usual to define the combination

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}. \tag{2.75}$$

This is known as the *fine structure constant*. It is dimensionless and takes the value

$$\alpha \approx \frac{1}{137}. \tag{2.76}$$

Our discussion above requires $\alpha < 2$. We see that Nature happily meets this requirement.

2.3.2 The Force Between Electric Dipoles

As an application of our formula for electrostatic energy, we can compute the force between two, far separated dipoles. We place the first dipole, $\mathbf{P}_1$, at the origin. It gives rise to a potential

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\frac{\mathbf{P}_1\cdot\mathbf{r}}{r^3}. \tag{2.77}$$

Now, at some distance away, we place a second dipole. We'll take this to consist of a charge Q at position $\mathbf{r}$ and a charge $-Q$ at position $\mathbf{r}-\mathbf{D}$, with $d \ll r$. The resulting dipole moment is $\mathbf{P}_2 = Q\mathbf{D}$. We're not interested in the energy stored in each individual dipole; only in the potential energy needed to bring the two dipoles together. This is given by (2.63),

$$\begin{aligned} U = Q(\phi(\mathbf{r}) - \phi(\mathbf{r}-\mathbf{D})) &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{P}_1\cdot\mathbf{r}}{r^3} - \frac{\mathbf{P}_1\cdot(\mathbf{r}-\mathbf{D})}{|\mathbf{r}-\mathbf{D}|^3}\right) \\ &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{P}_1\cdot\mathbf{r}}{r^3} - \mathbf{P}_1\cdot(\mathbf{r}-\mathbf{D})\left(\frac{1}{r^3} + \frac{3\mathbf{D}\cdot\mathbf{r}}{r^5} + \cdots\right)\right) \\ &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{P}_1\cdot\mathbf{D}}{r^3} - \frac{3(\mathbf{P}_1\cdot\mathbf{r})(\mathbf{D}\cdot\mathbf{r})}{r^5}\right), \end{aligned} \tag{2.78}$$

where, to get to the second line, we've Taylor expanded the denominator of the second term. This final expression can be written in terms of the second dipole moment. We find the nice, symmetric expression for the potential energy of two dipoles separated by distance $\mathbf{r}$,

$$U = \frac{1}{4\pi\epsilon_0}\left(\frac{\mathbf{P}_1\cdot\mathbf{P}_2}{r^3} - \frac{3(\mathbf{P}_1\cdot\mathbf{r})(\mathbf{P}_2\cdot\mathbf{r})}{r^5}\right). \tag{2.79}$$

But, we know from our first course on dynamics that the force between two objects is just given by $F = -\boldsymbol{\nabla}U$. We learn that the force between two dipoles is given by

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\boldsymbol{\nabla}\left(\frac{3(\mathbf{P}_1\cdot\mathbf{r})(\mathbf{P}_2\cdot\mathbf{r})}{r^5} - \frac{\mathbf{P}_1\cdot\mathbf{P}_2}{r^3}\right). \tag{2.80}$$

The strength of the force, and even its sign, depends on the orientation of the two dipoles. If $\mathbf{P}_1$ and $\mathbf{P}_2$ lie parallel to each other and to $\mathbf{r}$ then the resulting force is attractive. If $\mathbf{P}_1$ and $\mathbf{P}_2$ point in opposite directions, and lie parallel to $\mathbf{r}$, then the force is repulsive. The expression above allows us to compute the general force.

2.4 Conductors

Let's now throw something new into the mix. A *conductor* is a region of space which contains charges that are free to move. Physically, think "metal". We want to ask what happens to the story of electrostatics in the presence of a conductor. There are a number of things that we can say straight away:

- Inside a conductor we must have $\mathbf{E} = 0$. If this isn't the case, the charges would move. But we're interested in electrostatic situations where nothing moves.
- Since $\mathbf{E} = 0$ inside a conductor, the electrostatic potential ϕ must be constant throughout the conductor.
- Since $\mathbf{E} = 0$ and $\boldsymbol{\nabla}\cdot\mathbf{E} = \rho/\epsilon_0$, we must also have $\rho = 0$. This means that the interior of the conductor can't carry any charge.
- Conductors can be neutral, carrying both positive and negative charges which balance out. Alternatively, conductors can have net charge. In this case, any net charge must reside at the surface of the conductor.
- Since ϕ is constant, the surface of the conductor must be an equipotential. This means that any $\mathbf{E} = -\boldsymbol{\nabla}\phi$ is perpendicular to the surface. This also fits nicely with the discussion above since any component of the electric field that lies tangential to the surface would make the surface charges move.
- If there is surface charge σ anywhere in the conductor then, by our previous discontinuity result (2.23), together with the fact that $\mathbf{E} = 0$ inside, the electric field just outside the conductor must be

$$\mathbf{E} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{n}}. \tag{2.81}$$

Problems involving conductors are of a slightly different nature than those we've discussed up to now. The reason is that we don't know from the start where the charges are, so we don't know what charge distribution ρ that we should be solving for. Instead, the electric fields from other sources will cause the charges inside the conductor to shift around until they reach equilibrium in such a way that $\mathbf{E} = 0$ inside the conductor. In general, this will mean that even neutral conductors end up with some surface charge, negative in some areas, positive in others, just enough to generate an electric field inside the conductor that precisely cancels that due to external sources.

2.4.0.1 An Example: A Conducting Sphere

To illustrate the kind of problem that we have to deal with, it's probably best just to give an example. Consider a constant background electric field. (It could, for example, be generated by two charged plates of the kind we looked at in Subsection 2.1.4). Now place a neutral, spherical conductor inside this field. What happens?

We know that the conductor can't suffer an electric field inside it. Instead, the mobile charges in the conductor will move: the negative ones to one side; the positive ones to the other. The sphere now becomes *polarised*. These charges counteract the background electric field such that $\mathbf{E} = 0$ inside the conductor, while the electric field outside impinges on the sphere at right-angles. The end result must look qualitatively like Fig. 2.14.

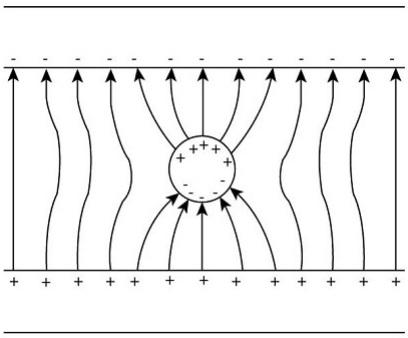


Fig. 2.14: A spherical conductor placed inside the field generated by two charged plates.

We'd like to understand how to compute the electric field in this, and related, situations. We'll give the answer in Subsection 2.4.4.

2.4.0.2 An Application: Faraday Cage

Consider some region of space that doesn't contain any charges, surrounded by a conductor. The conductor sits at constant $\phi = \phi_0$ while, since there are no charges inside, we must have $\nabla^2\phi = 0$. But this means that $\phi = \phi_0$ everywhere. This is because, if it didn't then there would be a maximum or minimum of ϕ somewhere inside. And we know from the discussion in Subsection 2.2.5 that this can't happen. Therefore, inside a region surrounded by a conductor, we must have $\mathbf{E} = 0$.

This is a very useful result if you want to shield a region from electric fields. In this context, the surrounding conductor is called a *Faraday cage*. As an application, if you're worried that they're trying to read your mind with electromagnetic waves, then you need only wrap your head in tin foil and all concerns should be alleviated.

2.4.1 Capacitors

Let's now solve for the electric field in some conductor problems. The simplest examples are *capacitors*. These are a pair of conductors, one carrying charge Q, the other charge $-Q$.

2.4.1.1 Parallel Plate Capacitor

To start, we'll take the conductors to have flat, parallel surfaces as shown in the figure. We usually assume that the distance d between the surfaces is much smaller than $\sqrt{A}$, where A is the area of the surface. This means that we can neglect the effects that arise around the edge of plates and we're justified in assuming that the electric field between the two plates is the same as it would be if the plates were infinite in extent. The problem reduces to the same one that we considered in Subsection 2.1.4. The electric field necessarily vanishes inside the conductor while, between the plates we have the result (2.26),

$$\mathbf{E} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{z}}, \tag{2.82}$$

where $\sigma = Q/A$ and we have assumed the plates are separated in the z-direction. We define the *capacitance* C to be

$$C = \frac{Q}{V}, \tag{2.83}$$

where V is the *voltage* or *potential difference* which is, as the name suggests, the difference in the potential ϕ on the two conductors. Since $E = -\,\mathrm{d}\phi/\mathrm{d}z$ is constant, we must have

$$\phi = -Ez + c \quad \Longrightarrow \quad V = \phi(0) - \phi(d) = Ed = \frac{Qd}{A\epsilon_0}, \tag{2.84}$$

and the capacitance for parallel plates of area A, separated by distance d, is

$$C = \frac{A\epsilon_0}{d}. \tag{2.85}$$

Because V was proportional to Q, the charge has dropped out of our expression for the capacitance. Instead, C depends only on the geometry of the set-up. This is a general property; we will soon see another example.

Capacitors are usually employed as a method to store electrical energy. We can see how much. Using our result (2.69), we have

$$U = \frac{\epsilon_0}{2}\int \mathrm{d}^3x\, \mathbf{E}\cdot\mathbf{E} = \frac{A\epsilon_0}{2}\int_0^d \mathrm{d}z \left(\frac{\sigma}{\epsilon_0}\right)^2 = \frac{Q^2}{2C}. \tag{2.86}$$

This is the energy stored in a parallel plate capacitor.

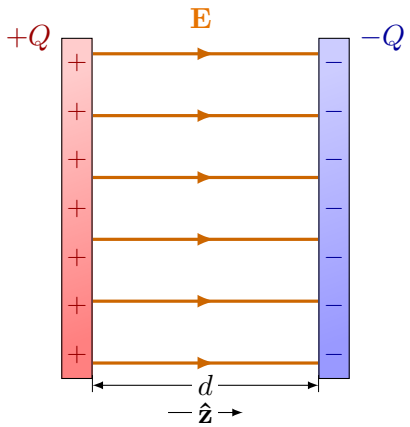


Fig. 2.15: Parallel plate capacitor.

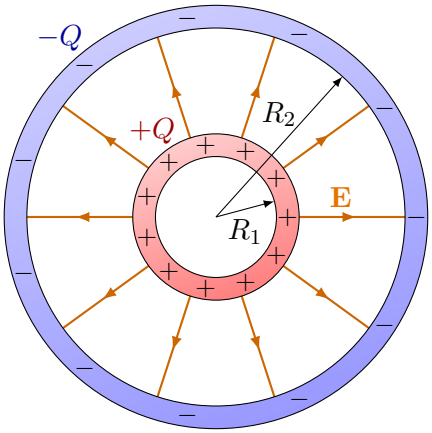


Fig. 2.16: Concentric sphere capacitor.

2.4.1.2 Concentric Sphere Capacitor

Consider a spherical conductor of radius R_1. Around this we place another conductor in the shape of a spherical shell with inner surface lying at radius R_2. We add charge $+Q$ to the sphere and $-Q$ to the shell. From our earlier discussion of charged spheres and shells, we know that the electric field between the two conductors must be

$$\mathbf{E} = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}, \quad R_1 < r < R_2. \tag{2.87}$$

Correspondingly, the potential is

$$\phi = \frac{Q}{4\pi\epsilon_0 r}, \quad R_1 < r < R_2, \tag{2.88}$$

and the capacitance is given by $C = 4\pi\epsilon_0 R_1 R_2/(R_2 - R_1)$.

2.4.2 Boundary Value Problems

Until now, we've thought of conductors as carrying some fixed charge Q. These conductors then sit at some constant potential ϕ. If there are other conductors in the vicinity that

carry a different charge then, as we've seen above, there will be some fixed potential difference, $V = \Delta\phi$ between them.

However, we can also think of a subtly different scenario. Suppose that we instead fix the potential ϕ in a conductor. This means that, whatever else happens, whatever other charges are doing all around, the conductor remains at a fixed ϕ. It never deviates from this value.

Now, this sounds a bit strange. We've seen above that the electric potential of a conductor depends on the distance to other conductors and also on the charge it carries. If ϕ remains constant, regardless of what objects are around it, then it must mean that the charge on the conductor is not fixed. And that's indeed what happens. Having conductors at fixed ϕ means that charge can flow in and out of the conductor. We implicitly assume that there is some background reservoir of charge which the conductor can dip into, taking and giving charge so that ϕ remains constant.

We can think of this reservoir of charge as follows: suppose that, somewhere in the background, there is a huge conductor with some charge Q which sits at some potential ϕ. To fix the potential of any other conductor, we simply attach it to one of this big reservoir-conductor. In general, some amount of charge will flow between them. The big conductor doesn't miss it, while the small conductor makes use of it to keep itself at constant ϕ.

The simplest example of the situation above arises if you connect your conductor to the planet Earth. By convention, this is taken to have $\phi = 0$ and it ensures that your conductor also sits at $\phi = 0$. Such conductors are said to be *grounded*. In practice, one may ground a conductor inside a chip in your cell phone by attaching it the metal casing.

Mathematically, we can consider the following problem. Take some number of objects, S_i. Some of the objects will be conductors at a fixed value of ϕ_i. Others will carry some fixed charge Q_i. This will rearrange itself into a surface charge σ_i such that $\mathbf{E} = 0$ inside while, outside the conductor, $\mathbf{E} = 4\pi\sigma\hat{\mathbf{n}}$. Our goal is to understand the electric field that threads the space between all of these objects. Since there is no charge sitting in this space, we need to solve the Laplace equation

$$\nabla^2\phi = 0, \tag{2.89}$$

subject to one of two boundary conditions

- Dirichlet Boundary Conditions: The value of ϕ is fixed on a given surface S_i.
- Neumann Boundary Conditions: The value of $\boldsymbol{\nabla}\phi \cdot \hat{\mathbf{n}}$ is fixed perpendicular to a given surface $\mathcal{S}_i$.

Notice that, for each $\mathcal{S}_i$, we need to decide which of the two boundary conditions we want. We don't get to chose both of them. We then have that with either Dirichlet or Neumann boundary conditions chosen on each surface $\mathcal{S}_i$, the Laplace equation has a unique solution. A proof is given in Appendix A.3.

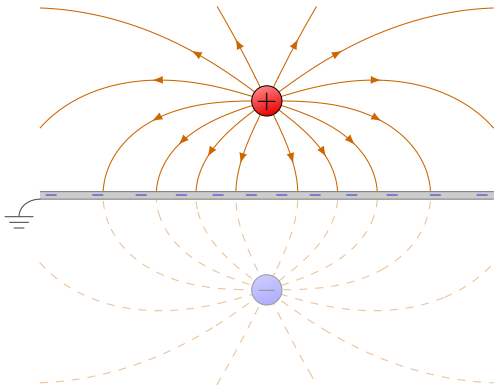


Fig. 2.17: A particle near a conducting plane looks like a dipole.

2.4.3 Method of Images

For particularly simple situations, there is a rather cute method that we can use to solve problems involving conductors. Although this technique is somewhat limited, it does give us some good intuition for what's going on. It's called the *method of images*.

2.4.3.1 A Charged Particle near a Conducting Plane

Consider a conductor which fills all of space $x < 0$. We'll ground this conductor so that $\phi = 0$ for $x < 0$. Then, at some point $x = d > 0$, we place a charge q. What happens?

We're looking for a solution to the Poisson equation with a delta-function source at $\mathbf{x} = \mathbf{D} = (d, 0, 0)$, together with the requirement that $\phi = 0$ on the plane $x = 0$. From our discussion in the previous section, there's a unique solution to this kind of problem. We just have to find it.

Here's the clever trick. Forget that there's a conductor at $x < 0$. Instead, suppose that there's a charge $-q$ placed opposite the real charge at $x = -d$. This is called the *image charge*. The potential for this pair of charges is just the potential

$$\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{\sqrt{(x-d)^2+y^2+z^2}} - \frac{q}{\sqrt{(x+d)^2+y^2+z^2}}\right). \tag{2.90}$$

By construction, this has the property that $\phi = 0$ for $x = 0$ and it has the correct source at $\mathbf{x} = (d, 0, 0)$. Therefore, this must be the right solution when $x \geq 0$. A cartoon of this is shown in Fig. 2.17. Of course, it's the wrong solution inside the conductor where the electric field vanishes. But that's trivial to fix: we just replace it with $\phi = 0$ ofr $x < 0$.

With the solution (2.90) in hand, we can now dispense with the image charge and explore what's really going on. We can easily compute the electric field from (2.90). If we focus on the electric field in the x-direction, it is

$$E_x = -\frac{\partial\phi}{\partial x} = \frac{q}{4\pi\epsilon_0}\left(\frac{x-d}{|\mathbf{r}-\mathbf{D}|^3} - \frac{x+d}{|\mathbf{r}+\mathbf{D}|^3}\right), \quad x \geq 0. \tag{2.91}$$

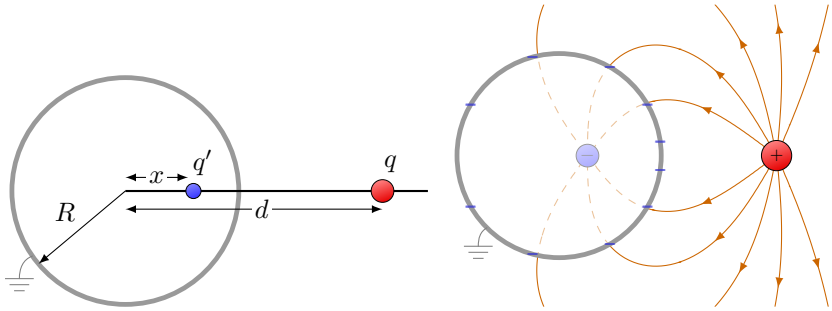


Fig. 2.18: A charged particle near a conducting sphere looks like an unbalanced dipole.

Meanwhile, $E_x = 0$ for $x < 0$. The discontinuity of E_x at the surface of the conductor determines the induced surface charge (2.81). It is

$$\sigma = E_x \epsilon_0|_{x=0} = -\frac{q}{2\pi} \frac{d}{(d^2 + y^2 + z^2)^{3/2}}. \tag{2.92}$$

We see that the surface charge is mostly concentrated on the plane at the point closest to the real charge. As you move away, it falls off as $1/(y^2 + z^2)^{3/2}$. We can compute the total induced surface charge by doing a simple integral,

$$q_{\text{induced}} = \int \mathrm{d}y\, \mathrm{d}z\, \sigma = -q. \tag{2.93}$$

The charge induced on the conductor is actually equal to the image charge. This is always true when we use the image charge technique.

Finally, as far as the real charge $+q$ is concerned, as long as it sits at $x > 0$, it feels an electric field which is identical in all respects to the field due to an image charge $-q$ embedded in the conductor. This means, in particular, that it will experience a force

$$\mathbf{F} = -\frac{q^2}{16\pi\epsilon_0 d^2}\hat{\mathbf{x}}. \tag{2.94}$$

This force is attractive, pulling the charge towards the conductor.

2.4.3.2 A Charged Particle Near a Conducting Sphere

We can play a similar game for a particle near a grounded, conducting sphere. The details are only slightly more complicated. We'll take the sphere to sit at the origin and have radius R. The particle has charge q and sits at $\mathbf{x} = \mathbf{D} = (d, 0, 0)$, with $d > R$. Our goal is to place an image charge q' somewhere inside the sphere so that $\phi = 0$ on the surface.

There is a way to derive the answer using conformal transformations. However, here we'll just state it. You should choose a particle of charge $q' =$, placed at $x = R_2/d$ and, by symmetry, $y = z = 0$. A cartoon of this is shown in Fig. 2.18.

The resulting potential is

$$\phi = \frac{q}{4\pi\epsilon_0}\left(\frac{1}{\sqrt{(x-d)^2 + y^2 + z^2}} - \frac{R}{d}\frac{1}{\sqrt{(x - R^2/d)^2 + y^2 + z^2}}\right). \tag{2.95}$$

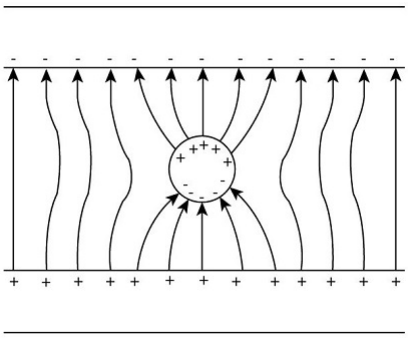


Fig. 2.19: A charged particle near a conducting sphere looks like an unbalanced dipole.

With a little algebra, you can check that $\phi = 0$ whenever $x^2 + y^2 + z^2 = R^2$. With a little more algebra, you can easily determine the induced surface charge and check that, when integrated over the sphere, we indeed have $q_{\text{induced}} = q'$. Once again, our charge experiences a force towards the conductor.

Above we've seen how to treat a grounded sphere. But what if we instead have an isolated conductor with some fixed charge, Q? It's easy to adapt the problem above. We simply add the necessary excess charge $Q - q'$ as an image that sits at the origin of the sphere. This will induce an electric field which emerges radially from the sphere. Because of the principle of superposition, we just add this to the previous electric field and see that it doesn't mess up the fact that the leectric field is perpendicular to the surface. This is now our solution.

2.4.4 Many Many More Problems

There are many mroe problems that you can cook up involving conductors, charges and electrostatics. Very few of them can be solved by the image charge method. Instead, you need to develop a number of basic tools of mathematical physics. A fairly comprehensive treatment of this can be found in the first 100 or so pages of Jackson.

For now, I would just like to leave you with the solution to the example that kicked off this section: what happens if you take a conducting sphere and place it in a constant electric field? This problem isn't quite solved by the image charge method. But it's solved by something similar: an image dipole.

We'll work in spherical polar coordinates and chose the original, constant electric field to point in the $\hat{\mathbf{z}}$ direction,

$$\mathbf{E}_0 = E_0\hat{\mathbf{z}} \quad \Longrightarrow \quad \phi_0 = -E_0 z = -E_0 r\cos\theta. \tag{2.96}$$

Take the conducting sphere to have radius R and be centered on the the origin. Let's add to this an image dipole with potential (2.45). We'll place the dipole at the origin, and orient it along the $\mathbf{z}$ axis like in Fig. 2.19.

The resulting potential is

$$\phi = -E_0\left(r - \frac{R^3}{r^2}\right)\cos\theta. \tag{2.97}$$

Since we've added a dipole term, we can be sure that this still solves the Laplace equation outside the conductor (See Appendix A.2 for further details on the form of this solution). Moreover, by construction, $\phi = 0$ when $r = R$. This is all we wanted from our solution. The induced surface charge can again be computed by evaluating the electric field just outside

$$\sigma = -\epsilon_0\frac{\partial\phi}{\partial r} = \epsilon_0\, E_0\left(1 + \frac{2R^3}{r^3}\right)\bigg|_{r=R}\cos\theta = 3\epsilon_0 E_0\cos\theta. \tag{2.98}$$

We see that the surface charge is positive in one hemisphere and negative in the other. The total induced charge averages to zero.

2.5 Dielectrics

Up to this point we have been concerned with the behaviour of static electric fields in free space, although we did allow perfectly conducting boundaries. We would now like to understand the behaviour of static electric fields in the presence of insulating materials, such as crystals and plastics. We will find that, with some small modification, the techniques that have been described previously can be applied directly even when dielectric bodies are present.

2.5.1 Isotropic Dielectrics

Faraday found that when an insulator is placed between the plates of a capacitor, held at a constant potential difference, the charge on the plates *increases* (alternatively, for constant charges, the potential *decreases*, as demonstrated in the lecture). In fact, if the insulator completely fills the space between the plates, the charge increases by a factor ϵ, which is called the *relative dielectric constant* or *relative permittivity* of the material. This is sometimes written ϵ_r or k instead of ϵ. The dielectric constant of most materials is of order 1 to 10, but it can be as high as 1000.

We are interested in insulating materials, and therefore, by definition, there is no free charge in the material itself. All of the charge is bound, and we shall assume that, overall, the material is neutral. Hence, it is crucially important to distinguish between *bound* charge and *free* charge.

Despite the absence of any free charges in an insulator, an applied electric field can cause positive and negative bound charge to separate, such that a dipole moment is induced. Microscopically, the electrons and the nucleus are displaced in different directions. For moderate fields in most materials, the induced dipole moment in a unit cell, or atom, is linearly proportional to the applied field.

An isotropic material is one for which the magnitude of the induced dipole moment does not depend on the orientation of the field with respect to the material. Most materials are anisotropic at some level; for example, sapphire.

In a region where the electric field is uniform, and the material homogeneous, charge *only appears on the external surfaces*. Since we are considering insulators, this effect is not due to free charge being displaced, but is caused by the cancellation of the separated charge internal to the material. A model to understand this effect conceptually is discussed in the following.

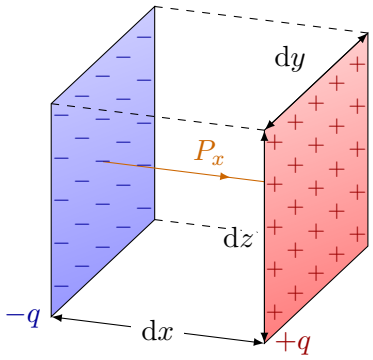


Fig. 2.20: The dipole moment of an infinitesimally small volume.

Consider some infinitesimally small volume $\mathrm{d}^3\mathbf{r}$, as shown in Fig. 2.20. The application of an external electric field causes positive charge $+q$ to be displaced towards one face, and negative charge $-q$ towards the opposite face. As a consequence there is an excess of charge at the two faces, even though, overall, the volume is neutral.

If a field is applied in the x-direction, the dipole moment is given by

$$p_x = q\,\mathrm{d}x = \sigma_x\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,, \tag{2.99}$$

where σ_x is the surface charge density on the $\hat{\mathbf{i}}$-directed surface. We could also have applied a field in the y and z-directions:

$$\begin{aligned} p_x &= \sigma_x\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,, \\ p_y &= \sigma_y\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,, \\ p_z &= \sigma_z\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,. \end{aligned} \tag{2.100}$$

It is more usual to express a dipole moment as the *dipole moment per unit volume*, P. The concept of P is meaningful because if we, say, place two volume elements side by side, then the charge on the combined surfaces is twice that of the single surface, but the displacement is the same; if we place two volume elements end to end, the charges on the internal surface cancel, the charges on the other surfaces remain the same, but now the separation has doubled. Therefore, it does not matter *how* we combine the small volumes - the dipole moment increases proportionally to the volume.

For each of x, y and z, the dipole moments per unit volume become

$$\begin{aligned} P_x &= \sigma_x, \\ P_y &= \sigma_y, \\ P_z &= \sigma_z. \end{aligned} \tag{2.101}$$

When a material is *anisotropic*, the polarisation density $\mathbf{P}$ depends on the direction of the inducing field, i.e., for a given $|\mathbf{E}|$, for each of x, y, and z, $\sigma_x = \sigma_y = \sigma_z$ is not fulfilled.

The expressions in Eq. (2.101) can now be combined into a single vector quantity, giving the vector dipole moment per unit volume $\mathbf{P}$:

$$\mathbf{P} = (P_x, P_y, P_z). \tag{2.102}$$

For example, say that the dipole moment associated with a single atom is $\mathbf{P}$, and that there are N atoms per unit volume, then in some small region the dipole moment per unit volume will be

$$\mathbf{P} = n\mathbf{P}, \tag{2.103}$$

where it is assumed that the medium is sufficiently diffuse that dipoles do not 'interact'.

If $\mathbf{P}$ is known, the polarisation (bound) charge density at the surface, with normal $\hat{\mathbf{n}}$, is given by

$$\sigma = |\mathbf{P}_\perp| = \mathbf{P} \cdot \hat{\mathbf{n}}. \tag{2.104}$$

Now consider a macroscopically large volume of material, with a uniform external field, as shown in Fig. 2.21. In this case, the charges on the internal surfaces cancel, and charge is only left on the external surfaces. We are thus left with the idea that when an electric field is applied to an insulating material, bound charge separates, such that equal and opposite charge appears on opposite faces.

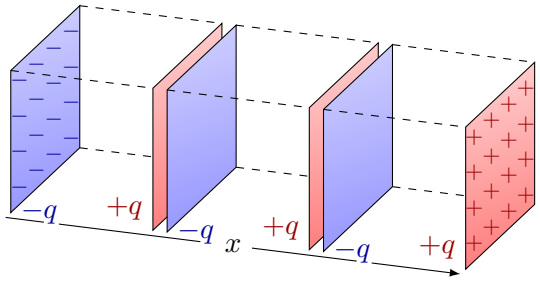


Fig. 2.21: The dipole moment of a macroscopically large volume.

When an insulating material is placed in a capacitor, a given internal electric field requires a given amount of charge on the plates, but we must compensate for the polarisation charge on the surfaces of the insulator, which tends to cancel the original charge. This must occur because the potential difference is fixed, and of course the separation between the plates is fixed, and therefore the electric field must be fixed ($E = V/d$). In other words, the free charge on the plates increases to offset the separated bound charge, which is consistent with Faraday's observation.

A self-consistent solution is reached when the total charge on the positive plate Q is equal to the free charge that is required to establish the field, $\mathbf{E}$, plus the additional free charge needed to offset the induced bound charge:

$$Q = \underbrace{\epsilon_0 |\mathbf{E}| A}_{\text{field without dielectric}} + \underbrace{\epsilon_0 \chi |\mathbf{E}| A}_{\text{offset induced bound charge}}, \tag{2.105}$$

where χ is the constant of proportionality that gives the surface polarisation charge for a given electric field.

Factorising,

$$Q = \epsilon_0(1+\chi)|\mathbf{E}|A. \tag{2.106}$$

By definition

$$C = \frac{Q}{V} = \frac{\epsilon_0(1+\chi)A}{d}. \tag{2.107}$$

Finally,

$$\epsilon = \frac{C_{\text{with dielectric}}}{C_{\text{without dielectric}}} = (1+\chi), \tag{2.108}$$

We thus arrive an important result: the relationship between the *relative permittivity* ϵ and the *susceptibility* χ:

$$\boxed{\epsilon = 1 + \chi.} \tag{2.109}$$

Of course we also have the capacitance in a capacitor filled with a dielectric:

$$C = \frac{Q}{V} = \frac{\epsilon_0\epsilon A}{d}. \tag{2.110}$$

The key point is that the charge Q always refers to the *free charge* on the plates - it does not include the polarisation (bound) charge on the surface of the dielectric. The presence of the bound charge is handled "automatically" through the introduction of the relative dielectric constant. Appreciating this point is central to understanding the classical description of the electrostatic behaviour of materials.

2.5.2 Polarisation Charge Density

It is now possible to use the basic concepts of polarisation, susceptibility, and relative dielectric constant in more sophisticated ways.

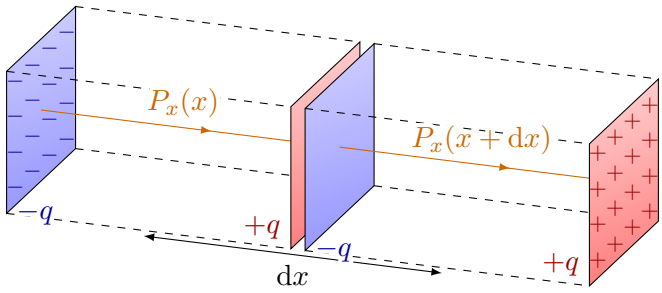


Fig. 2.22: Polarisation in a non-uniform field.

Consider what happens if the electric field is not uniform, which means that the polarisation is no longer uniform either, as shown in Fig. 2.22. In this case the separated, bound charge no longer cancels inside the material. The net internal, bound charge gives rise to a *polarisation* charge density ρ_p. The two internal surfaces (with $x = \text{const.}$), which in reality correspond to the same surface, therefore have surface charge density $\sigma = \mathbf{P}\cdot\hat{\mathbf{n}}$ and hence total charges

$$\begin{aligned} q_1 &= P_x(x)\,\mathrm{d}y\,\mathrm{d}z\,, \\ -q_2 &= -P_x(x+\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}z\,. \end{aligned} \tag{2.111}$$

Thus, the total charge on the internal surface due to the spatial dependence of P_x is

$$\begin{aligned} q &= q_1 - q_2 \\ &= [P_x(x) - P_x(x + \mathrm{d}x)]\,\mathrm{d}y\,\mathrm{d}z \\ &= -\frac{\partial P_x}{\partial x}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,, \end{aligned} \tag{2.112}$$

and therefore the charge per unit volume due to the spatially varying P_x is

$$\rho_{\mathrm{p},x} = -\frac{\partial P_x}{\partial x}. \tag{2.113}$$

Similar charges will appear if P_y or P_z changes with position, and therefore the total accumulated polarisation charge density is

$$\rho_{\mathrm{p}} = -\left[\frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z}\right], \tag{2.114}$$

and it follows for the *polarisation charge density*:

$$\boxed{\rho_{\mathrm{p}} = -\boldsymbol{\nabla}\cdot\mathbf{P}(\mathbf{r}),} \tag{2.115}$$

where the position dependence of the polarisation is shown explicitly.

The divergence of the polarisation per unit volume at a point gives the net polarisation charge density at that point. It is clear that bound charge is the source of $(-)\mathbf{P}$, in the same way that free charge is the source of $\mathbf{D}$.

2.5.3 Gauss' law for dielectric materials

Presented with (2.115) one is immediately inclined to apply the divergence theorem to some finite, closed volume. In other words,

$$\int_{\mathcal{V}} \mathrm{d}^3\mathbf{r}\,\rho_{\mathrm{p}} = -\int_{\mathcal{V}} \mathrm{d}^3\mathbf{r}\,\boldsymbol{\nabla}\cdot\mathbf{P} = -\oint_{\mathcal{S}} \mathrm{d}\mathbf{S}\cdot\mathbf{P} \tag{2.116}$$

The situation is shown in Fig. 2.23.

Previously, we saw that the dot product between the polarisation vector and the surface normal gives the surface charge density at the position where the dot product is evaluated ($\sigma = \mathbf{P}\cdot\hat{\mathbf{n}}$). The term on the RHS of (2.116) gives the total surface charge, but this must be equal and opposite to the internal bound charge from which it was separated. If there is no internal charge, because the electric field is everywhere uniform, then the surface integral of the normal component of the polarisation evaluates to zero.

In general, a volume of space may contain free charge ρ_{f} as well as bound charge ρ_{p} (see Fig. 2.24). For this volume, it is possible to apply Gauss' law, but now both of these charges must be taken into consideration.

According to Gauss' law we have

$$\int \mathrm{d}\mathbf{S}\cdot\mathbf{E} = \frac{1}{\epsilon_0}\int \mathrm{d}^3\mathbf{r}\,[\rho_{\mathrm{f}} + \rho_{\mathrm{p}}]. \tag{2.117}$$

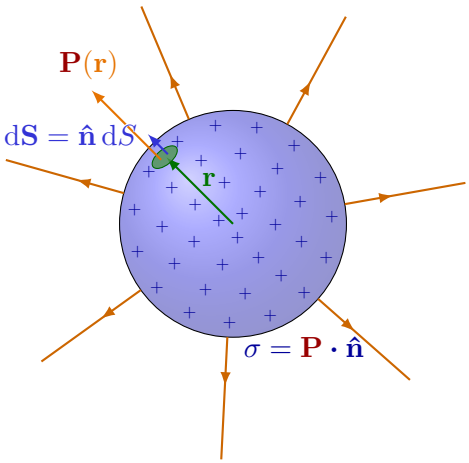


Fig. 2.23: Surface charge and polarisation.

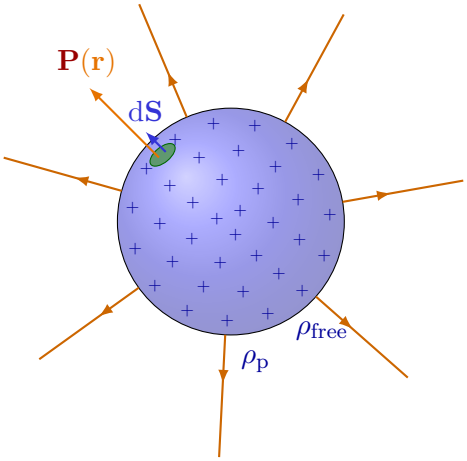


Fig. 2.24: Gauss' law with bound and free charge.

Applying the divergence theorem,

$$\begin{aligned}\int \mathrm{d}^3\mathbf{r}\,\boldsymbol{\nabla}\cdot\mathbf{E} &= \frac{1}{\epsilon_0}\int \mathrm{d}^3\mathbf{r}\,[\rho_{\mathrm{f}}+\rho_{\mathrm{p}}] \\ &= \frac{1}{\epsilon_0}\int \mathrm{d}^3\mathbf{r}\,[\rho_{\mathrm{f}}-\boldsymbol{\nabla}\cdot\mathbf{P}].\end{aligned} \tag{2.118}$$

Rearranging we get

$$\int \mathrm{d}^3\mathbf{r}\,\boldsymbol{\nabla}\cdot[\epsilon_0\mathbf{E}+\mathbf{P}] = \int \mathrm{d}^3\mathbf{r}\,\rho_{\mathrm{f}}. \tag{2.119}$$

Finally, because this expression must hold for all volumes,

$$\boldsymbol{\nabla}\cdot[\epsilon_0\mathbf{E}+\mathbf{P}] = \rho_{\mathrm{f}}. \tag{2.120}$$

We can now define the *electric displacement* $\mathbf{D}$ properly:

$$\boxed{\mathbf{D} = \epsilon_0\mathbf{E}+\mathbf{P}.} \tag{2.121}$$

We can thus conclude that the source of $\mathbf{D}$ are *free* charges and define *Gauss' law for dielectrics*:

$$\boxed{\boldsymbol{\nabla}\cdot\mathbf{D} = \rho_{\mathrm{f}}.} \tag{2.122}$$

Previously we assumed that the induced polarisation is proportional to the electric field,

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E}, \tag{2.123}$$

and therefore, using $\epsilon = (1 + \chi)$, we can define the *Electric displacement in terms of the electric field*:

$$\boxed{\mathbf{D}(\mathbf{r}) = \epsilon_0 \epsilon \mathbf{E}(\mathbf{r}),} \tag{2.124}$$

where position dependence has been referred to explicitly.

With the useful notation introduced above, we can now rewrite Eq. (2.120):

$$\boldsymbol{\nabla} \cdot [\epsilon_0 \epsilon \mathbf{E}] = \rho_{\mathrm{f}}. \tag{2.125}$$

This equation shows that (if ϵ is constant)

$$\boldsymbol{\nabla} \cdot [\epsilon_0 \mathbf{E}] = \frac{\rho_{\mathrm{f}}}{\epsilon}, \tag{2.126}$$

which indicates that the total charge enclosed is effectively reduced as a consequence of the induced bound surface charge.

Eq. (2.125) is a restatement of Gauss's law, but where the permittivity of free space ϵ_0 has been replaced by the permittivity of free space multiplied by the relative dielectric constant ϵ.

As described previously, the quantity $\mathbf{D}(\mathbf{r})$, which is a vector field, is called the *electric displacement* or the *electric flux density*. The notion of a flux density comes from our early lectures where we integrated the normal component of the electric field over a surface to create the concept of *electric flux*.

The key point is that the charge, $\rho_{\mathrm{f}}(\mathbf{r})$, in (2.122) relates to free charge, as it did before, but now the role of bound charge is included, indirectly, through the use of the *relative permittivity*. Whenever the relative permittivity is used in electrostatic field calculations, any reference to charge corresponds to free charge only: it does not include bound charge. It is remarkable that the effects of polarisation-charge separation can be taken care of solely by multiplying ϵ_0 by a simple multiplicative factor (for linear dielectrics).

2.5.3.1 A Word of Warning

The above formalism is used extensively in the study of electrostatic systems, and you are likely to go through life without ever worrying about anything different. A word of warning is, however, appropriate.

We have assumed that at each point in the material, the local electric field only induces a dipole. It does not induce a current, because no free charge is present, and neither does it induce higher-order poles, such as a quadrupole term. Remember that any charge distributions can be described by a hierarchy of terms, and so in principle any of these could be present. Thus, we have, implicitly, extracted the lowest-order term from a series of possibilities. Usually, this series converges so rapidly that it is sufficient to consider only the dipole term, but in certain materials this assumption can be violated.

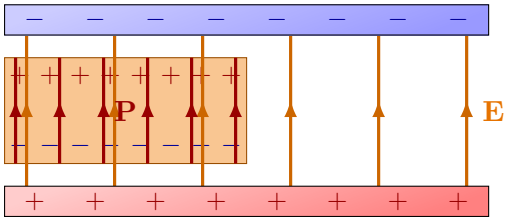


Fig. 2.25: The effect of a dielectric on the fields in a capacitor. $|\mathbf{E}| = V/d$, regardless of the presence of any dielectric. We then have $\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P}$, and so since the potential is held constant, when the dielectric material is introduced the free charge on the conducting surfaces increases to offset the induced bound charge, and to hold the potential constant.

In fact, the electric field induces a dipole, the gradient of the field induces a quadrupole, and the second derivative of the field induces an octupole. In almost all circumstances, the first term is dominant, and so this complication can be ignored. Indeed, this assumption is hidden in the majority of electromagnetic calculations.

2.5.4 Use of *D* and *E* in Electrostatic Problems

Usually, electrostatic problems come in two flavours, which are best approached in slightly different ways:

1. A collection of conducting surfaces is established and the potential differences between the surfaces are known, and remain fixed;
2. A collection of conducting surfaces is established, and the free charge on the surfaces is known, and remains fixed.

In the case of 1, and a homogeneous dielectric, the constant-potential surfaces determine the electric field $\mathbf{E}$ throughout the region, through Poisson's equation. For example, in the case of a parallel-plate capacitor, Fig. 2.25, $|\mathbf{E}| = V/d$, regardless of the presence of any dielectric. We then have $\mathbf{D} = \epsilon_0\epsilon\mathbf{E}$, and then Gauss' law gives the surface charge, $\sigma = |\mathbf{D}| = \epsilon_0\epsilon V/d$. As the potential is held constant, and dielectric material is introduced, the free charge on the conducting surfaces increases to offset the induced bound charge, and to hold the potential constant.

In the case of 2, a fixed distribution of free charge on the conductors determines $\mathbf{D}$ through Gauss's law $\boldsymbol{\nabla}\cdot\mathbf{D}(\mathbf{r}) = \rho_\mathrm{f}(\mathbf{r})$. For example, in the case of a parallel-plate capacitor $|\mathbf{D}| = \sigma$. The electric field is then given by $\mathbf{E} = \mathbf{D}/\epsilon_0\epsilon$, which in turn determines the potential difference, $V = |\mathbf{E}|d$. Therefore, $|\mathbf{E}| = \sigma/\epsilon_0\epsilon$, and $V = \sigma d/\epsilon_0\epsilon$. As the surface charge is held constant and dielectric material introduced, the voltage falls because the total electric field between the plates is smaller than would be expected on the basis of free charge alone.

Another way of thinking about the relationship between $\mathbf{D}$ and $\mathbf{E}$ is that free charge is the source of $\mathbf{D}$, such that the field lines associated with $\mathbf{D}$ can only start and end on free charge, whereas bound and free charges are the source of $\mathbf{E}$, and the field lines associated with $\mathbf{E}$ can begin and end on polarisation charge and free charge. This model

explains why, if the free charge is known, it is necessary to calculate $\mathbf{D}$ first, whereas if the potential difference is known, it is necessary to calculate $\mathbf{E}$ first. In this sense $\mathbf{E}$ is more fundamental than $\mathbf{D}$. Certainly, any test charge within a material will experience a force due to the total field $\mathbf{E}$, not merely due to the field that is associated with the free charge on the conductors.

2.5.5 Inhomogeneous Dielectrics and Boundary Conditions

We must now consider the case when ϵ is not homogeneous throughout the system of interest. First of all consider the relationships between the various field quantities on either side of a discontinuous change in ϵ. These relationships are called the boundary conditions.

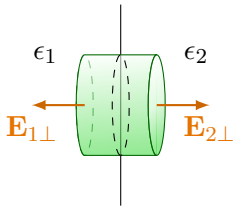


Fig. 2.26: The boundary between dissimilar dielectric materials.

To approach the problem of a boundary between two dissimilar materials, set up a "pillbox" that cuts across the surface , as shown in Fig. 2.26, and apply Gauss' law:

$$\int \mathbf{D}(\mathbf{r}) \cdot \mathrm{d}\mathbf{S} = \int \rho_{\mathrm{f}}(\mathbf{r})\, \mathrm{d}^3\mathbf{r} = 0, \tag{2.127}$$

where the second equality follows because there is no free surface charge at the boundary - unless there are surface imperfections.

Now shrink the parallel faces of the pillbox down, such that they are separated by an infinitesimally small distance. Also, the parallel faces of the pillbox are small enough that $\mathbf{D}$ is essentially constant over the area, and $\mathrm{d}\mathbf{S}$ points in opposite directions on the two surfaces. If the normal component of $\mathbf{D}$ in region 1 is called $\mathbf{D}_{1\perp}$ and the normal component of $\mathbf{D}$ in region 2 is $\mathbf{D}_{2\perp}$, then

$$A\mathbf{D}_{2\perp} - A\mathbf{D}_{1\perp} = \mathbf{0}, \tag{2.128}$$

where A is the area, from which it follows that

$$\mathbf{D}_{1\perp} = \mathbf{D}_{1\perp}. \tag{2.129}$$

We conclude that the normal component of $\mathbf{D}$ must be continuous across the boundary. This is independent of whether the relative permittivity changes or not. Only free charge is the source of $\mathbf{D}$. In turn, this shows that the normal component of the electric field must be discontinuous across the boundary, because $\mathbf{D} = \epsilon_0\epsilon\mathbf{E}$, which is self-consistent, because electric field ends on bound charge as well as free charge.

Now consider the boundary condition for the electric field. Place a closed loop around the boundary, as shown in Fig. 2.27, and apply Stokes's theorem,

$$\int \mathbf{E}(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 0. \tag{2.130}$$

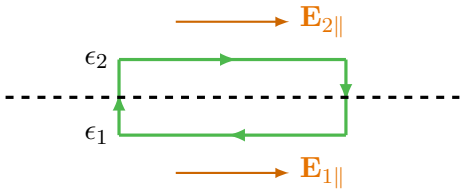


Fig. 2.27: The boundary between dissimilar dielectric materials.

If the sides of the loop are infinitesimally small, and the $\mathbf{E}$ field is constant along the sides, then we have

$$\mathbf{E}_{2\parallel}L - \mathbf{E}_{1\parallel}L = \mathbf{0}, \tag{2.131}$$

from which it follows

$$\mathbf{E}_{1\parallel} = \mathbf{E}_{12\parallel}. \tag{2.132}$$

We conclude that the parallel component of $\mathbf{E}$ must be continuous across the boundary, which is independent of whether the relative permittivity changes or not. The parallel component of $\mathbf{D}$ must be discontinuous across the boundary because $\mathbf{E} = \mathbf{D}/\epsilon_0\epsilon$.

In conclusion, across a dielectric boundary,

- The normal component of $\mathbf{D}$ is continuous ($\mathbf{D}_\perp$ continuous).
- The normal component of $\mathbf{E}$ is discontinuous.
- The parallel component of $\mathbf{E}$ is continuous ($\mathbf{E}_\parallel$ continuous).
- The parallel component of $\mathbf{D}$ is discontinuous.

2.5.6 The Behaviour of Field Lines at Dielectric Boundaries

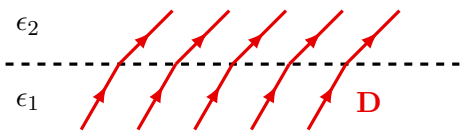


Fig. 2.28: The $\mathbf{D}$ fields at a dielectric boundary.

Consider how the boundary conditions on $\mathbf{D}$ and $\mathbf{E}$ affect the form of the field lines. For $\epsilon_2 > \epsilon_1$, it is obvious that the $\mathbf{D}$ field lines change direction as shown in Fig. 2.28, because $\mathbf{D}_{2\perp} = \mathbf{D}_{1\perp}$, and $\mathbf{D}_{2\parallel} > \mathbf{D}_{1\parallel}$.

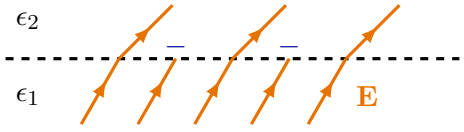


Fig. 2.29: The $\mathbf{E}$ fields at a dielectric boundary.

The E field has the same form, as shown in Fig. 2.29, because $\mathbf{E}_{2\parallel} = \mathbf{E}_{1\parallel}$, and $\mathbf{E}_{2\perp} < \mathbf{E}_{1\perp}$. In fact $\mathbf{D}$ and $\mathbf{E}$ point in the same direction for every point in space.

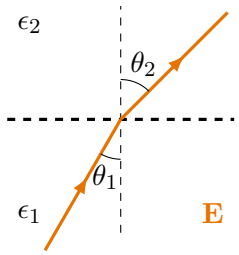


Fig. 2.30: Field lines at a dielectric boundary

The change in the direction of the field lines can be quantified as follows. For the **D** field, from $\mathbf{D}_\perp$ being constant across a boundary, using Fig. 2.30, we have

$$D_1 \cos\theta_1 = D_2 \cos\theta_2, \tag{2.133}$$

whereas for the **E** field, from $\mathbf{E}_\parallel$ being constant across a boundary:

$$E_1 \sin\theta_1 = E_2 \sin\theta_2. \tag{2.134}$$

Also $\mathbf{D} = \epsilon_0\epsilon\mathbf{E}$, so (2.133) becomes

$$\epsilon_0\epsilon_1 E_1 \cos\theta_1 = \epsilon_0\epsilon_2 E_2 \cos\theta_2. \tag{2.135}$$

Dividing by Eq. (2.134), we get for the change in the direction of the field lines at dielectric interfaces:

$$\epsilon_1 \cot\theta_1 = \epsilon_2 \cot\theta_2, \tag{2.136}$$

or

$$\frac{\cot\theta_2}{\cot\theta_1} = \frac{\epsilon_1}{\epsilon_2}. \tag{2.137}$$

We now have quantitatively derived the change in the direction of the field lines as a function of the relative permittivities of two materials across an interface. It is this change in direction that makes the analysis of general systems of dielectric bodies quite complicated. Of course, if the field lines are perpendicular to the surface, such as a parallel-plate capacitor with two dielectrics, or coaxial conductors with a coaxial dielectric filler, then the situation is straightforward to analyse. In fact, this is a key feature of all systems that can be analysed easily.

2.5.7 Boundary-Value Problems with Dielectrics

Once dielectric bodies having arbitrary shapes are combined with conducting surfaces having arbitrary shapes, electrostatic problems become complicated to solve, and the modern approach would be to set up a numerical model based on, say, Poisson's equation. Historically, other methods of solution had to be found, and this resulted in a rich structure of mathematical techniques. Indeed significant advances were made in mathematics itself because of the need to understand how to solve the integral and differential equations associated with electrostatic problems.

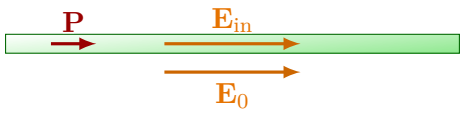


Fig. 2.31: Long thin rod parallel to a uniform field.

2.5.7.1 Long Thin Rod Parallel to a Uniform Field

Consider placing a long thin rod parallel to a uniform electric field of value $\mathbf{E}_0$: Fig. 2.31. Because the tangential component of the electric field must be continuous across the dielectric boundary, the internal electric field must be the same as the external electric field:

$$\mathbf{E}_{\text{in}} = \mathbf{E}_0. \tag{2.138}$$

We can therefore find the polarisation

$$\mathbf{P} = \epsilon_0 \mathbf{E}_0 \chi. \tag{2.139}$$

2.5.7.2 Thin Slab Perpendicular to a Uniform Field

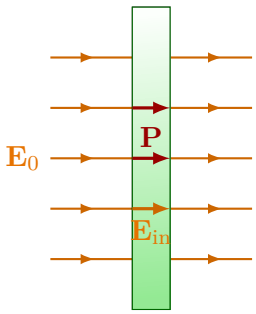


Fig. 2.32: Thin slab perpendicular to a uniform field.

Now consider a thin slab perpendicular to the field, as shown in Fig. 2.32. Because the slab is very large, we expect the internal electric field to be uniform and parallel to the external electric field. This is justified because, by symmetry, there can be no components in other directions. Since $\mathbf{D}$ and $\mathbf{E}$ are parallel, as shown in Subsection 2.5.6, also $\mathbf{D}$ is perpendicular to the boundary, i.e., $\mathbf{D} = \mathbf{D}_\perp$. Since $\mathbf{D}_\perp$ is continuous across the boundary, as we have seen in Subsection 2.5.5, it follows that

$$\epsilon_0 \epsilon \mathbf{E}_{\text{in}} = \epsilon \mathbf{E}_0, \tag{2.140}$$

giving

$$\mathbf{E}_{\text{in}} = \frac{\mathbf{E}_0}{\epsilon} = \frac{\mathbf{E}_0}{1+\chi}, \tag{2.141}$$

and finally

$$\mathbf{P} = \epsilon_0 \mathbf{E}_0 \frac{\chi}{1+\chi}. \tag{2.142}$$

These two examples serve to demonstrate that the polarisation and electric field inside a dielectric body are dependent on the shape of the body—as one would expect. What is

not so obvious is that for many simple shapes, such as a cylinder or sphere, the relationship always takes the form

$$\mathbf{P} = \epsilon_0 \mathbf{E}_0 \frac{\chi}{1 + n\chi}. \tag{2.143}$$

where $0 < n < 1$, and for a cylinder $n = 1/2$ and for a sphere $n = 1/3$, as we will see.

2.5.7.3 Dielectric Sphere in a Uniform Field

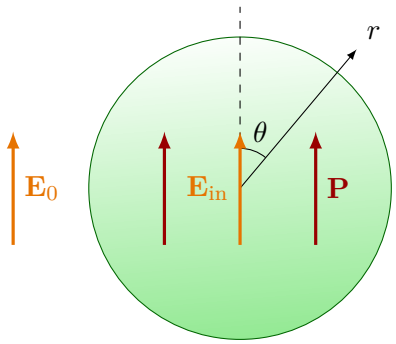


Fig. 2.33: Geometry of a dielectric *sphere* in a uniform field.

Consider a dielectric sphere with radius a in a uniform electric field, as shown in Fig. 2.33. Our Ansatz is to find the solution of Poisson's equation $\nabla^2 V = -(\rho_\mathrm{f} + \rho_\mathrm{p})/\epsilon_0$ that satisfies the boundary conditions (see Appendix A.2). Because there is no free charge in this problem – although formally free charge would be needed to create the parallel field – Poisson's equation reduces to $\nabla^2 V = -\rho_\mathrm{p}/\epsilon_0$.

Let us guess that the internal field is uniform, and that the external field is the original uniform field plus a dipole field generated by the surface polarisation charge on the sphere, which is similar to the case of the metallic sphere considered previously (cf. Subsection 2.4.4).

Using spherical polar coordinates, the potentials then become

$$V_\mathrm{in} = -E_\mathrm{in} r \cos\theta = -E_\mathrm{in} z \tag{2.144}$$

$$V_0 = -E_0 r \cos\theta + \frac{\kappa \cos\theta}{r^2}, \tag{2.145}$$

where κ is some constant of proportionality, whose value is to be found.

We require that the parallel component $E_\parallel$ be continuous across the boundary, where

$$E_\parallel = E_\theta = -\frac{1}{r}\frac{\partial V}{\partial \theta}\bigg|_{r=a}. \tag{2.146}$$

Therefore

$$-E_\mathrm{in} \sin\theta = -E_0 \sin\theta + \frac{\kappa \sin\theta}{r^3}\bigg|_{r=a}, \tag{2.147}$$

and

$$E_\mathrm{in} = E_0 - \frac{\kappa}{a^3}, \tag{2.148}$$

which can also be derived by requiring the potential to be continuous.

We also require that the normal component $D_\perp$ be continuous, where

$$D_\perp = -\epsilon_0 \epsilon \left.\frac{\partial V}{\partial r}\right|_{r=a}. \tag{2.149}$$

Now,

$$D_{\perp \text{in}} = \epsilon_0 \epsilon E_{\text{in}} \cos\theta, \tag{2.150}$$

$$D_{\perp 0} = \epsilon_0 E_0 \cos\theta + \epsilon_0 \frac{2\kappa \cos\theta}{a^3}, \tag{2.151}$$

where we have used the fact that $\epsilon = 1$ outside the dielectric. Therefore, using $D_{\perp \text{in}} = D_{\perp 0}$ and then Eq. (2.148)

$$\epsilon E_{\text{in}} = E_0 + \frac{2\kappa}{a^3} = \epsilon\left(E_0 + \frac{\kappa}{a^3}\right). \tag{2.152}$$

Hence we find

$$\kappa = \left(\frac{\epsilon - 1}{\epsilon + 2}\right) a^3 E_0, \tag{2.153}$$

and therefore can conclude considering Eq. (2.148) that the internal field is uniform:

$$\mathbf{E}_{\text{in}} = \frac{3}{\epsilon + 2}\mathbf{E}_0, \tag{2.154}$$

and finally that the polarisation is uniform, too:

$$\begin{aligned} \mathbf{P} &= \epsilon_0 E_{\text{in}} \chi \\ &= \frac{3\chi}{2 + \epsilon}\epsilon_0 \mathbf{E}_0 \\ &= \frac{\chi}{1 + \chi/3}\epsilon_0 \mathbf{E}_0. \end{aligned} \tag{2.155}$$

In summary, we find

$$\mathbf{P} = \frac{\chi}{1 + \chi/3}\epsilon_0 \mathbf{E}_0, \tag{2.156}$$

confirming our original guess about the relationship between polarisation and external field. Since it satisfies Poisson's equation and the boundary conditions, by the uniqueness theorem it is *the* solution, cf. Appendix A.2. The **D** field and the equipotentials are shown in Fig. 2.34.

2.5.7.4 Other Dielectric Bodies

Finding the electrostatic fields associated with other dielectric bodies is more complicated, and requires more sophisticated solutions of Poisson's equation. As representative examples, the solutions for oblate and prolate spheroids are shown in Fig. 2.35 and Fig. 2.36, respectively. Once the electric field is known, many other quantities follow, such as the couple exerted on a spheroid.

We shall not continue this topic further, but it is easy to gain an impression of the strength of the tools and the richness of the structures described.

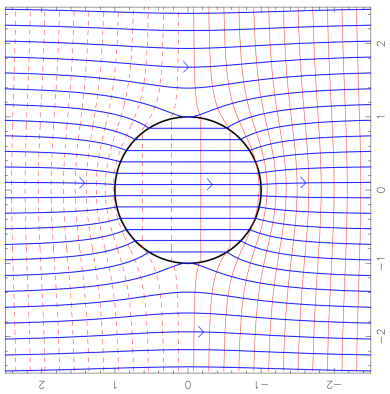


Fig. 2.34: Dielectric sphere in a uniform field. Lines of **D** are roughly horizontal (blue, dark), equipotentials are roughly vertical (red, light). In this example, $\epsilon = 4$.

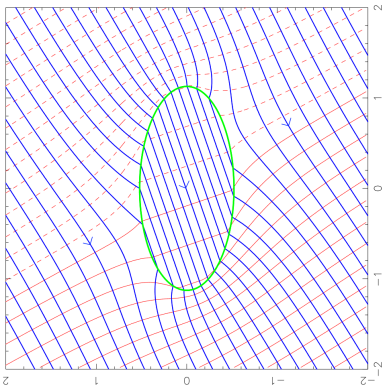


Fig. 2.35: Oblate dielectric spheroid in a uniform field. Lines of **D** slope down to the right (blue, dark), equipotentials slope up to the right (red, light).

2.5.8 Energy Density in Dielectrics

The only remaining issue is how to calculate the energy stored in an electric field when dielectric bodies are present. We have shown that the energy can be calculated from knowledge of the electric field alone when only free charge and conducting surfaces are present, but we have to be careful when dielectrics are introduced. The problem is that previously we calculated the energy required to bring up one charge at a time from infinity, and then summed all such contributions in order to find the energy needed to build a complete system. When dielectric bodies are present, however, not only is energy stored in the free charge that is moved, but energy is also stored in the bound charge that is separated. Rather than trying to reproduce the previous argument with induced charge, bringing up complete blocks of charge and placing them at the appropriate places one point at a time, which would be tedious, it is better to look at the other approach shown earlier, building up the complete charge distribution gradually, using a factor α going from 0 to 1.

We found in Eq. (2.67) that

$$U = \frac{1}{2}\int \mathrm{d}^3 r\, \rho(\mathbf{r}) V(\mathbf{r}). \tag{2.157}$$

In that derivation nothing was said that contradicts the argument when bound charge

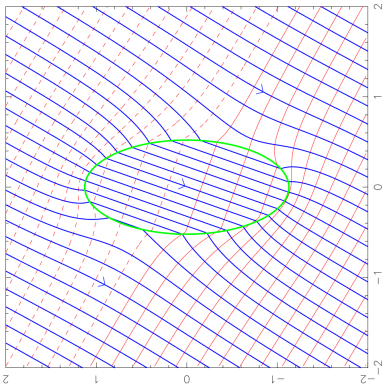


Fig. 2.36: Prolate dielectric spheroid in a uniform field. Lines of **D** slope down to the right (blue, dark), equipotentials slope up to the right (red, light).

is present, and hence we can apply a similar argumentation here. Bound (polarisation) charge is not explicitly included in $\rho(\mathbf{r})$, but rather gradually induced as the free charge is assembled.

We can now proceed efficiently, in almost exactly the same way as before in Subsection 2.3 for $\epsilon = 1$. We have

$$\begin{aligned} U &= \frac{1}{2}\int \mathrm{d}^3 r\, \rho(\mathbf{r}) V(\mathbf{r}) \\ &= \frac{1}{2}\int \mathrm{d}^3 r\, \boldsymbol{\nabla}\cdot\mathbf{D}(\mathbf{r}) V(\mathbf{r}) \end{aligned} \tag{2.158}$$

where Gauss' law has been used, but we also can integrate by parts using the vector identity

$$\boldsymbol{\nabla}\cdot(c\mathbf{A}) = c\boldsymbol{\nabla}\cdot\mathbf{A} + A\cdot\boldsymbol{\nabla}c, \tag{2.159}$$

where c is any scalar field, and $\mathbf{A}$ is any vector field, and therefore

$$U = \frac{1}{2}\int \mathrm{d}^3 r\, \{\boldsymbol{\nabla}\cdot[V(\mathbf{r})\mathbf{D}(\mathbf{r})] - \mathbf{D}(\mathbf{r})\cdot\boldsymbol{\nabla}V(\mathbf{r})\}. \tag{2.160}$$

Now look at each of the two RHS terms in turn.

The *first term* can be replaced, through the divergence theorem, by an integration over the surface containing the field,

$$\frac{1}{2}\int \mathrm{d}^3 r\, \boldsymbol{\nabla}\cdot[V(\mathbf{r})\mathbf{D}(\mathbf{r})] = \frac{1}{2}\int \mathrm{d}\mathbf{S}\cdot\mathbf{D}(\mathbf{r})V(\mathbf{r}). \tag{2.161}$$

The integration surface must, however, be taken at infinity, if it is to include all of the field. For a localised charge $V(\mathbf{r})$ falls as $1/r$ and $\mathbf{D}(\mathbf{r})$ falls as $1/r^2$, whereas the area increases as r^2, and therefore the integral on the RHS of Eq. (2.161) tends to zero for $r \to \infty$.

Only the *second term* remains, and hence the energy density in dielectrics is

$$\begin{aligned} U &= -\frac{1}{2}\int \mathrm{d}^3 r\, \mathbf{D}(\mathbf{r})\cdot\boldsymbol{\nabla}V(\mathbf{r}) \\ &= \frac{1}{2}\int \mathrm{d}^3 r\, \mathbf{D}(\mathbf{r})\cdot\mathbf{E}(\mathbf{r}). \end{aligned} \tag{2.162}$$

The argumentation above is similar to the dielectric-free version. In fact, for free space, where $\mathbf{D}(\mathbf{r}) = \epsilon_0 \mathbf{E}(\mathbf{r})$ it gives exactly the same result.

We conclude that the *energy stored in an electrostatic field* is given by the general equation

$$\boxed{U = \frac{1}{2} \int \mathrm{d}^3 r \, \mathbf{D}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r}).} \tag{2.163}$$

This relation is generally applicable, in free space or when insulating dielectric materials are present.

Chapter 3

Magnetostatics

Charges give rise to electric fields. Current give rise to magnetic fields. In this chapter, we will study the magnetic fields induced by steady currents. This means that we are again looking for time independent solutions to the Maxwell equations. We will also restrict to situations in which the charge density vanishes, so $\rho = 0$. We can then set $\mathbf{E} = 0$ and focus our attention only on the magnetic field. We're left with two Maxwell equations to solve:

$$\boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J}, \tag{3.1}$$

and

$$\boldsymbol{\nabla} \cdot \mathbf{B} = 0. \tag{3.2}$$

If you fix the current density $\mathbf{J}$, these equations have a unique solution. Our goal in this chapter is to find it.

3.0.0.1 Steady Currents

Before we solve (3.1) and (3.2), let's pause to think about the kind of currents that we're considering in this section. Because $\rho = 0$, there can't be any net charge. But, of course, we still want charge to be moving! This means that we necessarily have both positive and negative charges which balance out at all points in space. Nonetheless, these charges can move so there is a current even though there is no net charge transport.

This may sound artificial, but in fact it's exactly what happens in a typical wire. In that case, there is background of positive charge due to the lattice of ions in the metal. Meanwhile, the electrons are free to move. But they all move together so that at each point we still have $\rho = 0$. The continuity equation, which captures the conservation of electric charge, is

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{J} = 0. \tag{3.3}$$

Since the charge density is unchanging (and, indeed, vanishing), we have

$$\boldsymbol{\nabla} \cdot \mathbf{J} = 0. \tag{3.4}$$

Mathematically, this is just saying that if a current flows into some region of space, an equal current must flow out to avoid the build up of charge. Note that this is consistent with (3.1) since, for any vector field, $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \mathbf{B}) = 0$.

3.1 Ampére's Law

The first equation of magnetostatics,

$$\boldsymbol{\nabla} \times \mathbf{B} = \mu_0 \mathbf{J}, \tag{3.5}$$

is known as *Ampére's* law. As with many of these vector differential equations, there is an equivalent form in terms of integrals. In this case, we choose some open surface $\mathcal{S}$ with boundary $\mathcal{C} = \partial\mathcal{S}$. Integrating (3.5) over the surface, we can use Stokes' theorem to turn the integral of $\boldsymbol{\nabla} \times \mathbf{B}$ into a line integral over the boundary $\mathcal{C}$,

$$\int_{\mathcal{S}} \boldsymbol{\nabla} \times \mathbf{B} \cdot \mathrm{d}\mathbf{S} = \oint_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = \mu_0 \int_{\mathcal{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S}\,. \tag{3.6}$$

Recall that there's an implicit orientation in these equations. The surface $\mathcal{S}$ comes with a normal vector $\hat{\mathbf{n}}$ which points away from $\mathcal{S}$ in one direction. The line integral around the boundary is then done in the right-handed sense, meaning that if you stick the thumb of your right hand in the direction $\hat{\mathbf{n}}$ then your fingers curl in the direction of the line integral.

The integral of the current density over the surface $\mathcal{S}$ is the same thing as the total current I that passes through $\mathcal{S}$. Ampére's law in integral form then reads

$$\oint_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = \mu_0 I. \tag{3.7}$$

For most examples, this isn't sufficient to determine the form of the magnetic field; we'll usually need to invoke (3.2) as well. However, there is one simple example where symmetry considerations mean that (3.7) is all we need.

3.1.1 A Long Straight Wire

Consider an infinite, straight wire carrying current I. We'll take it to point in the $\hat{\mathbf{z}}$ direction. The symmetry of the problem is jumping up and down telling us that we need to use cylindrical polar coordinates, (r, ϕ, z), where $r = \sqrt{x^2 + y^2}$ is the radial distance away from the wire.

We take the open surface $\mathcal{S}$ to lie in the $x - y$ plane, centered on the wire. For the line integral in (3.7) to give something that doesn't vanish, it's clear that the magnetic field has to have some component that lies along the circumference of the disc.

But, by the symmetry of the problem, that's actually the only component that $\mathbf{B}$ can have: it must be of the form $\mathbf{B} = B(r)\hat{\boldsymbol{\varphi}}$. (If this was a bit too quick, we'll derive this more carefully soon). Any magnetic field of this form automatically satisfies the second Maxwell equation $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$. We need only worry about Ampére's law which tells us

$$\oint_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = B(r) \int_0^{2\pi} r\, \mathrm{d}\varphi = 2\pi r B(r) = \mu_0 I. \tag{3.8}$$

We see that the strength of the magnetic field is

$$\mathbf{B} = \frac{\mu_0 I}{2\pi r}\hat{\boldsymbol{\varphi}}. \tag{3.9}$$

The magnetic field circles the wire using the "right-hand rule": stick the thumb of your right hand in the direction of the current and your fingers curl in the direction of the magnetic field.

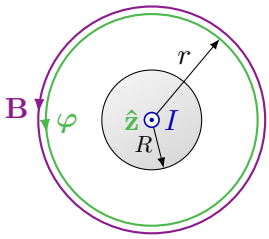


Fig. 3.1: An infinite, straight wire carrying current I and pointing in the $\hat{\mathbf{z}}$ direction.

Note that the simplest example of a magnetic field falls off as $1/r$. In contrast, the simplest example of an electric field – the point charge – falls of as $1/r^2$. You can trace this difference back to the geometry of the two situations. Because magnetic fields are sourced by currents, the simplest example is a straight line and the $1/r$ fall-off is because there are two transverse directions to the wire. Indeed, we saw in Subsection 2.1.3 that when we look at a line of charge, the electric field also drops off as $1/r$.

3.1.2 Surface Currents and Discontinuities

Consider the flat plane lying at $z = 0$ with a surface current density that we'll call $\mathbf{K}$. Note that $\mathbf{K}$ is the current per unit length, as opposed to $\mathbf{J}$ which is the current per unit area. You can think of the surface current as a bunch of wires, all lying parallel to each other.

We'll take the current to lie in the $\hat{\mathbf{x}}$-direction: $\mathbf{K} = K\hat{\mathbf{x}}$ as shown in Fig. 3.2.

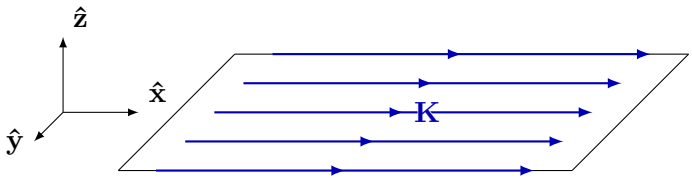


Fig. 3.2: A flat plane at $z = 0$ with a surface current density $\mathbf{K}$ in the $\hat{\mathbf{x}}$-direction.

From our previous result, we know that the $\mathbf{B}$ field should curl around the current in the right-handed sense. But, with an infinite number of wires, this can only mean that $\mathbf{B}$ is oriented along the $\hat{\mathbf{y}}$ direction. In fact, from the symmetry of the problem, it must look Fig. 3.3, with $\mathbf{B}$ pointing in the $-\hat{\mathbf{y}}$ direction when $z > 0$ and in the $+\hat{\mathbf{y}}$ direction when

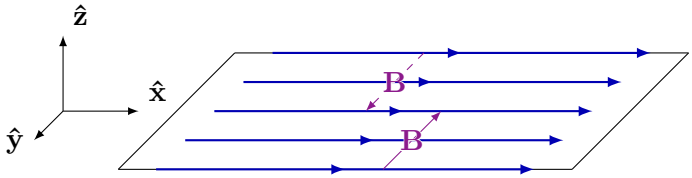


Fig. 3.3: Caption

$z < 0$. We write

$$\mathbf{B} = -B(z)\hat{\mathbf{y}}, \tag{3.10}$$

with $B(z) = -B(-z)$. We invoke Ampére's law using the open surface in Fig. 3.4 with

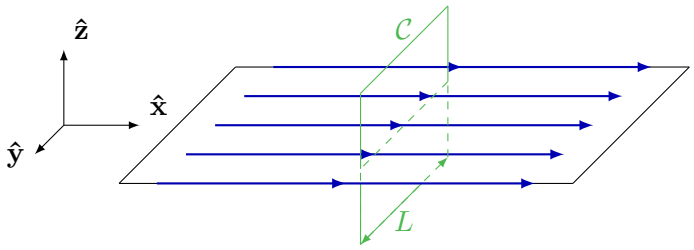


Fig. 3.4: A flat plane at $z = 0$ with a surface current density $\mathbf{K}$ in the $\hat{\mathbf{x}}$-direction.

length L in the y direction and extending to $\pm z$. We have

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = LB(z) - LB(-z) = 2LB(z) = \mu_0 KL, \tag{3.11}$$

so we find that the magnetic field is constant above an infinite plane of surface current

$$B(z) = \frac{mu_0 K}{2}, \quad z > 0. \tag{3.12}$$

This is rather similar to the case of the electric field in the presence of an infinite plane of surface charge.

The analogy with electrostatics continues. The magnetic field is not continuous across a plane of surface current. We have

$$B(z \to 0^+) - B(z \to 0^-) = \mu_0 K. \tag{3.13}$$

In fact, this is a general result that holds for any surface current $\mathbf{K}$. We can prove this statement by using the same curve that we used in Fig. 3.4 above and shrinking it until it barely touches the surface on both sides. If the normal to the surface is $\hat{\mathbf{n}}$ and $\mathbf{B}_\pm$ denotes the magnetic field on either side of the surface, then

$$\hat{\mathbf{n}} \times \mathbf{B}_+ - \hat{\mathbf{n}} \times \mathbf{B}_- = \mu_0 \mathbf{K}. \tag{3.14}$$

Meanwhile, the magnetic field normal to the surface is continuous. (To see this, you can use a Gaussian pillbox, together with the other Maxwell equation $\nabla \cdot \mathbf{B} = 0$).

When we looked at electric fields, we saw that the normal component was discontinuous in the presence of surface charge (2.23) while the tangential component is continuous. For magnetic fields, it's the other way around: the tangential component is discontinuous in the presence of surface currents.

3.1.2.1 A Solenoid

A *solenoid* consists of a surface current that travels around a cylinder. It's simplest to think of a single current-carrying wire winding many times around the outside of the cylinder. (Strictly speaking, the cross-sectional shape of the solenoid doesn't have to be a circle - it can be anything. But we'll stick with a circle here for simplicity). To make life easy, we'll assume that the cylinder is infinitely long. This just means that we can neglect effects due to the ends.

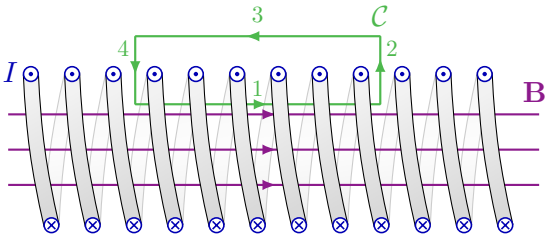


Fig. 3.5: Magnetic field **B** inside a solenoid.

We'll again use cylindrical polar coordinates, (r, φ, z), with the axis of the cylinder along $\hat{\mathbf{z}}$. By symmetry, we know that **B** will point along the z-axis. Its magnitude can depend only on the radial distance: $\mathbf{B} = B(r)\hat{\mathbf{z}}$. Once again, any magnetic field of this form immediately satisfies $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$.

We solve Ampére's law in differential form. Anywhere other than the surface of the solenoid, we have $\mathbf{J} = 0$ and

$$\boldsymbol{\nabla} \times \mathbf{B} = 0 \quad \Longrightarrow \quad \frac{\mathrm{d}B}{\mathrm{d}r} = 0 \quad \Longrightarrow \quad B(r) = \text{const.} \tag{3.15}$$

Outside the solenoid, we must have $B(r) = 0$ since $B(r)$ is constant and we know $B(r) \to 0$ as $r \to \infty$. To figure out the magnetic field inside the solenoid, we turn to the integral form of Ampére's law and consider the surface $\mathcal{S}$, bounded by the curve $\mathcal{C}$ shown in Fig. 3.5. Only the line 1 that runs inside the solenoid contributes to the line integral. We have

$$\oint_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = BL = \mu_0 INL, \tag{3.16}$$

where N is the number of windings of wire per unit length. We learn that inside the solenoid, the constant magnetic field is given by

$$\mathbf{B} = \mu_0 IN\hat{\mathbf{z}}. \tag{3.17}$$

Note that, since $K = IN$, this is consistent with our general formula for the discontinuity of the magnetic field in the presence of surface currents (3.14).

3.2 The Vector Potential

For the simple current distributions of the last section, symmetry considerations were enough to lead us to a magnetic field which automatically satisfied

$$\boldsymbol{\nabla} \cdot \mathbf{B} = 0. \tag{3.18}$$

But, for more general currents, this won't be the case. Instead we have to ensure that the second magnetostatic Maxwell equation is also satisfied.

In fact, this is simple to do. We are guaranteed a solution to $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ if we write the magnetic field as the curl of some vector field,

$$\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A} \tag{3.19}$$

Here $\mathbf{A}$ is called the *vector potential*. While magnetic fields that can be written in the form (3.19) certainly satisfy $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$, the converse is also true; any divergence-free magnetic field can be written as (3.19) for some $\mathbf{A}$.

(Actually, this previous sentence is only true if our space has a suitably simple topology. Since we nearly always think of space as $\mathbb{R}^3$ or some open ball on $\mathbb{R}^3$, we rarely run into subtleties. But if space becomes more interesting then the possible solutions to $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ also become more interesting. This is analogous to the story of the electrostatic potential that we mentioned briefly in Section 2.2).

Using the expression (3.19), Ampére's law becomes

$$\boldsymbol{\nabla} \times \mathbf{B} = -\nabla^2 \mathbf{A} + \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{B}) = \mu_0 \mathbf{J}, \tag{3.20}$$

where, in the first equality, we've used a standard identity from vector calculus. This is the equation that we have to solve to determine $\mathbf{A}$ and, through that, $\mathbf{B}$.

We can additionally substitute (3.19) into Maxwell's equations to work towards and expression for the electric field in terms of the vector potential,

$$\boldsymbol{\nabla} \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\boldsymbol{\nabla} \times \frac{\partial \mathbf{A}}{\partial t}, \tag{3.21}$$

and thus integrating this reveals,

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \boldsymbol{\nabla}\phi. \tag{3.22}$$

3.2.1 Magnetic Monopoles

Above, we dispatched with the Maxwell equation $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$ fairly quickly by writing $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$. But we never paused to think about what this equation is actually telling us. In fact, it has a very simple interpretation: it says that there are no magnetic charges. A point-like magnetic charge g would source the magnetic field, giving rise a $1/r^2$ fall-off

$$\mathbf{B} = \frac{g}{4\pi r^2}\hat{\mathbf{r}}. \tag{3.23}$$

An object with this behaviour is usually called a *magnetic monopole*. Maxwell's equations says that they don't exist. And we have never found one in Nature.

However, we could ask: how robust is this conclusion? Are we sure that magnetic monopoles don't exist? After all, it's easy to adapt Maxwell's equations to allow for presence of magnetic charges: we simply need to change (3.18) to read $\boldsymbol{\nabla} \cdot \mathbf{B} = \rho_m$ where ρ_m is the magnetic charge distribution. Of course, this means that we no longer get to use the vector potential $\mathbf{A}$. But is that such a big deal?

The twist comes when we turn to quantum mechanics. Because in quantum mechanics we're *obliged* to use the vector potential $\mathbf{A}$. Not only is the whole framework of electromagnetism in quantum mechanics based on writing things using $\mathbf{A}$, but it turns out that there are experiments that actually detect certain properties of $\mathbf{A}$ that are lost when we compute $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$. I won't explain the details here, but if you're interested then look up the "Aharonov-Bohm effect" in the lectures on Solid State Physics.

3.2.1.1 Monopoles After All?

To summarise, magnetic monopoles have never been observed. We have a law of physics (3.18) which says that they don't exist. And when we turn to quantum mechanics we need to use the vector potential $\mathbf{A}$ which automatically means that (3.18) is true. It sounds like we should pretty much forget about magnetic monopoles, right?

Well, no. There are actually very good reasons to suspect that magnetic monopoles do exist. The most important part of the story is due to Dirac. He gave a beautiful argument which showed that it is in fact possible to introduce a vector potential $\mathbf{A}$ which allows for the presence of magnetic charge, but only if the magnetic charge g is related to the charge of the electron e by

$$ge = 2\pi\hbar n, \quad \text{for} \quad n \in \mathbb{Z}. \tag{3.24}$$

This is known as the *Dirac quantisation condition*.

Moreover, following work in the 1970s by 't Hooft and Polyakov, we now realise that magnetic monopoles are ubiquitous in theories of particle physics. Our best current theory – the Standard Model – does not predict magnetic monopoles. But every theory that tries to go beyond the Standard Model, whether Grand Unified Theories, or String Theory or whatever, always ends up predicting that magnetic monopoles should exist. They're one of the few predictions for new physics that nearly all theories agree upon.

These days most theoretical physicists think that magnetic monopoles probably exist and there have been a number of experiments around the world designed to detect them. However, while theoretically monopoles seem like a good bet, their future observational status is far from certain. We don't know how heavy magnetic monopoles will be, but all evidence suggests that producing monopoles is beyond the capabilities of our current (or, indeed, future) particle accelerators. Our only hope is to discover some that Nature made for us, presumably when the Universe was much younger. Unfortunately, here too things seem against us. Our best theories of cosmology, in particular inflation, suggest that any monopoles that were created back in the Big Bang have long ago been diluted. At a guess, there are probably only a few floating around our entire observable Universe. The chances of one falling into our laps seem slim. But I hope I'm wrong.

3.2.2 Gauge Transformations

The choice of $\mathbf{A}$ in (3.19) is far from unique: there are lots of different vector potentials $\mathbf{A}$ that all give rise to the same magnetic field $\mathbf{B}$. This is because the curl of a gradient is automatically zero. This means that we can always add any vector potential of the form $\boldsymbol{\nabla}\chi$ for some function χ and the magnetic field remains the same,

$$\mathbf{A}' = \mathbf{A} + \boldsymbol{\nabla}\chi \quad \Longrightarrow \quad \boldsymbol{\nabla} \times \mathbf{A}' = \boldsymbol{\nabla} \times \mathbf{A}. \tag{3.25}$$

Such a change of $\mathbf{A}$ is called a *gauge transformation*. As we will see in Subsection 5.0.1, it is closely tied to the possible shifts of the electrostatic potential ϕ. Ultimately, such gauge transformations play a key role in theoretical physics. But, for now, we're simply going to use this to our advantage. Because, by picking a cunning choice of χ, it's possible to simplify our quest for the magnetic field.

We can always find a gauge transformation χ such that $\mathbf{A}'$ satisfies $\boldsymbol{\nabla}\cdot\mathbf{A}' = 0$. Making this choice is usually referred to as *Coulomb gauge*. Suppose that we've found some $\mathbf{A}$ which gives us the magnetic field that we want, so $\boldsymbol{\nabla}\times\mathbf{A} = \mathbf{B}$, but when we take the divergence we get some function $\boldsymbol{\nabla}\cdot\mathbf{B} = \psi(\mathbf{x})$. We instead choose $\mathbf{A}' = \mathbf{A} + \boldsymbol{\nabla}\chi$ which now has divergence

$$\boldsymbol{\nabla}\cdot\mathbf{A}' = \boldsymbol{\nabla}\cdot\mathbf{A} + \nabla^2\chi = \psi + \nabla^2\chi. \tag{3.26}$$

So if we want $\boldsymbol{\nabla}\cdot\mathbf{A}' = 0$, we just have to pick our gauge transformation χ to obey

$$\nabla^2\chi = -\psi. \tag{3.27}$$

But this is just the Poisson equation again. And we know from our discussion in Section 2.2 that there is always a solution. (For example, we can write it down in integral form using the Green's function).

3.2.2.1 Something a Little Misleading: The Magnetic Scalar Potential

There is another quantity that is sometimes used called the *magnetic scalar potential*, Ω. The idea behind this potential is that you might be interested in computing the magnetic field in a region where there are no currents and the electric field is not changing with time. In this case, you need to solve $\boldsymbol{\nabla}\times\mathbf{B} = 0$, which you can do by writing

$$\mathbf{B} = -\boldsymbol{\nabla}\Omega. \tag{3.28}$$

Now calculations involving the magnetic field really do look identical to those involving the electric field.

However, you should be wary of writing the magnetic field in this way. As we'll see in more detail in Subsection 5.0.1, we can *always* solve two of Maxwell's equations by writing $\mathbf{E}$ and $\mathbf{B}$ in terms of the electric potential ϕ and vector potential $\mathbf{A}$ and this formulation becomes important as we move onto more advanced areas of physics. In contrast, writing $\mathbf{B} = -\boldsymbol{\nabla}\Omega$ is only useful in a limited number of situations. The reason for this really gets to the heart of the difference between electric and magnetic fields: electric charges exist; magnetic charges don't!

3.2.3 Biot-Savart Law

We're now going to use the vector potential to solve for the magnetic field $\mathbf{B}$ in the presence of a general current distribution. From now, we'll always assume that we're working in Coulomb gauge and our vector potential obeys $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$. Then Ampére's law (3.20) becomes a whole lot easier: we just have to solve

$$\nabla^2\mathbf{A} = -\mu_0\mathbf{J}. \tag{3.29}$$

But this is just something that we've seen already. To see why, it's perhaps best to write it out in Cartesian coordinates. This then becomes three equations,

$$\nabla^2 A_i = -\mu_0 J_i, \quad (i = 1, 2, 3) \tag{3.30}$$

and each of these is the Poisson equation.

It's worth giving a word of warning at this point: the expression $\nabla^2\mathbf{A}$ is simple in Cartesian coordinates where, as we've seen above, it reduces to the Laplacian on each component. But, in other coordinate systems, this is no longer true. The Laplacian now also acts on the basis vectors such as $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\varphi}}$. So in these other coordinate systems, $\nabla^2\mathbf{A}$ is a little more of a mess. (You should probably use the identity $\nabla^2\mathbf{A} = -\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{A}) + \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{A})$) if you really want to compute in these other coordinate systems).

Anyway, if we stick to Cartesian coordinates then everything is simple. In fact, the resulting equations (3.30) are of exactly the same form that we had to solve in electrostatics. And, in analogy to (2.48), we know how to write down the most general solution using Green's functions. It is

$$A_i(x) = \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\, \frac{J_i(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}. \tag{3.31}$$

Or, if you're feeling bold, you can revert back to vector notation and write

$$\mathbf{A}(\mathbf{x}) = \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\, \frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}. \tag{3.32}$$

where you've just got to remember that the vector index on $\mathbf{A}$ links up with that on $\mathbf{J}$ (and not on $\mathbf{x}$ or $\mathbf{x}'$).

3.2.3.1 Checking Coulomb Gauge

We've derived a solution to (3.29), but this is only a solution to Ampére's equation (3.20) if the resulting $\mathbf{A}$ obeys the Coulomb gauge condition, $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$. Let's now check that it does. We have

$$\boldsymbol{\nabla}\cdot\mathbf{A}(\mathbf{x}) = \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\, \boldsymbol{\nabla}\cdot\left(\frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\right), \tag{3.33}$$

where you need to remember that the index of $\boldsymbol{\nabla}$ is dotted with the index of $\mathbf{J}$, but the derivative in $\boldsymbol{\nabla}$ is acting on $\mathbf{x}$, not on $\mathbf{x}'$ as that is the dummy variable of integration. We can write

$$\begin{aligned}\boldsymbol{\nabla}\cdot\mathbf{A}(\mathbf{x}) &= \frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\, \mathbf{J}(\mathbf{x}')\cdot\boldsymbol{\nabla}\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right)\\ &= -\frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\, \mathbf{J}(\mathbf{x}')\cdot\boldsymbol{\nabla}'\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right).\end{aligned} \tag{3.34}$$

Here we've done something clever. Now our $\boldsymbol{\nabla}'$ is differentiating with respect to $\mathbf{x}'$. To get this, we've used the fact that if you differentiate $1/|\mathbf{x}-\mathbf{x}'|$ with respect to $\mathbf{x}$ then you get the negative of the result from differentiating with respect to $\mathbf{x}'$. But since $\boldsymbol{\nabla}'$ sits inside an $\int \mathrm{d}^3x'$ integral, it's ripe for integrating by parts. This gives

$$\boldsymbol{\nabla}\cdot\mathbf{A}(\mathbf{x}) = -\frac{\mu_0}{4\pi}\int_{\mathcal{V}} \mathrm{d}^3x'\left[\boldsymbol{\nabla}'\cdot\left(\frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\right) - \boldsymbol{\nabla}'\cdot\mathbf{J}(\mathbf{x}')\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right)\right]. \tag{3.35}$$

The second term vanishes because we're dealing with steady currents obeying $\boldsymbol{\nabla}\cdot\mathbf{J} = 0$. The first term also vanishes if we take the current to be localised in some region of space,

$\hat{\mathcal{V}} \subset \mathcal{V}$ so that $\mathbf{J}(\mathbf{x}) = 0$ on the boundary $\partial\mathcal{V}$. We'll assume that this is the case. We conclude that

$$\boldsymbol{\nabla} \cdot \mathbf{A} = 0, \tag{3.36}$$

and (3.32) is indeed the general solution to the Maxwell equations (3.1) and (3.2) as we'd hoped.

3.2.3.2 The Magnetic Field

From the solution (3.32), it is simple to compute the magnetic field $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$. Again, we need to remember that the $\boldsymbol{\nabla}$ acts on the $\mathbf{x}$ in (3.32) rather than the $\mathbf{x}'$. We find

$$\mathbf{B}(\mathbf{x}) = \frac{\mu_0}{4\pi} \int_{\mathcal{V}} \mathrm{d}^3x' \, \frac{\mathbf{J}(\mathbf{x}') \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3}. \tag{3.37}$$

This is known as the *Biot-Savart law*. It describes the magnetic field due to a general current density.

There is a slight variation on (3.37) which more often goes by the name of the Biot-Savart law. This arises if the current is restricted to a thin wire which traces out a curve $\mathcal{C}$. Then, for a current density $\mathbf{J}$ passing through a small volume $\delta\mathcal{V}$, we write $\mathbf{J}\delta\mathcal{V} = (JA)\delta\mathbf{x}$ where A is the cross-sectional area of the wire and $\delta\mathbf{x}$ lies tangent to $\mathcal{C}$. Assuming that the cross-sectional area is constant throughout the wire, the current $I = JA$ is also constant. The Biot-Savart law becomes

$$\mathbf{B}(\mathbf{x}) = \frac{\mu_0}{4\pi} \int_{\mathcal{C}} \frac{\mathrm{d}\mathbf{x}' \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3}. \tag{3.38}$$

This describes the magnetic field due to the current I in the wire.

3.2.3.3 An Example: The Straight Wire Revisited

Of course, we already derived the answer for a straight wire in (3.9) without using this fancy vector potential technology. Before proceeding, we should quickly check that the Biot-Savart law reproduces our earlier result. As before, we'll work in cylindrical polar coordinates. We take the wire to point along the $\hat{\mathbf{z}}$ axis and use $r^2 = x^2 + y^2$ as our radial coordinate. This means that the line element along the wire is parametrised by $\mathrm{d}\mathbf{x}' = \hat{\mathbf{z}}\,\mathrm{d}z$ and, for a point $\mathbf{x}$ away from the wire, the vector $\mathrm{d}\mathbf{x}' \times (\mathbf{x} - \mathbf{x}')$ points along the tangent to the circle of radius r,

$$\mathrm{d}\mathbf{x}' \times (\mathbf{x} - \mathbf{x}') = r\hat{\boldsymbol{\varphi}}\,\mathrm{d}z\,. \tag{3.39}$$

So we have

$$\mathbf{B} = \frac{\mu_0 I \hat{\boldsymbol{\varphi}}}{4\pi} \int_{-\infty}^{+\infty} \mathrm{d}z \, \frac{r}{(r^2 + z^2)^{3/2}} = \frac{\mu_0 I}{2\pi r}\hat{\boldsymbol{\varphi}}, \tag{3.40}$$

which is the same result we found earlier (3.9).

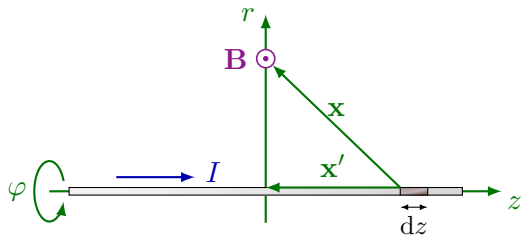


Fig. 3.6: An infinite, straight wire carrying current I and pointing in the $\hat{\mathbf{z}}$ direction

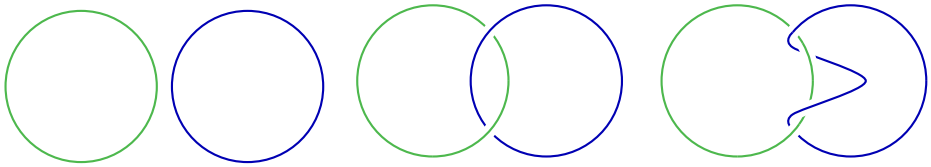


Fig. 3.7: Curves with linking number $n = 0$, $n = 1$ and $n = 2$.

3.2.4 A Mathematical Diversion: The Linking Number

There's a rather cute application of these ideas to pure mathematics. Consider two closed, non-intersecting curves, $\mathcal{C}$ and $\mathcal{C}'$, in $\mathbb{R}^3$. For each pair of curves, there is an integer $n \in \mathbb{Z}$ called the *linking number* which tells you how many times one of the curves winds around the other. For example, here are pairs of curves with linking number $|n| = 0$, 1 and 2.

To determine the sign of the linking number, we need to specify the orientation of each curve. In the last two figures above, the linking numbers are negative, if we traverse both red and blue curves in the same direction. The linking numbers are positive if we traverse one curve in a clockwise direction, and the other in an anti-clockwise direction.

Importantly, the linking number doesn't change as you deform either curve, provided that the two curves never cross. In fancy language, the linking number is an example of a topological invariant.

There is an integral expression for the linking number, first written down by Gauss during his exploration of electromagnetism. The Biot-Savart formula (3.38) offers a simple physics derivation of Gauss' expression. Suppose that the curve $\mathcal{C}$ carries a current I. This sets us a magnetic field everywhere in space. We will then compute $\oint_{\mathcal{C}'} \mathbf{B} \cdot d\mathbf{x}'$ around another curve $\mathcal{C}$. (If you want a justification for computing $\oint_{\mathcal{C}'} \mathbf{B} \cdot d\mathbf{x}'$ then you can think of it as the work done when transporting a magnetic monopole of unit charge around $\mathcal{C}$, but this interpretation isn't necessary for what follows.) The Biot-Savart formula gives

$$\oint_{\mathcal{C}'} \mathbf{B}(\mathbf{x}') \cdot d\mathbf{x}' = \frac{\mu_0 I}{4\pi} \oint_{\mathcal{C}'} d\mathbf{x}' \cdot \oint_{\mathcal{C}} \frac{d\mathbf{x}' \times (\mathbf{x}' - \mathbf{x})}{|\mathbf{x} - \mathbf{x}'|^3}, \tag{3.41}$$

where we've changed our conventions somewhat from (3.38): now $\mathbf{x}$ labels coordinates on $\mathcal{C}$ while $\mathbf{x}'$ labels coordinates on $\mathcal{C}'$.

Meanwhile, we can also use Stokes' theorem, followed by Ampére's law, to write

$$\oint_{\mathcal{C}'} \mathbf{B}(\mathbf{x}') \cdot d\mathbf{x}' = \int_{\mathcal{S}'} (\boldsymbol{\nabla} \times \mathbf{B}) \cdot d\mathbf{S} = \mu_0 \int_{\mathcal{S}'} \mathbf{J} \cdot d\mathbf{S}\,, \tag{3.42}$$

where $\mathcal{S}'$ is a surface bounded by $\mathcal{C}'$. The current is carried by the other curve, $\mathcal{C}$, which pierces $\mathcal{S}'$ precisely n times, so that

$$\oint_{\mathcal{C}'} \mathbf{B}(\mathbf{x}') \cdot \mathrm{d}\mathbf{x}' = \mu_0 \int_{\mathcal{S}'} \mathbf{J} \cdot \mathrm{d}\mathbf{S} = n\mu_0 I. \tag{3.43}$$

Comparing the two equations above, we arrive at Gauss' double-line integral expression for the linking number n,

$$n = \frac{1}{4\pi} \oint_{\mathcal{C}'} \mathrm{d}\mathbf{x}' \cdot \oint_{\mathcal{C}} \frac{\mathrm{d}\mathbf{x}' \times (\mathbf{x}' - \mathbf{x})}{|\mathbf{x} - \mathbf{x}'|^3}. \tag{3.44}$$

Note that our final expression is symmetric in $\mathcal{C}$ and $\mathcal{C}'$, even though these two curves played a rather different physical role in the original definition, with $\mathcal{C}$ carrying a current, and $\mathcal{C}'$ the path traced by some hypothetical monopole. To see that the expression is indeed symmetric, note that the triple product can be thought of as the determinant $\det(\mathbf{x}', \mathbf{x}, \mathbf{x}' - \mathbf{x})$. Swapping $\mathbf{x}$ and $\mathbf{x}'$ changes the order of the first two vectors and changes the sign of the third, leaving the determinant unaffected.

The formula (3.44) is rather pretty. It's not at all obvious that the right-hand-side doesn't change under (non-crossing) deformations of $\mathcal{C}$ and $\mathcal{C}'$; nor is it obvious that the right-hand-side must give an integer. Yet both are true, as the derivation above shows. This is the first time that ideas of topology sneak into physics. It's not the last.

3.3 Magnetic Dipoles

We've seen that the Maxwell equations forbid magnetic monopoles with a long-range $B \sim 1/r^2$ fall-off (3.23). So what is the generic fall-off for some distribution of currents which are localised in a region of space? In this section we will see that, if you're standing suitably far from the currents, you'll typically observe a dipole-like magnetic field.

3.3.1 A Current Loop

We start with a specific, simple example. Consider a circular loop of wire $\mathcal{C}$ of radius R carrying a current I. We can guess what the magnetic field looks like simply by patching together our result for straight wires: it must roughly take the shape shown in Fig. 3.8. However, we can be more accurate. Here we restrict ourselves only to the magnetic field far from the loop.

To compute the magnetic field far away, we won't start with the Biot-Savart law but instead return to the original expression for $\mathbf{A}$ given in (3.32). We're going to return to the notation in which a point in space is labelled as $\mathbf{r}$ rather than $\mathbf{x}$. (This is more appropriate for long-distance distance fields which are essentially an expansion in $r = |\mathbf{x}|$). The vector potential is then given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_{\mathcal{V}} \mathrm{d}^3 r' \, \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{3.45}$$

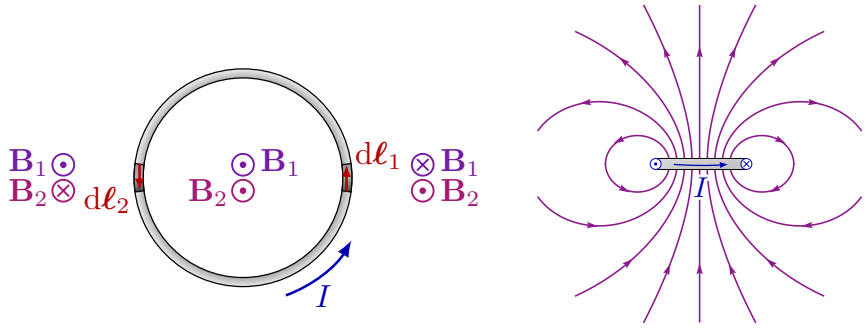


Fig. 3.8: The magnetic field for a circular loop of wire $\mathcal{C}$ of radius R carrying a current I

Writing this in terms of the current I (rather than the current density $\mathbf{J}$), we have

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi} \oint_{\mathcal{C}} \frac{\mathrm{d}\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}. \tag{3.46}$$

We want to ask what this looks like far from the loop. Just as we did for the electrostatic potential, we can Taylor expand the integrand using (2.50),

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \cdots . \tag{3.47}$$

So that

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi} \oint_{\mathcal{C}} \mathrm{d}\mathbf{r}' \left(\frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \cdots \right). \tag{3.48}$$

The first term in this expansion vanishes because we're integrating around a circle. This is just a reflection of the fact that there are no magnetic monopoles. For the second term, there's a way to write it in slightly more manageable form. To see this, let's introduce an arbitrary constant vector $\mathbf{g}$ and use this to look at

$$\oint_{\mathcal{C}} \mathrm{d}\mathbf{r}' \cdot \mathbf{g}(\mathbf{r}\cdot\mathbf{r}'). \tag{3.49}$$

Recall that, from the point of view of this integral, both $\mathbf{g}$ and $\mathbf{r}$ are constant vectors; it's the vector $\mathbf{r}'$ that we're integrating over. This is now the kind of line integral of a vector that allows us to use Stokes' theorem. We have

$$\oint_{\mathcal{C}} \mathrm{d}\mathbf{r}' \cdot \mathbf{g}(\mathbf{r}\cdot\mathbf{r}') = \int_{\mathcal{S}} \mathrm{d}\mathbf{S} \cdot \boldsymbol{\nabla} \times (\mathbf{g}(\mathbf{r}\cdot\mathbf{r}')) = \int_{\mathcal{S}} \mathrm{d}S_i \, \epsilon_{ijk} \partial'_j (g_k r_l r'_l), \tag{3.50}$$

where, in the final equality, we've resorted to index notation to help us remember what's connected to what. Now the derivative ∂' acts only on the r' and we get

$$\oint_{\mathcal{C}} \mathrm{d}\mathbf{r}' \cdot \mathbf{g}(\mathbf{r}\cdot\mathbf{r}') \int_{\mathcal{S}} \mathrm{d}S_i \, \epsilon_{ijk} g_k r_j = \mathbf{g} \cdot \int_{\mathcal{S}} \mathrm{d}\mathbf{S} \times \mathbf{r}. \tag{3.51}$$

But this is true for all constant vectors $\mathbf{g}$ which means that it must also hold as a vector identity once we strip away $\mathbf{g}$. We have

$$\oint_{\mathcal{C}} \mathrm{d}\mathbf{r}' \, (\mathbf{r}\cdot\mathbf{r}') = \boldsymbol{\mathcal{S}} \times \mathbf{r}, \tag{3.52}$$

where we've introduced the vector area $\boldsymbol{\mathcal{S}}$ of the surface $\mathcal{S}$ bounded by $\mathcal{C}$, defined as

$$\boldsymbol{\mathcal{S}} = \int_{\mathcal{S}} \mathrm{d}\mathbf{S} \tag{3.53}$$

If the boundary $C\mathcal{C}$ lies in a plane – as it does for us – then the vector $\boldsymbol{\mathcal{S}}$ points out of the plane.

Now let's apply this result to our vector potential (3.48). With the first term vanishing, we're left with

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\mathbf{r}}{r^3}, \tag{3.54}$$

where we've introduced the *magnetic dipole moment*

$$\mathbf{m} = I\boldsymbol{\mathcal{S}}. \tag{3.55}$$

This is our final, simple, answer for the long-range behaviour of the vector potential due to a current loop. It remains only to compute the magnetic field. A little algebra gives

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}-\mathbf{m}}{r^3}\right). \tag{3.56}$$

Now we see why $\mathbf{m}$ is called the magnetic dipole; this form of the magnetic field is exactly the same as the dipole electric field (2.47).

I stress that the $\mathbf{B}$ field due to a current loop and $\mathbf{E}$ field due to two charges don't look the same close up. But they have identical "dipole" long-range fall-offs.

3.3.2 General Current Distributions

We can now perform the same kind of expansion for a general current distribution $\mathbf{J}$ localised within some region of space. We use the Taylor expansion (2.50) in the general form of the vector potential (3.32),

$$A_i(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \mathrm{d}^3r'\,\frac{J_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} = \frac{\mu_0}{4\pi}\int \mathrm{d}^3r'\left(\frac{J_i(\mathbf{r}')}{r}+\frac{J_i(\mathbf{r}')(\mathbf{r}\cdot\mathbf{r}'}{r^3}+\cdots\right), \tag{3.57}$$

where we're using a combination of vector and index notation to help remember how the indices on the left and right-hand sides match up.

The first term above vanishes. Heuristically, this is because currents can't stop and end, they have to go around in loops. This means that the contribution from one part must be cancelled by the current somewhere else. To see this mathematically, we use the slightly odd identity

$$\partial_j(J_j r_i) = (\partial_j J_j)r_i + J_i = J_i, \tag{3.58}$$

where the last equality follows from the continuity condition $\boldsymbol{\nabla}\cdot\mathbf{J}=0$. Using this, we see that the first term in (3.57) is a total derivative (of $\partial/\partial r_i'$ rather than $\partial/\partial r_i$) which vanishes if we take the integral over $\mathbb{R}^3$ and keep the current localised within some interior region.

For the second term in (3.57) we use a similar trick, now with the identity

$$\partial_j(J_j r_i r_k) = (\partial_j J_j)r_i r_k + J_i r_k + J_i r_i = J_i r_k + J_k r_i. \tag{3.59}$$

Because $\mathbf{J}$ in (3.57) is a function of $\mathbf{r}'$, we actually need to apply this trick to the $J_i r_j'$ terms in the expression. We once again abandon the boundary term to infinity. Dropping

the argument of $\mathbf{J}$, we can use the identity above to write the relevant piece of the second term as

$$\int \mathrm{d}^3r'\, J_i r_j r'_j = \int \mathrm{d}^3r'\, \frac{r_j}{2}(J_i r'_j - J_j r'_i) = \int \mathrm{d}^3r'\, \frac{1}{2}(J_i(\mathbf{r}\cdot\mathbf{r}') - r'_i(\mathbf{J}\cdot\mathbf{r})). \tag{3.60}$$

But now this is in a form that is ripe for the vector product identity $\mathbf{a}\times(\mathbf{B}\times\mathbf{c}) = \mathbf{B}(\mathbf{a}\cdot\mathbf{c}) - \mathbf{c}(\mathbf{a}\cdot\mathbf{B})$. This means that we can rewrite this term as

$$\int \mathrm{d}^3r'\, \mathbf{J}(\mathbf{r}\cdot\mathbf{r}') = \frac{1}{2}\mathbf{r}\times\int \mathrm{d}^3r'\, \mathbf{J}\times\mathbf{r}'. \tag{3.61}$$

With this in hand, we see that the long distance fall-off of any current distribution again takes the dipole form (3.54)

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\mathbf{r}}{r^3}, \tag{3.62}$$

now with the magnetic dipole moment given by the integral,

$$\mathbf{m} = \frac{1}{2}\int \mathrm{d}^3r'\, \mathbf{r}'\times\mathbf{J}(\mathbf{r}'). \tag{3.63}$$

Just as in the electric case, the multipole expansion continues to higher terms. This time you need to use vector spherical harmonics. Just as in the electric case, if you want further details then look in Jackson.

3.4 Magnetic Forces

We've seen that a current produces a magnetic field. But a current is simply moving charge. And we know from the Lorentz force law that a charge q moving with velocity $\mathbf{v}$ will experience a force

$$\mathbf{F} = q\mathbf{v}\times\mathbf{B}. \tag{3.64}$$

This means that if a second current is placed somewhere in the neighbourhood of the first, then they will exert a force on one another. Our goal in this section is to figure out this force.

3.4.1 Force Between Currents

Let's start simple. Take two parallel wires carrying currents I_1 and I_2 respectively. We'll place them a distance d apart in the x-direction.

The current in the first wire sets up a magnetic field (3.9). So if the charges in the second wire are moving with velocity $\mathbf{v}$, they will each experience a force

$$\mathbf{F} = q\mathbf{v}\times\mathbf{B} = q\mathbf{v}\times\left(\frac{\mu_0 I_1}{2\pi d}\right)\hat{\mathbf{y}}, \tag{3.65}$$

where $\hat{\mathbf{y}}$ is the direction of the magnetic field experienced by the second wire as shown in Fig. 3.9. The next step is to write the velocity $\mathbf{v}$ in terms of the current $\mathbf{J}_2$ in the second

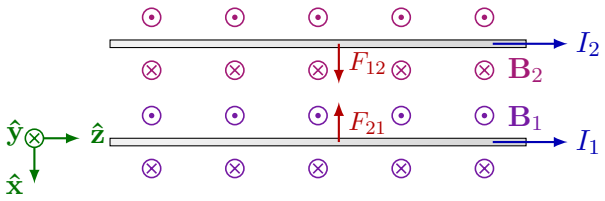


Fig. 3.9: Caption

wire. We did this in Section 1.1 when we first introduced the idea of currents: if there's a density n of these particles and each carries charge q, then the current density is

$$\mathbf{J}_2 = nq\mathbf{v}. \tag{3.66}$$

For a wire with cross-sectional area A, the total current is just $I_2 = J_2 A$. For our set-up, $\mathbf{J}_2 = J_2\hat{\mathbf{z}}$.

Finally, we want to compute the force on the wire per unit length, $\mathbf{f}$. Since the number of charges per unit length is nA and $\mathbf{F}$ is the force on each charge, we have

$$\mathbf{f} = nA\mathbf{F} = \left(\frac{\mu_0 I_1 I_2}{2\pi d}\right)\hat{\mathbf{z}} \times \hat{\mathbf{y}} = -\left(\frac{\mu_0 I_1 I_2}{2\pi d}\right)\hat{\mathbf{x}}. \tag{3.67}$$

This is our answer for the force between two parallel wires. If the two currents are in the same direction, so that $I_1 I_2 > 0$, the overall minus sign means that the force between two wires is attractive. For currents in opposite directions, with $I_1 I_2 < 0$, the force is repulsive.

3.4.1.1 The General Force Between Currents

We can extend our discussion to the force experienced between two current distributions $\mathbf{J}_1$ and $\mathbf{J}_2$. We start by considering the magnetic field $\mathbf{B}(\mathbf{r})$ due to the first current $\mathbf{J}_1$. As we've seen, the Biot-Savart law (3.37) tells us that this can be written as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \mathrm{d}^3 r' \, \frac{\mathbf{J}_1(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}. \tag{3.68}$$

If the current $\mathbf{J}_1$ is localised on a curve $\mathcal{C}_1$, then we can replace this volume integral with the line integral (3.38)

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0 I_1}{4\pi}\oint_{\mathcal{C}_1} \frac{\mathrm{d}\mathbf{r}_1 \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}. \tag{3.69}$$

Now we place a second current distribution $\mathbf{J}_2$ in this magnetic field. It experiences a force per unit area given by (1.9), so the total force is

$$\mathbf{F} = \int \mathrm{d}^3 r \, \mathbf{J}_2(\mathbf{r}) \times \mathbf{B}(\mathbf{r}). \tag{3.70}$$

Again, if the current $\mathbf{J}_2$ is restricted to lie on a curve $\mathcal{C}_2$, then this volume integral can be replaced by the line integral

$$\mathbf{F} = I_2 \oint_{\mathcal{C}_2} \mathrm{d}\mathbf{r} \times \mathbf{B}(\mathbf{r}), \tag{3.71}$$

and the force can now be expressed as a double line integral,

$$\mathbf{F} = \frac{\mu_0}{4\pi} I_1 I_2 \oint_{\mathcal{C}_1} \oint_{\mathcal{C}_2} \mathrm{d}\mathbf{r}_2 \times \left(\mathrm{d}\mathbf{r}_1 \times \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|^3} \right). \tag{3.72}$$

In general, this integral will be quite tricky to perform. However, if the currents are localised, and well-separated, there is a somewhat better approach where the force can be expressed purely in terms of the dipole moment of the current.

3.4.2 Force and Energy for a Dipole

We start by asking a slightly different question. We'll forget about the second current and just focus on the first: call it $\mathbf{J}(\mathbf{r})$. We'll place this current distribution in a magnetic field $\mathbf{B}(\mathbf{r})$ and ask: what force does it feel?

In general, there will be two kinds of forces. There will be a force on the centre of mass of the current distribution, which will make it move. There will also be a torque on the current distribution, which will want to make it re-orient itself with respect to the magnetic field. Here we're going to focus on the former. Rather remarkably, we'll see that we get the answer to the latter for free!

The Lorentz force experienced by the current distribution is

$$\mathbf{F} = \int_{\mathcal{V}} \mathrm{d}^3 r \, \mathbf{J}(\mathbf{r}) \times \mathbf{B}(\mathbf{r}). \tag{3.73}$$

We're going to assume that the current is localised in some small region $\mathbf{r} = \mathbf{R}$ and that the magnetic field $\mathbf{B}$ varies only slowly in this region. This allows us to Taylor expand

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}(\mathbf{R}) + (\mathbf{r} \cdot \boldsymbol{\nabla})\mathbf{B}(\mathbf{R}) + \cdots . \tag{3.74}$$

We then get the expression for the force

$$\mathbf{F} = -\mathbf{B}(\mathbf{R}) \times \int_{\mathcal{V}} \mathrm{d}^3 r \, \mathbf{J}(\mathbf{r}) + \int_{\mathcal{V}} \mathrm{d}^3 r \, \mathbf{J}(\mathbf{r}) \times [(\mathbf{r} \cdot \boldsymbol{\nabla})\mathbf{B}(\mathbf{R})] + \cdots . \tag{3.75}$$

The first term vanishes because the currents have to go around in loops; we've already seen a proof of this following equation (3.57). We're going to do some fiddly manipulations with the second term. To help us remember that the derivative $\boldsymbol{\nabla}$ is acting on $\mathbf{B}$, which is then evaluated at $\mathbf{R}$, we'll introduce a dummy variable $\mathbf{r}'$ and write the force as

$$\mathbf{F} = \int_{\mathcal{V}} \mathrm{d}^3 r \, \mathbf{J}(\mathbf{r}) \times \left[(\mathbf{r} \cdot \boldsymbol{\nabla}')\mathbf{B}(\mathbf{r}') \right] \Big|_{\mathbf{r}'=\mathbf{R}} . \tag{3.76}$$

Now we want to play around with this. First, using the fact that $\boldsymbol{\nabla} \times \mathbf{B} = 0$ in the vicinity of the second current, we're going to show, that we can rewrite the integrand as

$$\mathbf{J}(\mathbf{r}) \times \left[(\mathbf{r} \cdot \boldsymbol{\nabla}')\mathbf{B}(\mathbf{r}') \right] = -\boldsymbol{\nabla}' \times \left[(\mathbf{r} \cdot \mathbf{B}(\mathbf{r}'))\mathbf{J}(\mathbf{r}) \right]. \tag{3.77}$$

To see why this is true, it's simplest to rewrite it in index notation. After shuffling a couple of indices, what we want to show is:

$$\epsilon_{ijk} J_j(r) r_l \partial'_l B_k(r') = \epsilon_{ijk} J_j(r) r_l \partial'_k B_l(r'). \tag{3.78}$$

Or, subtracting one from the other,

$$\epsilon_{ijk} J_j(r) r_l (\partial'_l B_k(r') - \partial'_k B_l(r')) = 0. \tag{3.79}$$

But the terms in the brackets are the components of $\boldsymbol{\nabla} \times \mathbf{B}$ and so vanish. So our result is true and we can rewrite the force (3.76) as

$$\mathbf{F} = -\boldsymbol{\nabla}' \times \int_{\mathcal{V}} \mathrm{d}^3 r \; (\mathbf{r} \cdot \mathbf{B}(\mathbf{r}'))\mathbf{J}(\mathbf{r}) \Big|_{\mathbf{r}'=\mathbf{R}}. \tag{3.80}$$

Now we need to manipulate this a little more. We make use of the identity (3.61) where we replace the constant vector by $\mathbf{B}$. Thus, up to some relabelling, (3.61) is the same as

$$\int_{\mathcal{V}} \mathrm{d}^3 r \, (\mathbf{B} \cdot \mathbf{r})\mathbf{J} = \frac{1}{2}\mathbf{B} \times \int_{\mathcal{V}} \mathrm{d}^3 r \, \mathbf{J} \times \mathbf{r} = -\mathbf{B} \times \mathbf{m}, \tag{3.81}$$

where $\mathbf{m}$ is the magnetic dipole moment of the current distribution. Suddenly, our expression for the force is looking much nicer: it reads

$$\mathbf{F} = \boldsymbol{\nabla} \times (\mathbf{B} \times \mathbf{m}) \tag{3.82}$$

where we've dropped the $\mathbf{r}' = \mathbf{R}$ notation because, having lost the integral, there's no cause for confusion: the magnetic dipole $\mathbf{m}$ is a constant, while $\mathbf{B}$ varies in space. Now we invoke a standard vector product identity. Using $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$, this simplifies and we're left with a simple expression for the force on a dipole

$$\mathbf{F} = \boldsymbol{\nabla}(\mathbf{B} \cdot \mathbf{m}). \tag{3.83}$$

After all that work, we're left with something remarkably simple. Moreover, like many forces in Newtonian mechanics, it can be written as the gradient of a function. This function, of course, is the energy U of the dipole in the magnetic field,

$$U = -\mathbf{B} \cdot \mathbf{m}. \tag{3.84}$$

This is an important expression that will play a role in later courses in Quantum Mechanics and Statistical Physics. For now, we'll just highlight something clever: we derived (3.84) by considering the force on the centre of mass of the current. This is related to how U depends on $\mathbf{r}$. But our final expression also tells us how the energy depends on the orientation of the dipole $\mathbf{m}$ at fixed position. This is related to the torque. Computing the force gives us the torque for free. This is because, ultimately, both quantities are derived from the underlying energy.

3.4.2.1 The Force Between Dipoles

As a particular example of the force (3.83), consider the case where the magnetic field is set up by a dipole $\mathbf{m}_1$. We know that the resulting long-distance magnetic field is (3.63),

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3(\mathbf{m}_1 \cdot \hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}_1}{r^3}\right). \tag{3.85}$$

Now we'll consider how this affects the second dipole $\mathbf{m} = \mathbf{m}_2$. From (3.83), we have

$$\mathbf{F} = \frac{\mu_0}{4\pi}\boldsymbol{\nabla}\left(\frac{3(\mathbf{m}_1 \cdot \hat{\mathbf{r}})(\mathbf{m}_2 \cdot \hat{\mathbf{r}}) - \mathbf{m}_1 \cdot \mathbf{m}_2}{r^3}\right), \tag{3.86}$$

where $\mathbf{r}$ is the vector from $\mathbf{m}_1$ to $\mathbf{m}_2$. Note that the structure of the force is identical to that between two electric dipoles in (2.80). This is particularly pleasing because we used two rather different methods to calculate these forces. If we act with the derivative, we have

$$\mathbf{F} = \frac{3\mu_0}{4\pi r^4}[(\mathbf{m}_1 \cdot \hat{\mathbf{r}})\mathbf{m}_2 + (\mathbf{m}_2 \cdot \hat{\mathbf{r}})\mathbf{m}_1 + (\mathbf{m}_1 \cdot \mathbf{m}_2)\hat{\mathbf{r}} - 5(\mathbf{m}_1 \cdot \hat{\mathbf{r}})(\mathbf{m}_2 \cdot \hat{\mathbf{r}})\hat{\mathbf{r}}] \tag{3.87}$$

First note that if we swap $\mathbf{m}_1$ and $\mathbf{m}_2$, so that we also send $\mathbf{r} \to -\mathbf{r}$, then the force swaps sign. This is a manifestation of Newton's third law: every action has an equal and opposite reaction. Recall from Dynamics and Relativity lectures that we needed Newton's third law to prove the conservation of momentum of a collection of particles. We see that this holds for a bunch of dipoles in a magnetic field.

But there was also a second part to Newton's third law: to prove the conservation of angular momentum of a collection of particles, we needed the force to lie parallel to the separation of the two particles. And this is *not* true for the force (3.87). If you set up a collection of dipoles, they will start spinning, seemingly in contradiction of the conservation of angular momentum. What's going on?! Well, angular momentum is conserved, but you have to look elsewhere to see it. The angular momentum carried by the dipoles is compensated by the angular momentum carried by the magnetic field itself.

Finally, a few basic comments: the dipole force drops off as $1/r^4$, quicker than the Coulomb force. Correspondingly, it grows quicker than the Coulomb force at short distances. If $\mathbf{m}_1$ and $\mathbf{m}_2$ point in the same direction and lie parallel to the separation $\mathbf{R}$, then the force is attractive. If $\mathbf{m}_1$ and $\mathbf{m}_2$ point in opposite directions and lie parallel to the separation between them, then the force is repulsive. The expression (3.87) tells us the general result.

3.4.3 So What is a Magnet?

Until now, we've been talking about the magnetic field associated to electric currents. But when asked to envisage a magnet, most people would think if a piece of metal, possibly stuck to their fridge, possibly in the form of a bar magnet like the one shown in Fig. 3.10. How are these related to our discussion above?

These metals are permanent magnets. They often involve iron. They can be thought of as containing many microscopic magnetic dipoles, which align to form a large magnetic dipole $\mathbf{M}$. In a bar magnet, the dipole $\mathbf{M}$ points between the two poles. The iron filings in the picture trace out the magnetic field which takes the same form that we saw for the current loop in Section 3.3.

This means that the leading force between two magnets is described by our result (3.87). Suppose that $\mathbf{M}_1$, $\mathbf{M}_2$ and the separation $\mathbf{R}$ all lie along a line. If $\mathbf{M}_1$ and $\mathbf{M}_2$ point in the same direction, then the North pole of one magnet faces the South pole of another and (3.87) tells us that the force is attractive. Alternatively, if $\mathbf{M}_1$ and $\mathbf{M}_2$ point in opposite directions then two poles of the same type face each other and the force is repulsive. This, of course, is what we all learned as kids.

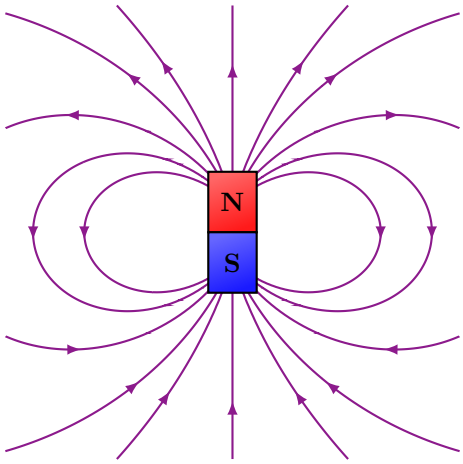


Fig. 3.10: Caption

The only remaining question is: where do the microscopic dipole moments $\mathbf{m}$ come from? You might think that these are due to tiny electric atomic currents but this isn't quite right. Instead, they have a more fundamental origin. The electric charges – which are electrons – possess an inherent angular momentum called *spin*. Roughly you can think of the electron as spinning around its own axis in much the same way as the Earth spins. But, ultimately, spin is a quantum mechanical phenomenon and this classical analogy breaks down when pushed too far. The magnitude of the spin is:

$$s = \frac{1}{2}\hbar, \tag{3.88}$$

where, recall, $\hbar$ has the same dimensions as angular momentum.

We can push the classical analogy of spin just a little further. Classically, an electrically charged spinning ball would give rise to a magnetic dipole moment. So one may wonder if the spinning electron also gives rise to a magnetic dipole. The answer is yes. It is given by

$$\mathbf{m} = g\frac{e}{2m}\mathbf{s}, \tag{3.89}$$

where e is the charge of the electron and m is its mass. The number g is dimensionless and called, rather uninspiringly, the *g-factor*. It has been one of the most important numbers in the history of theoretical physics, with several Nobel prizes awarded to people for correctly calculating it! The classical picture of a spinning electron suggests $g = 1$. But this is wrong. The first correct prediction (and, correspondingly, first Nobel prize) was by Dirac. His famous relativistic equation for the electron gives

$$g = 2. \tag{3.90}$$

Subsequently it was observed that Dirac's prediction is not quite right. The value of g receives corrections. The best current experimental value is

$$g = 2.00231930419922 \pm (1.5 \times 10^{-12}). \tag{3.91}$$

Rather astonishingly, this same value can be computed theoretically using the framework of quantum field theory (specifically, quantum electrodynamics). In terms of precision, this is one of the great triumphs of theoretical physics.

There is much much more to the story of magnetism, not least what causes the magnetic dipoles **m** to align themselves in a material. The details involve quantum mechanics and are beyond the scope of this course.

3.5 Magnetic Materials

To this point we have not mentioned magnetic materials, and yet the behaviour of a magnetic system will clearly change if a magnetisable material is introduced. As in the case of electrically polarisable materials, where induced electric dipoles change the behaviour of electric field lines, magnetically polarisable materials change the behaviour of magnetic field lines. There are, essentially, three different kinds of magnetic material:

- diamagnetic,
- paramagnetic,
- ferromagnetic.

Their different properties stem from the precise nature of the magnetic dipole moments of the atoms and molecules that make up the materials.

There are two main contributions to the magnetic dipole moment of an atom, or molecule: (i) the orbital motions of electrons correspond to current loops and thus form dipoles; (ii) the electrons themselves have intrinsic magnetic dipole moments arising from their quantum-mechanical spin. The sum of these two contributions determines the overall magnetic properties of most materials:

- Closed current loops react *diamagnetically*, which means that they oppose any applied external magnetic field, where *oppose* means that the net field is reduced, as a consequence of the dipole field combining with the external field.
- Some materials react *paramagnetically*, which means that permanent dipole moments of the material align themselves with the external field, leading to an enhancement of the field.
- Some materials behave *ferromagnetically*, which means that they behave in a way that is similar to paramagnetic materials, but the effect is considerably enhanced.

Regardless of the precise physical origin of the magnetic properties of materials, we require a general way of describing the macroscopic behaviour of materials when they are exposed to an external magnetic field.

3.5.1 Magnetisation Currents

The key concept is that magnetic materials acquire a net magnetic dipole moment when placed in a field. Assume, for simplicity, that this net dipole moment occurs as a result of

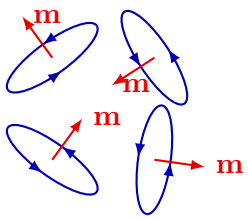


Fig. 3.11: Misaligned dipoles with no net dipole moment.

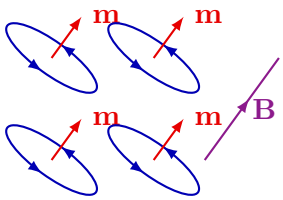


Fig. 3.12: Aligned dipoles with a net dipole moment.

microscopic circulating currents that align when a field is applied. If all of the dipoles are arranged randomly, as shown in Fig. 3.11, there is no net circulating current, and there is no net dipole moment. If the dipoles are aligned, as in Fig. 3.12, there is a net circulating current, and a net dipole moment. How can we describe how these individual dipoles add together to create an overall magnetic field?

Consider the situation where four loops of wire are connected together with a common edge, which points in the direction of $\hat{\mathbf{y}}$ (see Fig. 3.13). The four loops all contribute to the current I_y shown. The dipole moment due to I_1 is

$$m_z = I_1 \, \mathrm{d}x \, \mathrm{d}y \,, \tag{3.92}$$

and the dipole moment due to I_1' is

$$m_z' = I_1' \, \mathrm{d}x \, \mathrm{d}y \,. \tag{3.93}$$

We know, however, that

$$m_z' m_z + \frac{\partial m_z}{\partial x} \, \mathrm{d}x \,, \tag{3.94}$$

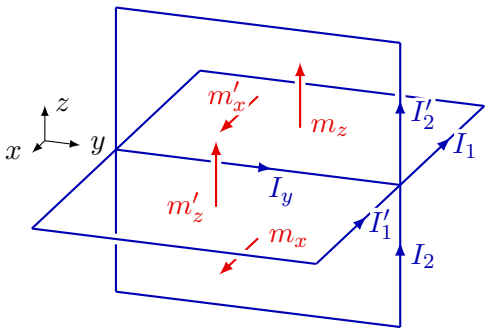


Fig. 3.13: Four rigidly connected current loops.

and therefore

$$\begin{aligned} m_z - m_z' &= -\frac{\partial m_z}{\partial x}\,\mathrm{d}x \\ &= (I_1 - I_1')\,\mathrm{d}x\,\mathrm{d}y \\ &= I_y^1\,\mathrm{d}x\,\mathrm{d}y\,, \end{aligned} \tag{3.95}$$

where I_y^1 is the net current in the central path due to I_1 and I_1'. Hence we find for I_y^1:

$$I_y^1\,\mathrm{d}y = -\frac{\partial m_z}{\partial x}. \tag{3.96}$$

Similarly, for the other pair of loops, the dipole moment due to I_2 is

$$m_x = I_2\,\mathrm{d}y\,\mathrm{d}z\,, \tag{3.97}$$

and the dipole moment due to I_2' is

$$m_x' = I_2'\,\mathrm{d}y\,\mathrm{d}z\,. \tag{3.98}$$

We know that

$$m_x' = m_x + \frac{\partial m_x}{\partial z}\,\mathrm{d}z\,, \tag{3.99}$$

and therefore

$$\begin{aligned} m_x - m_x' &= -\frac{\partial m_x}{\partial z}\,\mathrm{d}z \\ &= (I_2 - I_2')\,\mathrm{d}y\,\mathrm{d}z \\ &= -I_y^2\,\mathrm{d}y\,\mathrm{d}z\,, \end{aligned} \tag{3.100}$$

where I_y^2 is the net current in the central path due to I_2 and I_2'. Hence we find for I_y^2:

$$I_y^2\,\mathrm{d}y = \frac{\partial m_x}{\partial z}. \tag{3.101}$$

The *total current* on the central path due to all four loops I_y is

$$I_y = I_y^1 + I_y^2 = \frac{1}{\mathrm{d}y}\left[\frac{\partial m_x}{\partial z} - \frac{\partial m_z}{\partial x}\right], \tag{3.102}$$

which we can convert to the current density

$$J_y = \frac{1}{\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z}\left[\frac{\partial m_x}{\partial z} - \frac{\partial m_z}{\partial x}\right]. \tag{3.103}$$

We introduce a new quantity called the *magnetisation*, denoted by $\mathbf{M}$, which is the magnetic dipole moment per unit volume:

$$\mathbf{M} = \mathbf{m}\frac{1}{\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z}. \tag{3.104}$$

The magnetic dipole moment per unit volume, $\mathbf{M}$, which is a vector field, is very similar to the electric dipole moment per unit volume $\mathbf{P}$, introduced previously in the context of electrostatic systems.

The total dipole moment of an object, say $\mathbf{m}_{\text{tot}}$, can be found by integrating the dipole moment per unit volume over the object:

$$\mathbf{m}_{\text{tot}} = \int \mathrm{d}^3\mathbf{r}\,\mathbf{M}. \tag{3.105}$$

In other words, if an object is broken down into infinitesimally small elements, the dipole moment of the complete system is found by adding together the contributions from the elements.

If the individual dipole moments point in random directions, the vector sum over the whole object leads to a zero net dipole moment. If they are aligned, even partially, a net dipole moment results.

Returning to the original problem, Eq. (3.103) can be written

$$J_y = \left[\frac{\partial M_x}{\partial z} - \frac{\partial M_z}{\partial x}\right]. \tag{3.106}$$

The same procedure can be carried out for two other sets of loops having common paths in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions, resulting in a total of three equations:

$$J_x = \left[\frac{\partial M_z}{\partial y} - \frac{\partial M_y}{\partial z}\right], \tag{3.107}$$

$$J_y = \left[\frac{\partial M_x}{\partial z} - \frac{\partial M_z}{\partial x}\right], \tag{3.108}$$

$$J_z = \left[\frac{\partial M_y}{\partial x} - \frac{\partial M_x}{\partial y}\right], \tag{3.109}$$

from which the *magnetisation current density* is finally written

$$\boxed{\mathbf{J}_{\text{m}} = \boldsymbol{\nabla} \times \mathbf{M}} \tag{3.110}$$

We have introduced a subscript on the current to show that it represents the intrinsic magnetic behaviour of the individual atoms and molecules. It must be distinguished from, and is in addition to, the field produced by the movement of free charge. In fact it can be a fictitious current introduced merely to represent a magnetic dipole moment that has physical origins other than current. Again, this model is analogous to the electrostatic case, where we distinguished between externally applied free charge, and the separation of bound (polarisation) charge, which was associated entirely with the materials used in the system.

It is clear that if the magnetisation is uniform, the curl, which is calculated through spatial derivatives, is zero, and the magnetisation current is zero. This is equivalent to the current on neighbouring loops summing to zero on the common path.

3.5.2 Surface Magnetisation Currents

If the magnetisation is uniform within some object, then the magnetisation currents must reside on the surface: conversely, for example, the current in the wires of a solenoid can be regarded as a *surface* current that produces a uniform magnetic field.

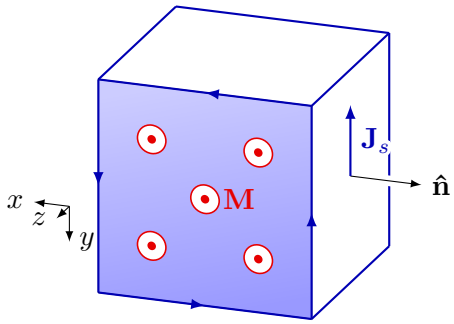


Fig. 3.14: Sheet current density $\mathbf{J}_s$ arising from the magnetisation $\mathbf{M}$.

To show this, consider a uniformly magnetised block (see Fig. 3.14). All the little current loops corresponding to the magnetic dipoles cancel each other out inside, leaving a sheet of current flowing around the surface. We call this a surface current density $\mathbf{J}_s$ [$\mathrm{A\,m^{-1}}$], like the current density but with one dimension shrunk to zero. So it is rather like surface charge σ flowing in the surface. The current flowing in the x-direction in a surface in the xz plane scales only with the extent in the z-direction.

Let the block have sides of length $\mathrm{d}x$, $\mathrm{d}y$ and $\mathrm{d}z$, and magnetisation M_z in the z-direction. Then the current I in the surface whose cross-section is shown is given by $I = |\mathbf{J}_s|\,\mathrm{d}z$ and $M_z\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = I\,\mathrm{d}x\,\mathrm{d}y$. So $m_z = |\mathbf{J}_s|$, and $\mathbf{J}_s$ is perpendicular to $\mathbf{M}$ and to the unit normal to the surface $\hat{\mathbf{n}}$.

The *surface current density* is therefore

$$\boxed{\mathbf{J}_s = \mathbf{M} \times \hat{\mathbf{n}}.} \tag{3.111}$$

As an aside reminder, the surface current density is introduced merely to represent a magnetic dipole moment that may have physical origins other than current.

3.5.3 Magnetic Field Strength

The total current in a material comprises the true conduction current, which is associated with the bulk movement of free charge, and the magnetisation current, which represents the intrinsic internal magnetic behaviour of the atoms and molecules.

Ampére's law can be invoked by simply taking into account this extra current:

$$\boldsymbol{\nabla} \times \mathbf{B} = \mu_0[\mathbf{J}_{\mathrm{free}} + \mathbf{J}_{\mathrm{m}}], \tag{3.112}$$

where we explicitly distinguish between free current $\mathbf{J}_{\mathrm{free}}$ and magnetisation current $\mathbf{J}_{\mathrm{m}}$. Remember that $\mathbf{B}$ simply gives the force on some small test current placed in the system, so it is perfectly reasonable, and necessary, to add on the magnetisation currents when calculating $\mathbf{B}$.

Simple manipulation using Eq. (3.110) gives

$$\boldsymbol{\nabla} \times [\mathbf{B} - \mu_0\mathbf{M}] = \mu_0\mathbf{J}_{\mathrm{free}}. \tag{3.113}$$

Define the *magnetic field strength* $\mathbf{H}$ in magnetic materials, according to

$$\mu_0\mathbf{H} = \mathbf{B} - \mu_0\mathbf{M} \tag{3.114}$$

i.e.,

$$\boxed{\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}),} \tag{3.115}$$

such that

$$\boldsymbol{\nabla} \times \mathbf{H} = \mathbf{J}_{\text{free}}. \tag{3.116}$$

The magnetic field strength has the advantage that it is related directly to the conduction current flowing at a point, regardless of the magnetic properties of the material in which the current is flowing. Of course if we wish to know the magnetic field $\mathbf{B}$, we would need to know $\mathbf{M}$.

Using Stokes' theorem, we have

$$\begin{aligned}\int \mathrm{d}\mathbf{S} \cdot \boldsymbol{\nabla} \times \mathbf{H}(\mathbf{r}) &= \oint_{\mathcal{C}=\partial\mathcal{S}} \mathrm{d}\mathbf{l} \cdot \mathbf{H}(\mathbf{r}), \\ &= \int \mathrm{d}\mathbf{S} \cdot \mathbf{J}_{\text{free}}(\mathbf{r}), \\ &= I,\end{aligned} \tag{3.117}$$

where specific reference to position has been included for clarity. Therefore

$$\oint \mathrm{d}\mathbf{l} \cdot \mathbf{H}(\mathbf{r}) = I. \tag{3.118}$$

From now on we will always assume that $\mathbf{J}(\mathbf{r})$, without the subscript, corresponds to conduction (free) current.

Equations (3.116) and (3.118) constitute Ampére's law when magnetic materials are present. They are entirely consistent with our previous definition of Ampére's law, when no magnetisable material was present. Therefore, if we calculate the line integral of the magnetic field strength $\mathbf{H}$ around any closed path, the result is equal to the current flowing through any surface that has the path as its bounding edge. It can be seen that the magnetic field strength is intimately related to the flow of conduction current. When written in this way, Ampére's law implicitly accounts for the properties of the materials making up the system.

For the purpose of calculating forces, we require $\mathbf{B}$, which in turn requires us to know $\mathbf{M}$, as appreciated by inspection of (3.114). In reality, as external current is applied and increased, $\mathbf{H}$ increases, and $\mathbf{M}$, the dipole moment per unit volume, increases, as the state of magnetisation of the material changes.

For many materials, at least in the small-field limit, $\mathbf{M}$ is linearly proportional to $\mathbf{H}$, and we define

$$\mathbf{M} = \chi_{\text{m}}\mathbf{H}, \tag{3.119}$$

where χ_{m} is a constant of proportionality called the *magnetic susceptibility*.

Thus, Eq. (3.114) becomes

$$\mu_0\mathbf{H} = \mathbf{B} - \mu_0\mathbf{M} \quad \Longrightarrow \quad \mu_0\mathbf{H} = \mathbf{B} - \mu_0\chi_{\text{m}}\mathbf{H}, \tag{3.120}$$

which on rearranging produces

$$\mathbf{B} = \mu_0 \underbrace{(1 + \chi_\mathrm{m})}_{\equiv \mu} \mathbf{H}. \tag{3.121}$$

Now we can rewrite the relationship between the *magnetic field* and the *magnetic field strength*:

$$\mathbf{B} = \mu_0 \mu \mathbf{H}, \tag{3.122}$$

where the *relative permeability* μ of the material is given by

$$\mu = 1 + \chi_\mathrm{m}, \tag{3.123}$$

note that μ is sometimes written as μ_r.

Also note that μ implicitly accounts for the change in the magnetic state of the material that results from a magnetic field $\mathbf{H}$ being applied. It is common practice to use μ when describing the behaviour of magnetic fields, and χ_m when describing the detailed properties of magnetic materials.

For most insulators, which are non-magnetic, $\mu \approx 1$, whereas for magnetic materials, μ can take on very large values. The magnetic susceptibility can take on both positive and negative values, which was not true of the electric susceptibility. In fact, materials are defined in the following way:

- *Paramagnetic* materials have a positive susceptibility.
- *Diamagnetic* materials have a negative susceptibility.
- *Ferromagnetic* materials have an exceedingly high positive susceptibility.

Ferromagnetic materials are also often non-linear for relatively low field strengths, meaning that a $\mathbf{B} - \mathbf{H}$ curve needs to be used to describe the properties of a material, rather than just using a single constant of proportionality. This non-linearity gives rise to saturation in transformers, where the waveform of the current at the output does not follow the waveform of the current at the input. In fact, most magnetic materials are highly hysteretic, which means that a graph (see Fig. 3.15) of $\mathbf{B}$ against $\mathbf{H}$ describes a loop as the applied magnetic field is increased, decreased, and reversed in sign.

A further complication is that certain materials may be associated with a magnetic field even in the absence of external excitation. In this case, the material is said to be *permanently magnetised*. The material has a constant dipole moment per unit volume, sometimes denoted $\mathbf{M}_0$, which is not at all related to the applied field $\mathbf{H}$.

Finally, $\mathbf{B}$ and $\mathbf{H}$ do not have to lie in the same direction, in which case it is necessary to relate all of the components of $\mathbf{B}$ to all of the components of $\mathbf{H}$ through a matrix, or tensor. It is clear that the magnetic properties of materials can be quite complicated.

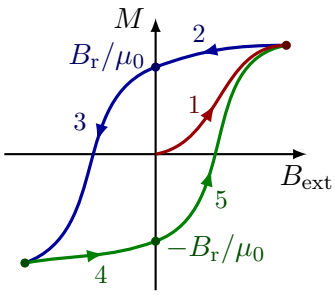


Fig. 3.15: (1): The material follows a non-linear magnetisation curve when magnetised from a zero field value. (2) and (4): When driving magnetic field drops to zero, the ferromagnetic material retains a considerable degree of magnetisation. (3) The driving magnetic field must be reversed and increased to a large value to drive the magnetisation to zero again.

3.5.4 Inhomogeneous Magnetic Materials and Boundary Conditions

Magnetic systems rarely have homogeneous magnetic properties, and therefore we are led to looking at situations where the magnetic properties change, possibly abruptly, at some boundary. As in the case of electrostatic problems, we must consider the magnetic field **B** and the magnetic field strength **H** separately.

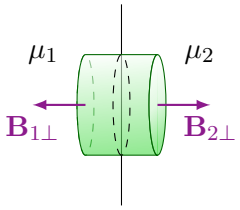


Fig. 3.16: A pillbox on the boundary between dissimilar magnetic materials.

Set up a 'pillbox' that cuts across the boundary between two different magnetic materials, as shown in Fig. 3.16. We know that, according to Maxwell's equation (3.2) for the magnetic field **B**,

$$\oint_S \mathrm{d}\mathbf{S} \cdot \mathbf{B}(\mathbf{r}) = 0. \tag{3.124}$$

Shrink the parallel faces of the pillbox down so that they are separated by an infinitesimally small distance. Also, make the parallel faces small enough so that $\mathbf{B}(\mathbf{r})$ is essentially constant over the surfaces. If the normal component of **B** in region 1 is called $\mathbf{B}_{1\perp}$, and the normal component of **B** in region 2 is called $\mathbf{B}_{2\perp}$, then

$$A\mathbf{B}_{2\perp} - A\mathbf{B}_{1\perp} = 0, \tag{3.125}$$

where A is the area, from which it follows that

$$\mathbf{B}_{1\perp} = \mathbf{B}_{2\perp}. \tag{3.126}$$

The normal component of **B** is continuous across a boundary, regardless of whether the relative permeability changes or not. The normal component of **H** must change discontinuously across a boundary because $\mathbf{B} = \mu_0\mu\mathbf{H}$.

Now set up an imaginary loop that cuts across the boundary between two dissimilar magnetic materials, as shown in Fig. 3.17. Be cause there is no conduction (free) current

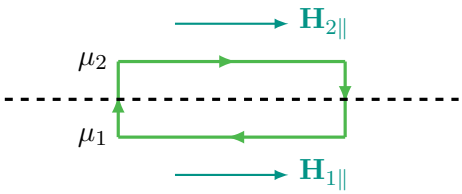


Fig. 3.17: A loop on the boundary between dissimilar magnetic materials

on the surface

$$\oint \mathrm{d}\mathbf{l} \cdot \mathbf{H}(\mathbf{r}) = 0. \tag{3.127}$$

If the sides of the loop, having length L, are brought infinitely close together,

$$L\mathbf{H}_{2\parallel} - L\mathbf{H}_{1\parallel} = 0, \tag{3.128}$$

from which it follows that

$$\mathbf{H}_{1\parallel} = \mathbf{H}_{2\parallel}. \tag{3.129}$$

The tangential component of the magnetic field strength $\mathbf{H}$ is continuous across a boundary, regardless of whether the relative permeability changes or not. The tangential component of $\mathbf{B}$ must change discontinuously across a boundary because $\mathbf{B} = \mu_0\mu\mathbf{H}$.

In conclusion, across a magnetic boundary,

- The normal component of $\mathbf{B}$ is continuous ($\mathbf{B}_\perp$ continuous).
- The normal component of $\mathbf{H}$ is discontinuous.
- The parallel component of $\mathbf{H}$ is continuous ($\mathbf{H}_\parallel$ continuous).
- The parallel component of $\mathbf{B}$ is discontinuous.

These observations are comparable to the boundary conditions for celectric fields.

It seems that magnetic fields have much in common with electric fields, and it is worthwhile comparing the similarities, and highlighting the differences:

- The magnetic field (magnetic flux density) $\mathbf{B}$ is the fundamental quantity in magnetostatics, in the sense that it gives directly the force on a test current.

  The electric field $\mathbf{E}$ is the fundamental quantity in electrostatics because it gives the force on a test charge.

- The magnetic field strength $\mathbf{H}$ is defined for the convenience of including magnetic materials, and it is related to $\mathbf{B}$ through $\mathbf{B} = \mu_0\mu\mathbf{H}$.

  The electric displacement (electric flux density) $\mathbf{D}$ is defined for the purpose of including dielectric materials, and it is related to the electric field $\mathbf{E}$ through $\mathbf{D} = \epsilon\epsilon_0\mathbf{E}$.

- In terms of the boundary conditions, $\mathbf{B}$ and $\mathbf{D}$ are alike, and $\mathbf{H}$ and $\mathbf{E}$ are alike.

- Because the normal components of **B** and **D** are continuous across a boundary, they can be imagined, for conceptual purposes, to have the properties of fluxes, which must be conserved.

In the case of electric fields, we saw that free charge is the source of **D**, such that field lines associated with **D** can only start and end on free charge, whereas field lines associated with **E** can start and end on either free charge or bound charge. In the case of magnetic fields, the movement of free charge, in the sense of electrical current, is the source of **H**, whereas both electrical current and magnetisation current are sources of **B**. There seem to be magnetisation currents on the outer surface of a magnetic material to which the external magnetic field lines are pinned. We can also see how the direction of magnetic field lines changes at a magnetic boundary in the same way that the direction of electric field lines changes at a dielectric boundary.

3.5.5 Boundary-Value Problems with Magnetic Materials

How do we calculate the form of the magnetic field lines when electric currents and magnetic bodies are present? The modern approach would be to carry out numerical solutions on a computer, but systems with certain symmetries can be solved relatively easily. Few examples will be considered in the following, with increasing level of complexity.

3.5.5.1 Long Thin Rod Parallel to Uniform Field

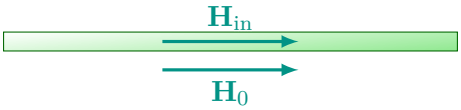


Fig. 3.18: Long thin rod in a parallel magnetic field.

Consider a long thin rod of magnetic material in a uniform field, as shown in Fig. 3.18. Because the tangential component of **H** must be continuous across the boundary, the internal magnetic field strength must be the same as the external field:

$$\mathbf{H}_{\text{in}} = \mathbf{H}_0, \tag{3.130}$$

from which it follows that

$$\mathbf{B}_{\text{in}} = \mu \mathbf{B}_0, \tag{3.131}$$

For paramagnetic and ferromagnetic materials ($\mu > 1$), the number of flux lines per unit area is larger on the inside than on the outside; the field is concentrated in the material.

For diamagnetic materials ($\mu < 1$), the number of flux lines per unit area is smaller on the inside than on the outside. Flux appears to be expelled.

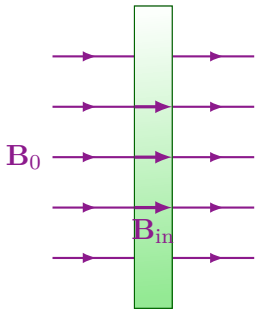


Fig. 3.19: Thin slab perpendicular to a uniform magnetic field.

3.5.5.2 Thin Slab Perpendicular to Uniform Field

Consider a slab of magnetic material perpendicular to a uniform field (much thinner than its width so that field lines remain parallel), as shown in Fig. 3.19. Because the perpendicular component of $\mathbf{B}$ must be continuous across the boundary, the internal magnetic field must be the same as the external magnetic field:

$$\mathbf{B}_{\text{in}} = \mathbf{B}_0, \tag{3.132}$$

which is essentially the conservation of flux across the boundary. It follows that

$$\mathbf{H}_{\text{in}} = \frac{1}{\mu}\mathbf{H}_0. \tag{3.133}$$

For paramagnetic and ferromagnetic materials the magnetic field strength inside the material is less than that outside the material.

For diamagnetic materials, the magnetic field strength internally is greater than that externally.

3.5.5.3 Magnetisable Sphere in a Uniform Field

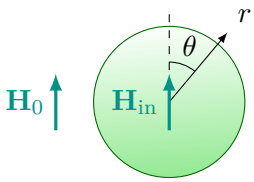


Fig. 3.20: Magnetisable sphere in a uniform magnetic field.

Consider a magnetisable (either diamagnetic or paramagnetic) sphere in a uniform external field, as shown in Fig. 3.20. This problem can be solved in an almost identical way to that of the dielectric sphere. We shall use the magnetic scalar potential $\phi_{\text{m}}(\mathbf{r})$, which is related to the magnetic field strength through

$$\mathbf{H}(\mathbf{r}) = -\boldsymbol{\nabla}\phi_{\text{m}}(\mathbf{r}). \tag{3.134}$$

Notice that because $\mathbf{H}$ is finite at the boundary, the normal derivative of ϕ_{m} is finite, and therefore ϕ_{m} must be continuous.

As in the case of a dielectric sphere, let us assume that the internal field is uniform, and the external field is the externally applied uniform field plus a dipole field generated by the magnetisation current. The potential becomes

$$\phi_{\rm m}(\mathbf{r}) = -H_{\rm in} r\cos\theta \quad \text{for} \quad r < a \tag{3.135}$$

$$\phi_{\rm m}(\mathbf{r}) = -H_0 r\cos\theta + \frac{A\cos\theta}{r^2} \quad \text{for} \quad r > a, \tag{3.136}$$

where A is some constant of proportionality whose value is to be found.

Because the potential is continuous across the boundary, or using $H_{\parallel}$ continuous,

$$H_{\rm in} a\cos\theta = H_0 a\cos\theta - \frac{A\cos\theta}{a^2}, \tag{3.137}$$

or,

$$H_{\rm in} = H_0 - \frac{A}{a^3}. \tag{3.138}$$

We also require that the normal component of $\mathbf{B}$ be continuous, where

$$B_\perp = B_r = -\mu\mu_0 \frac{\partial \phi_{\rm m}}{\partial r}, \tag{3.139}$$

at $r = a$, which gives, using Eq. (3.135)

$$\mu\mu_0 H_{\rm in} = \mu_0\left(H_0 + \frac{2A}{a^3}\right). \tag{3.140}$$

Solving for A gives

$$A = \frac{\mu - 1}{\mu + 2} H_0 a^3, \tag{3.141}$$

from which it follows that

$$H_{\rm in} = \frac{3}{\mu + 2} H_0. \tag{3.142}$$

This result is the same as that for the dielectric sphere with $\epsilon \to \mu$.

3.5.5.4 Uniformly Magnetised Cylinder

Consider a cylinder that is uniformly magnetised along its length. In fact, this is a reasonable model for a bar magnet. In this case, magnetisation currents flow around the surface of the cylinder, in much the same way as the current in the wires of a solenoid. As a consequence, a cylindrical bar magnet has the same $\mathbf{B}$-field as a short solenoid (see Fig. 3.21).

3.6 Units of Electromagnetism

More than any other subject, electromagnetism is awash with different units. In large part this is because electromagnetism has such diverse applications and everyone from

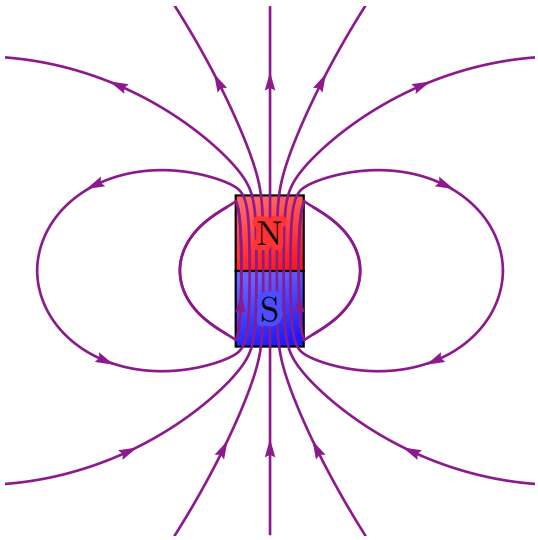


Fig. 3.21: The B-field of a magnetised cylinder.

astronomers, to electrical engineers, to particle physicists needs to use it. But it's still annoying. Here we explain the basics of SI units.

The SI unit of charge is the *Coulomb*. As of 2019[1], the Coulomb is defined in terms of the charge $-e$ carried by the electron. This is taken to be exactly

$$e = 1.602176634 \times 10^{-19}\text{C}. \tag{3.143}$$

If you rub a balloon on your sweater, it picks up a charge of around 10^{-6}C or so. A bolt of lightening deposits a charge of about 15C. The total charge that passes through an AA battery in its lifetime is about 5000C.

The SI unit of current is the *Ampere*, denoted A. It is defined as one Coulomb of charge passing every second. The current that runs through single ion channels in cell membranes is about 10^{-12}A. The current that powers your toaster is around 1A to 10A. There is a current in the Earth's atmosphere, known as the Birkeland current, which creates the aurora and varies between 10^5A and 10^6A. Galactic size currents in so-called Seyfert galaxies (particularly active galaxies) have been measured at a whopping 10^{18}A.

The electric field is measured in units of $\text{N}\,\text{C}^{-1}$. The electrostatic potential ϕ has units of *Volts*, denoted V, where the 1 Volt is the potential difference between two infinite, parallel plates, separated by 1m, which create an electric field of $1\text{N}\,\text{C}^{-1}$.

[1]Prior to 2019, a reluctance to rely on fundamental physics meant that the definitions were a little more tortuous. The Ampere was taken to be the base unit, and the Coulomb was defined as the amount of charge transported by a current of 1A in a second. The Ampere, in turn, was defined to be the current carried by two straight, parallel wires when separated by a distance of 1m, in order to experience an attractive force-per-unit-length of $2 \times 10^{-7}\text{N}\,\text{m}^{-1}$. (Recall that a Newton is the unit of force needed to accelerate 1kg at $1\text{m}\,\text{s}^{-1}$.) From our result (3.67), we see that if we plug in $I_1 = I_2 = 1\text{A}$ and $d = 1\text{m}$ then this force is $f = \mu_0/2\pi\text{A}^2\,\text{m}^{-1}$ This definition is the reason that μ_0 has the strange-looking value $\mu_0 = 4\pi \times 10^{-7}\text{m}\,\text{kg}\,\text{C}^{-2}$. The new definitions of SI units means that we can no longer say with certainty that $\mu_0 = 4\pi \times 10^{-7}\text{m}\,\text{kg}\,\text{C}^{-2}$, but this only holds up to the experimental accuracy of a dozen significant figures or so. For our purposes, the main lesson to draw from this is that, from the perspective of fundamental physics, SI units are arbitrary and a little daft.

A nerve cell sits at around 10^{-2}V. An AA battery sits at 1.5V. The largest manmade voltage is 10^7V produced in a van der Graaf generator. This doesn't compete well with what Nature is capable of. The potential difference between the ends of a lightening bolt can be 10^8. The voltage around a pulsar (a spinning neutron star) can be 10^{15}V.

The unit of a magnetic field is the *Tesla*, denoted T. A particle of charge 1C, passing through a magnetic field of 1T at $1\text{m}\,\text{s}^{-1}$ will experience a force of 1N. From the examples that we've seen above it's clear that 1C is a lot of charge. Correspondingly, 1T is a big magnetic field. Our best instruments (SQUIDs) can detect changes in magnetic fields of 10^{-18}T. The magnetic field in your brain is 10^{-12}T. The strength of the Earth's magnetic field is around 10^{-5}T while a magnet stuck to your fridge has about 10^{-3}T. The strongest magnetic field we can create on Earth is around 100T. Again, Nature beats us quite considerably. The magnetic field around neutron stars can be between 10^6T and 10^9T. (There is an exception here: in "heavy ion collisions", in which gold or lead nuclei are smashed together in particle colliders, it is thought that magnetic fields comparable to those of neutron stars are created. However, these magnetic fields are fleeting and small. They are stretch over the size of a nucleus and last for a millionth of a second or so).

As the above discussion amply demonstrates, SI units are based entirely on historical convention rather than any deep underlying physics. A much better choice is to pick units of charge such that we can discard ϵ_0 and μ_0. There are two commonly used frameworks that do this, called *Lorentz-Heaviside* units and *Gaussian* units. I should warn you that the Maxwell equations take a slightly different form in each.

To fully embrace natural units, we should also set the speed of light $c = 1$. (See the rant in the Dynamics and Relativity lectures). However we can't set everything to one. There is one combination of the fundamental constants of Nature which is dimensionless. It is known as the *fine structure constant*,

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}, \tag{3.144}$$

and takes value $\alpha \approx 1/137$. Ultimately, this is the correct measure of the strength of the electromagnetic force. It tells us that, in units with $\epsilon_0 = \hbar = c = 1$, the natural, dimensionless value of the charge of the electron is $e \approx 0.3$.

Chapter 4

Electrodynamics

For static situations, Maxwell's equations split into the equations of electrostatics, (2.1) and (2.2), and the equations of magnetostatics, (3.1) and (3.2). The only hint that there is a relationship between electric and magnetic fields comes from the fact that they are both sourced by charge: electric fields by stationary charge; magnetic fields by moving charge. In this section we will see that the connection becomes more direct when things change with time.

4.1 Faraday's Law of Induction

One of the Maxwell equations relates time varying magnetic fields to electric fields,

$$\boldsymbol{\nabla} \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0. \tag{4.1}$$

This equation tells us that if you change a magnetic field, you'll create an electric field. In turn, this electric field can be used to accelerate charges which, in this context, is usually thought of as creating a current in wire. The process of creating a current through changing magnetic fields is called *induction*.

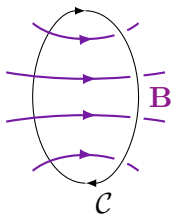


Fig. 4.1: Caption

We'll consider a wire to be a conductor, stretched along a stationary, closed curve, $\mathcal{C}$, as shown in the figure. We will refer to closed wires of this type as a "circuit". We integrate both sides of (4.1) over a surface $\mathcal{S}$ which is bounded by $\mathcal{C}$,

$$\int_{\mathcal{S}} (\boldsymbol{\nabla} \times \mathbf{E}) \cdot \mathrm{d}\mathbf{S} = -\int_{\mathcal{S}} \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{S}\,. \tag{4.2}$$

By Stokes theorem, we can write this as

$$\int_{\mathcal{C}} \mathbf{E} \cdot \mathrm{d}\mathbf{r} = -\int_{\mathcal{S}} \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{S} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{S}} \mathbf{B} \cdot \mathrm{d}\mathbf{S}\,. \tag{4.3}$$

Recall that the line integral around $\mathcal{C}$ should be in the right-handed sense; if the fingers on your right-hand curl around $\mathcal{C}$ then your thumb points in the direction of $\mathrm{d}\mathbf{S}$. (This means that in the figure $\mathrm{d}\mathbf{S}$ points in the same direction as $\mathbf{B}$). To get the last equality above, we need to use the fact that neither $\mathcal{C}$ nor $\mathcal{S}$ change with time. Both sides of this

equation are usually given names. The integral of the electric field around the curve $\mathcal{C}$ is called the *electromotive force*, $\mathcal{E}$, or *emf* for short,

$$\mathcal{E} = \int_{\mathcal{C}} \mathbf{E} \cdot \mathrm{d}\mathbf{r} \,. \tag{4.4}$$

It's not a great name because the electromotive force is not really a force. Instead it's the tangential component of the force per unit charge, integrated along the wire. Another way to think about it is as the work done on a unit charge moving around the curve $\mathcal{C}$. If there is a non-zero emf present then the charges will be accelerated around the wire, giving rise to a current.

The integral of the magnetic field over the surface $\mathcal{S}$ is called the magnetic *flux* Φ through $\mathcal{S}$,

$$\Phi = \int_{\mathcal{S}} \mathbf{B} \cdot \mathrm{d}\mathbf{S} \,. \tag{4.5}$$

The Maxwell equation (4.1) can be written as

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{4.6}$$

In this form, the equation is usually called *Faraday's Law*. Sometimes it is called the flux rule.

Faraday's law tells us that if you change the magnetic flux through $\mathcal{S}$ then a current will flow. There are a number of ways to change the magnetic field. You could simply move a bar magnet in the presence of circuit, passing it through the surface $\mathcal{S}$; or you could replace the bar magnet with some other current density, restricted to a second wire $\mathcal{C}'$, and move that; or you could keep the second wire $\mathcal{C}'$ fixed and vary the current in it, perhaps turning it on and off. All of these will induce a current in $\mathcal{C}$.

However, there is then a secondary effect. When a current flows in $\mathcal{C}$, it will create its own magnetic field. We've seen how this works for steady currents in Chapter 3. This induced magnetic field will always be in the direction that opposes the change. This is called *Lenz's law*. If you like, "Lenz's law" is really just the minus sign in Faraday's law (4.6).

We can illustrate this with a simple example. Consider the case where $\mathcal{C}$ is a circle, lying in a plane. We'll place it in a uniform $\mathbf{B}$ field and then make $\mathbf{B}$ smaller over time, so $\dot{\Phi} < 0$. By Faraday's law, $\mathcal{E} > 0$ and the current will flow in the right-handed direction around $\mathcal{C}$ as shown in Fig 4.2. But now you can wrap your right-hand in a different way: point your thumb in the direction of the current and let your fingers curl to show you the direction of the induced magnetic field. These are the circles drawn in Fig 4.2. You see that the induced current causes $\mathbf{B}$ to increase inside the loop, counteracting the original decrease.

Lenz's law is rather like a law of inertia for magnetic fields. It is necessary that it works this way simply to ensure energy conservation: if the induced magnetic field aided the process, we'd get an unstable runaway situation in which both currents and magnetic fields were increasing forever.

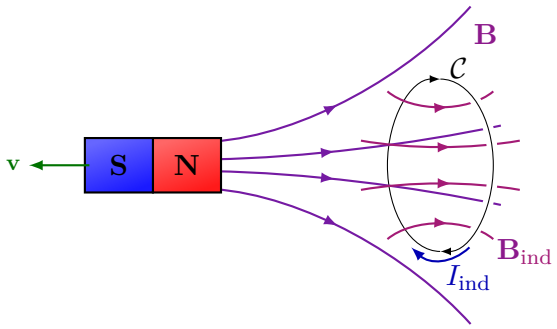


Fig. 4.2: Caption

4.1.1 Faraday's Law for Moving Wires

There is another, related way to induce currents in the presence of a magnetic field: you can keep the field fixed, but move the wire. Perhaps the simplest example is shown in Fig. 4.3: it's a rectangular circuit, but where one of the wires is a metal bar that can slide backwards and forwards. This whole set-up is then placed in a magnetic field, which passes up, perpendicular through the circuit.

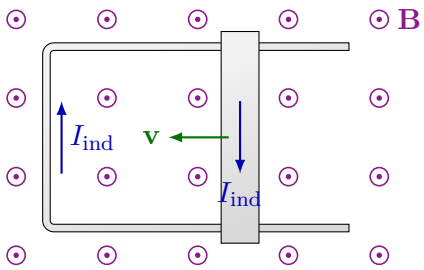


Fig. 4.3: Moving circuit.

Slide the bar to the left with speed v. Each charge q in the bar experiences a Lorentz force qvB, pushing it in the y-direction. This results in an emf which, now, is defined as the integrated force per charge. In this case, the resulting emf is

$$\mathcal{E} = vBd, \tag{4.7}$$

where d is the length of the moving bar. But, because the area inside the circuit is getting smaller, the flux through $\mathcal{C}$ is also decreasing. In this case, it's simple to compute the change of flux: it is

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -vBd. \tag{4.8}$$

We see that once again the change of flux is related to the emf through the flux rule

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{4.9}$$

Note that this is the same formula (4.6) that we derived previously, but the physics behind it looks somewhat different. In particular, we used the Lorentz force law and didn't need the Maxwell equations.

As in our previous example, the emf will drive a current around the loop $\mathcal{C}$. And, just as in the previous example, this current will oppose the motion of the bar. In this case,

it is because the current involves charges moving with some speed u around the circuit. These too feel a Lorentz force law, now pushing the bar back to the right. This means that if you let the bar go, it will not continue with constant speed, even if the connection is frictionless. Instead it will slow down. This is the analog of Lenz's law in the present case. We'll return to this example in Section 4.1.3 and compute the bar's subsequent motion.

4.1.1.1 The General Case

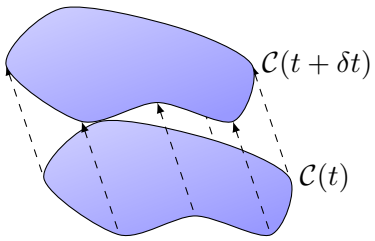


Fig. 4.4: Moving Circuits.

There is a nice way to include both the effects of time-dependent magnetic fields and the possibility that the circuit $\mathcal{C}$ changes with time. We consider the moving loop $\mathcal{C}(t)$, as shown in Fig. 4.4. Now the change in flux through a surface $\mathcal{S}$ has two terms: one because B may be changing, and one because $\mathcal{C}$ is changing. In a small time δt, we have

$$\begin{aligned}\delta\Phi = \Phi(t+\delta t) - \Phi(t) &= \int_{\mathcal{S}(t+\delta t)} \mathbf{B}(t+\delta t)\cdot \mathrm{d}\mathbf{S} - \int_{\mathcal{S}(t)} \mathbf{B}(t)\cdot \mathrm{d}\mathbf{S} \\ &= \int_{\mathcal{S}(t)} \frac{\partial \mathbf{B}}{\partial t}\delta t \cdot \mathrm{d}\mathbf{S} + \left[\int_{\mathcal{S}(t+\delta t)} - \int_{\mathcal{S}(t)}\right] \mathbf{B}(t)\cdot \mathrm{d}\mathbf{S} + \mathcal{O}\left(\delta t^2\right). \quad (4.10)\end{aligned}$$

We can do something with the middle terms. Consider the closed surface created by $\mathcal{S}(t)$ and $\mathcal{S}(t+\delta t)$, together with the cylindrical region swept out by $\mathcal{C}(t)$ which we call $\mathcal{S}_c$. Because $\boldsymbol{\nabla}\cdot\mathbf{B} = 0$, the integral of $\mathbf{B}(t)$ over any closed surface vanishes. But $\int_{\mathcal{S}(t+\delta t)} - \int_{\mathcal{S}(t)}$ is the top and bottom part of the closed surface, with the minus sign just ensuring that the integral over the bottom part $\mathcal{S}(t)$ is in the outward direction. This means that we must have

$$\left[\int_{\mathcal{S}(t+\delta t)} - \int_{\mathcal{S}(t)}\right] \mathbf{B}(t)\cdot \mathrm{d}\mathbf{S} = -\int_{\mathcal{S}_c} \mathbf{B}(t)\cdot \mathrm{d}\mathbf{S}\,. \quad (4.11)$$

For the integral over $\mathcal{S}_c$, we can write the surface element as

$$\mathrm{d}\mathbf{S} = (\mathrm{d}\mathbf{r}\times\mathbf{v})\delta t, \quad (4.12)$$

where $\mathrm{d}\mathbf{r}$ is the line element along $\mathcal{C}(t)$ and $\mathbf{v}$ is the velocity of a point on $\mathcal{C}$. We find that the expression for the change in flux can be written as

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = \lim_{\delta t\to 0}\frac{\delta\Phi}{\delta t} = \int_{\mathcal{S}(t)} \frac{\partial \mathbf{B}}{\partial t}\cdot \mathrm{d}\mathbf{S} - \int_{\mathcal{C}(t)} (\mathbf{v}\times\mathbf{B})\cdot \mathrm{d}\mathbf{r}\,, \quad (4.13)$$

where we've taken the liberty of rewriting $(\mathrm{d}\mathbf{r}\times\mathbf{v})\cdot\mathbf{B} = \mathrm{d}\mathbf{r}\cdot(\mathbf{v}\times\mathbf{B})$. Now we use the Maxwell equation (4.1) to rewrite the $\partial\mathbf{B}/\partial t$ in terms of the electric field. This gives us our final expression

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -\int_{\mathcal{C}} (\mathbf{E} + \mathbf{v}\times\mathbf{B})\cdot \mathrm{d}\mathbf{r}\,, \quad (4.14)$$

where the right-hand side now includes the force tangential to the wire from both electric fields and also from the motion of the wire in the presence of magnetic fields. The electromotive force should be defined to include both of these contributions,

$$\mathcal{E} = \int_{\mathcal{C}} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \mathrm{d}\mathbf{r}\,, \tag{4.15}$$

and we once again get the flux rule $\mathcal{E} - \mathrm{d}\Phi/\mathrm{d}t$.

4.1.2 Inductance and Magnetostatic Energy

In Section 2.3, we computed the energy stored in the electric field by considering the work done in building up a collection of charges. But we didn't repeat this calculation for the magnetic field in Chapter 3. The reason is that we need the concept of emf to describe the work done in building up a collection of currents.

Suppose that a constant current I flows along some curve $\mathcal{C}$. From the results of Chapter 3 we know that this gives rise to a magnetic field and hence a flux $\Phi = \int_{\mathcal{S}} \mathbf{B} \cdot \mathrm{d}\mathbf{S}$ through the surface $\mathcal{S}$ bounded by $\mathcal{C}$. Now increase the current I. This will increase the flux Φ. But we've just learned that the increase in flux will, in turn, induce an emf around the curve $\mathcal{C}$. The minus sign of Lenz's law ensures that this acts to resist the change of current. The work needed to build up a current is what's needed to overcome this emf.

4.1.2.1 Inductance

If a current I flowing around a curve $\mathcal{C}$ gives rise to a flux $\Phi = \int_{\mathcal{S}} \mathbf{B} \cdot \mathrm{d}\mathbf{S}$ then the *inductance* L of the circuit is defined to be

$$L = \frac{\Phi}{I} \tag{4.16}$$

The inductance is a property only of our choice of curve $\mathcal{C}$.

4.1.2.2 An Example: The Solenoid

A solenoid consists of a cylinder of length l and cross-sectional area A. We take $l \gg \sqrt{A}$ so that any end-effects can be neglected. A wire wrapped around the cylinder carries current I and winds N times per unit length. We previously computed the magnetic field through the centre of the solenoid to be (3.17)

$$B = \mu_0 I N. \tag{4.17}$$

This means that a flux through a single turn is $\Phi_0 = \mu_0 I N A$. The solenoid consists of Nl turns of wire, so the total flux is

$$\Phi = \mu_0 I N^2 A l = \mu_0 I N^2 V, \tag{4.18}$$

with $V = Al$ the volume inside the solenoid. The inductance of the solenoid is therefore

$$L = \mu_0 N^2 V. \tag{4.19}$$

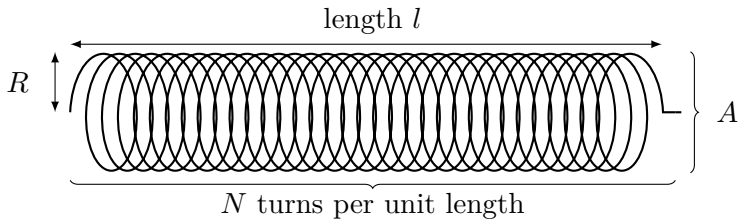


Fig. 4.5: A solenoid consisting of a cylinder of length l and cross-sectional area A, with $l \gg \sqrt{A}$ so that any end-effects can be neglected. A wire wrapped around the cylinder carries current I and winds N times per unit length.

4.1.2.3 Magnetostatic Energy

The definition of inductance is useful to derive the energy stored in the magnetic field. Let's take our circuit $\mathcal{C}$ with current I. We'll try to increase the current. The induced emf is

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -L\frac{\mathrm{d}I}{\mathrm{d}t}. \tag{4.20}$$

As we mentioned above, the induced emf can be thought of as the work done in moving a unit charge around the circuit. But we have current I flowing which means that, in time δt, a charge $I\delta t$ moves around the circuit and the amount of work done is

$$\delta W = \mathcal{E} I \delta t = -LI\frac{\mathrm{d}I}{\mathrm{d}t}\delta t \quad \Longrightarrow \quad \frac{\mathrm{d}W}{\mathrm{d}t} = -LI\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{L}{2}\frac{\mathrm{d}I^2}{\mathrm{d}t}. \tag{4.21}$$

The work needed to build up the current is just the opposite of this. Integrating over time, we learn that the total work necessary to build up a current I along a curve with inductance L is

$$W = \frac{1}{2}LI^2 = \frac{1}{2}I\Phi. \tag{4.22}$$

Following our discussion for electric energy in (2.4), we identify this with the energy U stored in the system. We can write it as

$$U = \frac{1}{2}I\int_{\mathcal{S}} \mathbf{B}\cdot\mathrm{d}\mathbf{S} = \frac{1}{2}I\int_{\mathcal{S}} \boldsymbol{\nabla}\times\mathbf{A}\cdot\mathrm{d}\mathbf{S} = \frac{1}{2}I\oint_{\mathcal{C}} \mathbf{A}\cdot\mathrm{d}\mathbf{r} = \frac{1}{2}\int \mathrm{d}^3x\,\mathbf{J}\cdot\mathbf{A}, \tag{4.23}$$

where, in the last step, we've used the fact that the current density $\mathbf{J}$ is localised on the curve $\mathcal{C}$ to turn the integral into one over all of space. At this point we turn to the Maxwell equation $\boldsymbol{\nabla}\times\mathbf{B} = \mu_0\mathbf{J}$ to write the energy as

$$U = \frac{1}{2\mu_0}\int \mathrm{d}^3x\,(\boldsymbol{\nabla}\times\mathbf{B})\cdot\mathbf{A} = \frac{1}{2\mu_0}\int \mathrm{d}^3x\,[\boldsymbol{\nabla}\cdot(\mathbf{B}\times\mathbf{A}) + \mathbf{B}\cdot(\boldsymbol{\nabla}\times\mathbf{A})]. \tag{4.24}$$

We assume that $\mathbf{B}$ and $\mathbf{A}$ fall off fast enough at infinity so that the first term vanishes. We're left with the simple expression

$$U = \frac{1}{2\mu_0}\int \mathrm{d}^3x\,\mathbf{B}\cdot\mathbf{B}. \tag{4.25}$$

Combining this with our previous result (2.69) for the electric field, we have the energy stored in the electric and magnetic fields,

$$U = \int \mathrm{d}^3x\left(\frac{\epsilon_0}{2}\mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\mathbf{B}\cdot\mathbf{B}\right). \tag{4.26}$$

This is a nice result. But there's something a little unsatisfactory behind our derivation of (4.26). First, we reiterate a complaint from Section 2.3: we had to approach the energy in both the electric and magnetic fields in a rather indirect manner, by focussing not on the fields but on the work done to assemble the necessary charges and currents. There's nothing wrong with this, but it's not a very elegant approach and it would be nice to understand the energy directly from the fields themselves. One can do better by using the Lagrangian approach to Maxwell's equations which we turn to in Section 5.1.

Second, we computed the energy for the electric fields and magnetic fields alone and then simply added them. We can't be sure, at this point, that there isn't some mixed contribution to the energy such as $\mathbf{E}\cdot\mathbf{B}$. It turns out that there are no such terms. Again, we'll postpone a proof of this until Section 5.1.

4.1.3 Resistance

You may have noticed that our discussion above has been a little qualitative. If the flux changes, we have given expressions for the induced emf $\mathcal{E}$ but we have not given an explicit expression for the resulting current. And there's a good reason for this: it's complicated.

The presence of an emf means that there is a force on the charges in the wire. And we know from Newtonian mechanics that a force will cause the charges to accelerate. This is where things start to get complicated. Accelerating charges will emit waves of electromagnetic radiation, a process that you will explore later. Relatedly, there will be an opposition to the formation of the current through the process that we've called Lenz's law.

So things are tricky. What's more, in real wires and materials there is yet another complication: friction. Throughout these lectures we have modelled our charges as if they are moving unimpeded, whether through the vacuum of space or through a conductor. But that's not the case when electrons move in real materials. Instead, there's stuff that gets in their way: various messy impurities in the material, or sound waves (usually called phonons in this context) which knock them off-course, or even other electrons. All these effects contribute to a friction force that acts on the moving electrons. The upshot of this is that the electrons do not accelerate forever. In fact, they do not accelerate for very long at all. Instead, they very quickly reach an equilibrium speed, analogous to the "terminal velocity" that particles reach when falling in gravitational field while experiencing air resistance. In many circumstances, the resulting current I is proportional to the applied emf. This relationship is called *Ohm's law*. It is

$$\mathcal{E} = IR. \tag{4.27}$$

The constant of proportionality R is called the resistance. The emf is $\mathcal{E} = \int \mathbf{E}\cdot d\mathbf{x}$. If we write $\mathbf{E} = -\boldsymbol{\nabla}\phi$, then $\mathcal{E} = V$, the potential difference between two ends of the wire. This gives us the version of Ohm's law that is familiar from school: $V = IR$.

The resistance R depends on the size and shape of the wire. If the wire has length L and cross-sectional area A, we define the *resistivity* as $\rho = AR/L$. (It's the same Greek letter that we earlier used to denote charge density. They're not the same thing. Sorry

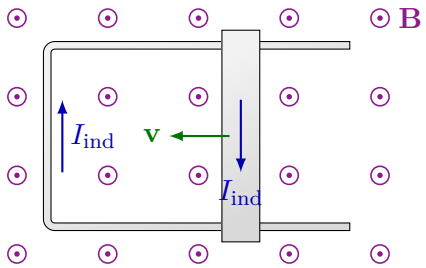


Fig. 4.6: Moving circuit.

for any confusion!) The resistivity has the advantage that it's a property of the material only, not its dimensions. Alternatively, we talk about the conductivity $\sigma = 1/\rho$. (This is the same Greek letter that we previously used to denote surface charge density. They're not the same thing either.) The general form of Ohm's law is then

$$\mathbf{J} = \sigma \mathbf{E}. \tag{4.28}$$

Unlike the Maxwell equations, Ohm's law does not represent a fundamental law of Nature. It is true in many, perhaps most, materials. But not all. There is a very simple classical model, known as the *Drude model*, which treats electrons as billiard balls experiencing linear drag which gives rise to Ohm's law.. But a proper derivation of Ohm's law needs quantum mechanics and a more microscopic understanding of what's happening in materials. Needless to say, this is (way) beyond the scope of this course. So, at least in this small section, we will take Ohm's law (4.27) as an extra input in our theory.

When Ohm's law holds, the physics is very different. Now the applied force (or, in this case, the emf) is proportional to the velocity of the particles rather than the acceleration. It's like living in the world that Aristotle envisaged rather than the one Galileo understood. But it also means that the resulting calculations typically become much simpler.

4.1.3.1 An Example

Let's return to our previous example of a sliding bar of length d and mass m which forms a circuit, sitting in a magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$. But now we will take into account the effect of electrical resistance. We take the resistance of the sliding bar to be R. But we'll make life easy for ourselves and assume that the resistance of the rest of the circuit is negligible.

There are two dynamical degrees of freedom in our problem: the position x of the sliding bar and the current I that flows around the circuit. We take $I > 0$ if the current flows along the bar in the positive $\hat{\mathbf{y}}$ direction. The Lorentz force law tells us that the force on a small volume of the bar is $\mathbf{F} = IB\hat{\mathbf{y}} \times \hat{\mathbf{z}}$. The force on the whole bar is therefore

$$\mathbf{F} = IBd\hat{\mathbf{x}} \tag{4.29}$$

The equation of motion for the position of the wire is then

$$m\ddot{x} = IBd. \tag{4.30}$$

Now we need an equation that governs the current $I(t)$. If the total emf around the circuit comes from the induced emf, we have

$$\mathcal{E} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -Bd\dot{x}. \tag{4.31}$$

Ohm's law tells us that $\mathcal{E} = IR$. Combining these, we get a simple differential equation for the position of the bar

$$m\ddot{x} = -\frac{B^2d^2}{R}\dot{x}, \tag{4.32}$$

which we can solve to see that any initial velocity of the bar, v, decays exponentially:

$$\dot{x}(t) = -ve^{-B^2d^2t/mR}. \tag{4.33}$$

Note that, in this calculation we neglected the magnetic field created by the current. It's simple to see the qualitative effect of this. If the bar moves to the left, so $\dot{x} < 0$, then the flux through the circuit decreases. The induced current is $I > 0$ which increases **B** inside the circuit which, in accord with Lenz's law, attempts to counteract the reduced flux.

In the above derivation, we assumed that the total emf around the circuit was provided by the induced emf. This is tantamount to saying that no current flows when the bar is stationary. But we can also relax this assumption and include in our analysis an emf $\mathcal{E}_0$ across the circuit (provided, for example, by a battery) which induces a current $I_0 = \mathcal{E}_0 d/R$. Now the total emf is

$$\mathcal{E} = \mathcal{E}_0 + \mathcal{E}_{\text{induced}} = \mathcal{E}_0 - Bd\dot{x}. \tag{4.34}$$

The total current is again given by Ohms law $I = \mathcal{E}/R$. The position of the bar is now governed by the equation

$$m\ddot{x} = -\frac{Bd}{R}(\mathcal{E}_0 - Bd\dot{x}). \tag{4.35}$$

Again, it's simple to solve this equation.

4.1.3.2 Joule Heating

In Subsection 4.1.2, we computed the work done in changing the current in a circuit $\mathcal{C}$. This ignored the effect of resistance. In fact, if we include the resistance of a wire then we need to do work just to keep a constant current. This should be unsurprising. It's the same statement that, in the presence of friction, we need to do work to keep an object moving at a constant speed.

Let's return to a fixed circuit $\mathcal{C}$. As we mentioned above, if a battery provides an emf $\mathcal{E}_0$, the resulting current is $I = \mathcal{E}_0/R$. We can now run through arguments similar to those that we saw when computing the magnetostatic energy. The work done in moving a unit charge around $\mathcal{C}$ is $\mathcal{E}_0$ which means that amount of work necessary to keep a current I moving for time δt is

$$\delta W = \mathcal{E}_0 I\delta t = I^2R\delta t. \tag{4.36}$$

We learn that the power (work per unit time) dissipated by a current passing through a circuit of resistance R is $\mathrm{d}W/\mathrm{d}t = I^2R$. This is not energy that can be usefully stored

like the magnetic and electric energy (4.26); instead it is lost to friction which is what we call *heat*. (The difference between heat and other forms of energy is explained in the Thermodynamics section in the Statistical Physics notes). The production of heat by a current is called *Joule heating* or, sometimes, *Ohmic heating*.

4.2 One Last Thing: The Displacement Current

We've now worked our way through most of the Maxwell equations. We've looked at Gauss' law (which is really equivalent to Coulomb's law)

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \tag{4.37}$$

and the law that says there are no magnetic monopoles

$$\nabla \cdot \mathbf{B} = 0, \tag{4.38}$$

and Ampére's law

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}, \tag{4.39}$$

and now also Faraday's law

$$\nabla \times + \frac{\partial \mathbf{B}}{\partial t} = 0. \tag{4.40}$$

In fact, there's only one term left to discuss. When fields change with time, there is an extra term that appears in Amp'ere's law, which reads in full:

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right). \tag{4.41}$$

This extra term is called the *displacement current*. It's not a great name because it's not a current. Nonetheless, as you can see, it sits in the equation in the same place as the current which is where the name comes from.

So what does this extra term do? Well, something quite remarkable. But before we get to this, there's a story to tell you.

The first four equations above (4.37), (4.38), (4.38) and (4.40) - which include Ampére's law in unmodified form – were arrived at through many decades of painstaking experimental work to try to understand the phenomena of electricity and magnetism. Of course, it took theoretical physicists and mathematicians to express these laws in the elegant language of vector calculus. But all the hard work to uncover the laws came from experiment.

The displacement current term is different. This was arrived at by pure thought alone. This is one of Maxwell's contributions to the subject and, in part, why his name now lords over all four equations. He realised that the laws of electromagnetism captured by (4.37) to (4.40) are not internally consistent: the displacement current term *has* to be there. Moreover, once you add it, there are astonishing consequences.

4.2.1 Why Ampére's Law is Not Enough

We'll look at the consequences in the next section. But for now, let's just see why the unmodified Ampére's law (4.39) is inconsistent. We simply need to take the divergence to find

$$\mu_0 \nabla \cdot \mathbf{J} = \nabla \cdot (\nabla \times \mathbf{B}) = 0. \tag{4.42}$$

This means that any current that flows into a given volume has to also flow out. But we know that's not always the case. To give a simple example, we can imagine putting lots of charge in a small region and watching it disperse. Since the charge is leaving the central region, the current does not obey $\nabla \cdot \mathbf{J} = 0$, seemingly in violation of Ampére's law.

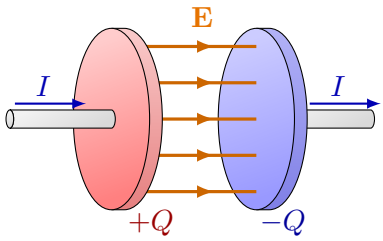


Fig. 4.7: Caption

There is a standard thought experiment involving circuits which is usually invoked to demonstrate the need to amend Ampére's law. This is shown in Fig. 4.7. The idea is to cook up a situation where currents are changing over time. To do this, we hook it up to a capacitor - which can be thought of as two conducting plates with a gap between them — to a circuit of resistance R. The circuit includes a switch. When the switch is closed, the current will flow out of the capacitor and through the circuit, ultimately heating up the resistor.

So what's the problem here? Let's try to compute the magnetic field created by the current at some point along the circuit using Amp'ere's law. We can take a curve $\mathcal{C}$ that surrounds the wire and surface $\mathcal{S}$ with boundary $\mathcal{C}$. If we chose $\mathcal{S}$ to be the obvious choice, cutting through the wire, then the calculation is the same as we saw in Section 3.1. We have

$$\int_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = \mu_0 I, \tag{4.43}$$

where I is the current through the wire which, in this case, is changing with time.

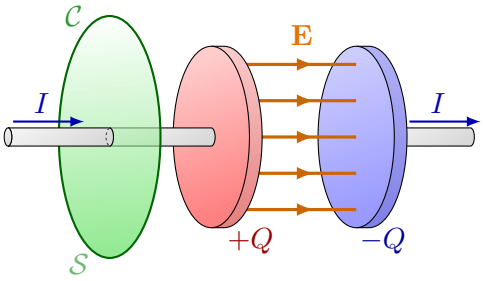


Fig. 4.8: This choice of surface suggests there is a magnetic field.

Suppose, however, that we instead decided to bound the curve $\mathcal{C}$ with the surface $\mathcal{S}'$, which now sneaks through the gap between the capacitor plates. Now there is no current

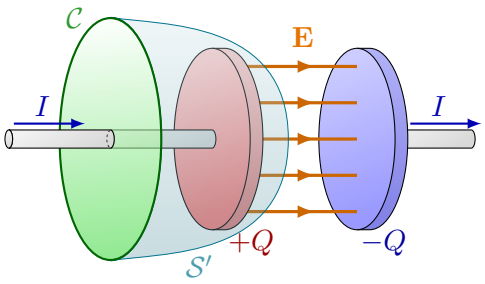


Fig. 4.9: This choice of surface suggests there is none.

passing through $\mathcal{S}'$, so if we were to use Ampére's law, we would conclude that there is no magnetic field

$$\int_{\mathcal{C}} \mathbf{B} \cdot \mathrm{d}\mathbf{r} = 0. \tag{4.44}$$

This is in contradiction to our first calculation (4.43). So what's going on here? Well, Ampére's law only holds for steady currents that are not changing with time. And we've deliberately put together a situation where I is time dependent to see the limitations of the law.

4.2.1.1 Adding the Displacement Current

Let's now see how adding the displacement current (4.41) fixes the situation. We'll first look at the abstract issue that Ampére's law requires $\boldsymbol{\nabla}\cdot\mathbf{J} = 0$. If we add the displacement current, then taking the divergence of (4.41) gives

$$\mu_0\left(\boldsymbol{\nabla}\cdot\mathbf{J} + \epsilon_0\boldsymbol{\nabla}\cdot\frac{\mathrm{d}\mathbf{E}}{\mathrm{d}t}\right) = \boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\mathbf{B}) = 0. \tag{4.45}$$

But, using Gauss's law, we can write $\epsilon_0\boldsymbol{\nabla}\cdot\mathbf{E} = \rho$, so the equation above becomes

$$\boldsymbol{\nabla}\cdot\mathbf{J} + \frac{\mathrm{d}\rho}{\mathrm{d}t} = 0, \tag{4.46}$$

which is the continuity equation that tells us that electric charge is locally conserved. It's only with the addition of the displacement current that Maxwell's equations become consistent with the conservation of charge.

Now let's return to our puzzle of the circuit and capacitor. Without the displacement current we found that $\mathbf{B} = 0$ when we chose the surface $\mathcal{S}'$ which passes between the capacitor plates. But the displacement current tells us that we missed something, because the build up of charge on the capacitor plates leads to a time-dependent electric field between the plates. For static situations, we computed this in (2.26): it is

$$E = \frac{Q}{\epsilon_0 A}, \tag{4.47}$$

where A is the area of each plate and Q is the charge that sits on each plate, and we are ignoring the edge effects which is acceptable as long as the size of the plates is much

bigger than the gap between them. Since Q is increasing over time, the electric field is also increasing

$$\frac{\partial E}{\partial t} = \frac{1}{\epsilon_0 A}\frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{1}{\epsilon_0 A} I(t). \tag{4.48}$$

So now if we repeat the calculation of $\mathbf{B}$ using the surface $\mathcal{S}'$, we find an extra term from (4.41) which gives

$$\int_{\mathcal{C}} \mathbf{B}\cdot \mathrm{d}\mathbf{r} = \int_{\mathcal{S}'} \mu_0\epsilon_0 \frac{\partial E}{\partial t} = \mu_0 I. \tag{4.49}$$

This is the same answer (4.43) that we found using Amp'ere's law applied to the surface $\mathcal{S}$.

Great. So we see why the Maxwell equations need the extra term known as the displacement current. Now the important thing is: what do we do with it? As we'll now see, the addition of the displacement current leads to one of the most wonderful discoveries in physics: the explanation for light.

4.3 And There Was Light

The emergence of light comes from looking for solutions of Maxwell's equations in which the electric and magnetic fields change with time, even in the absence of any external charges or currents. This means that we're dealing with the Maxwell equations in vacuum:

$$\begin{aligned} \boldsymbol{\nabla}\cdot\mathbf{E} = 0 \quad \text{and} \quad \boldsymbol{\nabla}\times\mathbf{B} &= \mu_0\epsilon_0\frac{\partial \mathbf{E}}{\partial t} \\ \boldsymbol{\nabla}\cdot\mathbf{B} = 0 \quad \text{and} \quad \boldsymbol{\nabla}\times\mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \end{aligned} \tag{4.50}$$

The essence of the physics lies in the two Maxwell equations on the right: if the electric field shakes, it causes the magnetic field to shake which, in turn, causes the electric field to shake, and so on. To derive the equations governing these oscillations, we start by computing the second time derivative of the electric field,

$$\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\mathbf{B}) = \boldsymbol{\nabla}\times\frac{\partial\mathbf{B}}{\partial t} = -\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{E}). \tag{4.51}$$

To complete the derivation, we need the identity

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{E}) - \nabla^2\mathbf{E}. \tag{4.52}$$

But, the first of Maxwell equations tells us that $\boldsymbol{\nabla}\cdot\mathbf{E} = 0$ in vacuum, so the first term above vanishes. We find that each component of the electric field satisfies,

$$\frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} - \nabla^2\mathbf{E} = 0. \tag{4.53}$$

This is the wave equation. The speed of the waves, c, is given by

$$c = \sqrt{\frac{1}{\mu_0\epsilon_0}} \tag{4.54}$$

Identical manipulations hold for the magnetic field. We have

$$\frac{\partial^2 \mathbf{B}}{\partial t^2} = -\frac{\partial}{\partial t}(\boldsymbol{\nabla} \times \mathbf{E}) = -\boldsymbol{\nabla} \times \frac{\partial \mathbf{E}}{\partial t} = -\frac{1}{\mu_0 \epsilon_0} \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{B}) = \frac{1}{\mu_0 \epsilon_0} \nabla^2 \mathbf{B}, \tag{4.55}$$

where, in the last equality, we have made use of the vector identity (4.51), now applied to the magnetic field $\mathbf{B}$, together with the Maxwell equation $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$. We again find that each component of the magnetic field satisfies the wave equation,

$$\frac{1}{c^2} \frac{\partial^2 \mathbf{B}}{\partial t^2} - \nabla^2 \mathbf{B} = 0. \tag{4.56}$$

The waves of the magnetic field travel at the same speed c as those of the electric field. What is this speed? At the very beginning of these lectures we provided the numerical values of the electric constant

$$\epsilon_0 = 8.854187817 \times 10^{-12} \mathrm{m}^{-3}\,\mathrm{kg}^{-1}\,\mathrm{s}^3\,\mathrm{C}^2, \tag{4.57}$$

and the magnetic constant,

$$\mu_0 = 4\pi \times 10^{-7} \mathrm{m}\,\mathrm{kg}\,\mathrm{C}^{-2}. \tag{4.58}$$

Plugging in these numbers gives the speed of electric and magnetic waves to be

$$c = 299792458 \mathrm{m}\,\mathrm{s}^{-1}. \tag{4.59}$$

But this is something that we've seen before. It's the speed of light! This, of course, is because these electromagnetic waves *are* light. The simple calculation that we have just seen represents one of the most important moments in physics. Not only are electric and magnetic phenomena unified in the Maxwell equations, but now optics – one of the oldest fields in science – is seen to be captured by these equations as well.

4.3.1 Solving the Wave Equation

We've derived two wave equations, one for $\mathbf{E}$ and one for $\mathbf{B}$. We can solve these independently, but it's important to keep in our mind that the solutions must also obey the original Maxwell equations. This will then give rise to a relationship between $\mathbf{E}$ and $\mathbf{B}$. Let's see how this works.

We'll start by looking for a special class of solutions in which waves propagate in the x-direction and do not depend on y and z. These are called *plane-waves* because, by construction, the fields $\mathbf{E}$ and $\mathbf{B}$ will be constant in the (y, z) plane for fixed x and t.

The Maxwell equation $\boldsymbol{\nabla} \cdot \mathbf{E} = 0$ tells us that we must have E_x constant in this case. Any constant electric field can always be added as a solution to the Maxwell equations so, without loss of generality, we'll choose this constant to vanish. We look for solutions of the form

$$\mathbf{E} = (0, E(x,t), 0), \tag{4.60}$$

where $\mathbf{E}$ satisfies the wave equation (4.53) which is now

$$\frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} - \nabla^2 E = 0. \tag{4.61}$$

The most general solution to the wave equation takes the form

$$E(x,t) = f(x-ct) + g(x+ct). \tag{4.62}$$

Here $f(x-ct)$ describes a wave profile which moves to the right with speed c. (Because, as t increases, x also has to increase to keep f constant). Meanwhile, $g(x+ct)$ describes a wave profile moving to the left with the speed c.

The most important class of solutions of this kind are those which oscillate with a single frequency ω. Such waves are called *monochromatic*. For now, we'll focus on the right-moving waves and take the profile to be the sine function. (We'll look at the option to take cosine waves or other shifts of phase in a moment when we discuss polarisation). We have

$$E = E_0 \sin\left[\omega\left(\frac{x}{c} - t\right)\right]. \tag{4.63}$$

We usually write this as

$$E = E_0 \sin(kx - \omega t), \tag{4.64}$$

where k is the *wavenumber*. The wave equation (4.53) requires that it is related to the frequency by

$$\omega^2 = c^2 k^2. \tag{4.65}$$

Equations of this kind, expressing frequency in terms of wavenumber, are called *dispersion relations*. Because waves are so important in physics, there's a whole bunch of associated quantities which we can define. They are:

- The quantity ω is more properly called the *angular frequency* and is taken to be positive. The actual frequency $f = \omega/2\pi$ measures how often a wave peak passes you by. But because we will only talk about ω, we will be lazy and just refer to this as frequency.
- The *period* of oscillation is $T = 2\pi/\omega$.
- The *wavelength* of the wave is $\lambda = 2\pi/k$. This is the property of waves that you first learn about in kindergarten. The wavelength of visible light is between $\lambda \sim 3.9\times10^{-7}$m and 7×10^{-7}m. At one end of the spectrum, gamma rays have wavelength $\lambda \sim 10^{-12}$m and X-rays around $\lambda \sim 10^{-10}$m to 10^{-8}m. At the other end, radio waves have $\lambda \sim$ 1cm to 10km. Of course, the electromagnetic spectrum doesn't stop at these two ends. Solutions exist for all λ.

  Although we grow up thinking about wavelength, moving forward the wavenumber k will turn out to be a more useful description of the wave.
- E_0 is the amplitude of the wave.

So far we have only solved for the electric field. To determine the magnetic field, we use $\boldsymbol{\nabla}\cdot\mathbf{B} = 0$ to tell us that B_x is constant and we again set $B_x = 0$. We know that the other components B_y and B_z must obey the wave equation (4.56). But their behaviour is dictated by what the electric field is doing through the Maxwell equation $\boldsymbol{\nabla}\times\mathbf{E} = -\partial\mathbf{B}/\partial t$. This tells us that

$$\mathbf{B} = (0,0,B), \tag{4.66}$$

with

$$\frac{\partial B}{\partial t} = -\frac{\partial E}{\partial x} = -kE_0 \cos(kx - \omega t). \tag{4.67}$$

We find

$$B = \frac{E_0}{c} \sin(kx - \omega t). \tag{4.68}$$

We see that the electric **E** and magnetic **B** fields oscillate in phase, but in orthogonal directions. And both oscillate in directions which are orthogonal to the direction in which the wave travels.

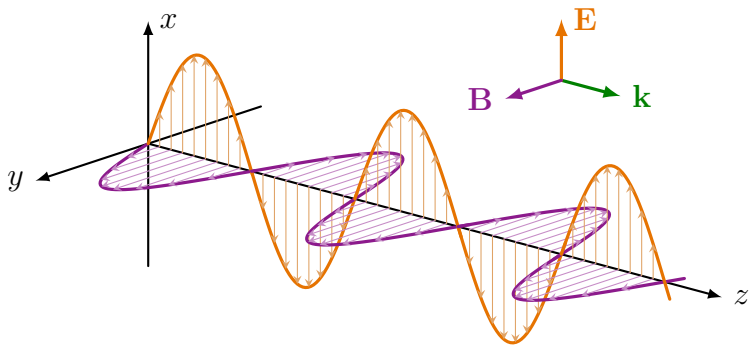


Fig. 4.10: Caption

Because the Maxwell equations are linear, we're allowed to add any number of solutions of the form (4.64) and (4.68) and we will still have a solution. This sometimes goes by the name of the *principle of superposition*. (We mentioned it earlier when discussing electrostatics). This is a particularly important property in the context of light, because it's what allow light rays travelling in different directions to pass through each other. In other words, it's why we can see anything at all.

The linearity of the Maxwell equations also encourages us to introduce some new notation which, at first sight, looks rather strange. We will often write the solutions (4.64) and (4.68) in complex notation,

$$\mathbf{E} = E_0 \hat{\mathbf{y}} e^{i(kx-\omega t)}, \quad \mathbf{B} = \frac{E_0}{c} \hat{\mathbf{z}} e^{i(kx-\omega t)}. \tag{4.69}$$

This is strange because the physical electric and magnetic fields should certainly be real objects. You should think of them as simply the real parts of the expressions above. But the linearity of the Maxwell equations means both real and imaginary parts of **E** and **B** solve the Maxwell equations. And, more importantly, if we start adding complex **E** and **B** solutions, then the resulting real and imaginary pieces will also solve the Maxwell equations. The advantage of this notation is simply that it's typically easier to manipulate complex numbers than lots of cos and sin formulae.

However, you should be aware that this notation comes with some danger: whenever you compute something which isn't linear in **E** and **B** – for example, the energy stored in the fields, which is a quadratic quantity – you can't use the complex notation above; you need to take the real part first.

4.3.2 Polarisation

Above we have presented a particular solution to the wave equation. Let's now look at the most general solution with a fixed frequency ω. This means that we look for solutions within the ansatz,

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \quad \text{and} \quad \mathbf{B} = \mathbf{B}_0 e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}, \tag{4.70}$$

where, for now, both $\mathbf{E}_0$ and $\mathbf{B}_0$ could be complex-valued vectors to include phase. (Again, we only get the physical electric and magnetic fields by taking the real part of these equations). The vector $\mathbf{k}$ is called the *wavevector*. Its magnitude, $|\mathbf{k}| = k$, is the wavenumber and the direction of $\mathbf{k}$ points in the direction of propagation of the wave. The expressions (4.70) already satisfy the wave equations (4.53) and (4.56) if ω and $\mathbf{k}$ obey the dispersion relation $\omega^2 = c^2k^2$.

The wavevector can also be complex $\mathbf{k} \to \mathbf{k} + i\boldsymbol{\kappa}$ giving damped waves. If

$$\left\{\begin{matrix}\mathbf{k} \parallel \boldsymbol{\kappa} \\ \mathbf{k} \nparallel \boldsymbol{\kappa}\end{matrix}\right\} \text{ the wave is } \left\{\begin{matrix}\text{homogeneous} \\ \text{inhomogeneous}\end{matrix}\right\}. \tag{4.71}$$

If there is free charge present, $\boldsymbol{\nabla}\cdot\mathbf{D} \neq 0 \implies \mathbf{k}\cdot\mathbf{D} \neq 0$, so $\mathbf{D}$ and $\mathbf{E}$ can have components parallel to $\mathbf{k}$. For the special case of $\epsilon = 0$ and in the absence of free charge, Eq. (4.72) is satisfied for $\mathbf{k} \not\perp \mathbf{E}$, and indeed for $\mathbf{k} \parallel \mathbf{E}$. Plasma waves (see Subsection 4.4.3) are longitudinal.

In isotropic materials, if there are no free charges or currents, we get further constraints on $\mathbf{E}_0$, $\mathbf{B}_0$ and $\mathbf{k}$ from the original Maxwell equations. These are

$$\boldsymbol{\nabla}\cdot\mathbf{D} = 0 \implies \mathbf{k}\cdot\mathbf{E}_0 = 0, \tag{4.72}$$

$$\boldsymbol{\nabla}\cdot\mathbf{B} = 0 \implies \mathbf{k}\cdot\mathbf{B}_0 = 0, \tag{4.73}$$

$$\boldsymbol{\nabla}\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \implies \mathbf{k}\times\mathbf{E}_0 = \omega\mathbf{B}_0. \tag{4.74}$$

Let's now interpret these equations:

4.3.2.1 Linear Polarisation

Suppose that we take $\mathbf{E}_0$ and $\mathbf{B}_0$ to be real. The first two equations above say that both $\mathbf{E}_0$ and $\mathbf{B}_0$ are orthogonal to the direction of propagation. The last of the equations (4.74) above says that $\mathbf{E}_0$ and $\mathbf{B}_0$ are also orthogonal to each other. You can check that the fourth Maxwell equation doesn't lead to any further constraints. Using the dispersion relation $\omega = ck$, the last constraint above can be written as

$$\hat{\mathbf{k}}\times(\mathbf{E}_0/c) = \mathbf{B}_0. \tag{4.75}$$

This means that the three vectors $\hat{\mathbf{k}}$, $\mathbf{E}_0/c$ and $\mathbf{B}_0$ form a right-handed orthogonal triad. Waves of this form are said to be *linearly polarised*. The electric and magnetic fields oscillate in fixed directions, both of which are transverse to the direction of propagation.

From Eqs. (4.72-4.74) we find for real fields and $\mathbf{k}$

$$\boxed{\mathbf{B} \perp \mathbf{k} \text{ and } \mathbf{E}, \quad \mathbf{D} \perp \mathbf{k} \text{ and } \mathbf{H},} \tag{4.76}$$

while $\mathbf{E}$ and $\mathbf{H}$ are *not necessarily* orthogonal to $\mathbf{k}$.

For isotropic materials, ϵ and μ are simple scalars (however this is not the general case, see Section 6.3) so that $\mathbf{B} \parallel \mathbf{H}$ and $\mathbf{D} \parallel \mathbf{E}$.

4.3.2.2 Circular and Elliptic Polarisation

Suppose that we now take $\mathbf{E}_0$ and $\mathbf{B}_0$ to be complex. The actual electric and magnetic fields are just the real parts of (4.70), but now the polarisation does not point in a fixed direction. To see this, write

$$\mathbf{E}_0 = \boldsymbol{\alpha} - i\boldsymbol{\beta}. \tag{4.77}$$

The real part of the electric field is then

$$\mathbf{E} = \boldsymbol{\alpha}\cos(\mathbf{k}\cdot\mathbf{x} - \omega t) + \boldsymbol{\beta}\sin(\mathbf{k}\cdot\mathbf{x} - \omega t), \tag{4.78}$$

with Maxwell equations ensuring that $\boldsymbol{\alpha}\cdot\mathbf{k} = \boldsymbol{\beta}\cdot\mathbf{k} = 0$. If we look at the direction of $\mathbf{E}$ at some fixed point in space, say the origin $\mathbf{x} = 0$, we see that it doesn't point in a fixed direction. Instead, it rotates over time within the plane spanned by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (which is the plane perpendicular to $\mathbf{k}$).

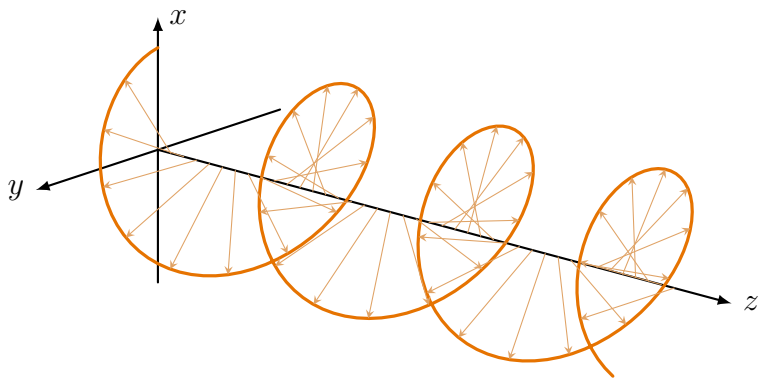


Fig. 4.11: Caption

A special case arises when the phase of $\mathbf{E}_0$ is $e^{i\pi/4}$, so that $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}|$, with the further restriction that $\boldsymbol{\alpha}\cdot\boldsymbol{\beta} = 0$. Then the direction of $\mathbf{E}$ traces out a circle over time in the plane perpendicular to $\mathbf{k}$. This is called *circular polarisation*. The polarisation is said to be *right-handed* if $\boldsymbol{\beta} = \hat{\mathbf{k}} \times \boldsymbol{\alpha}$ and *left-handed* if $\boldsymbol{\beta} = -\hat{\mathbf{k}} \times \boldsymbol{\alpha}$.

In general, the direction of $\mathbf{E}$ at some point in space will trace out an ellipse in the plane perpendicular to the direction of propagation $\mathbf{k}$. Unsurprisingly, such light is said to have *elliptic polarisation*.

4.3.2.3 General Wave

A general solution to the wave equation consists of combinations of waves of different wavenumbers and polarisations. It is naturally expressed as a Fourier decomposition by

summing over solutions with different wavevectors,

$$\mathbf{E}(\mathbf{x},t) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\mathbf{E}(\mathbf{k})e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}. \tag{4.79}$$

Here, the frequency of each wave depends on the wavevector by the now-familiar dispersion relation $\omega = ck$.

4.3.3 An Application: Reflection off a Conductor

There are lots of things to explore with electromagnetic waves and we will see many examples later in the course. For now, we look at a simple application: we will reflect waves off a conductor. We all know from experience that conductors, like metals, look shiny. Here we'll see why.

Suppose that the conductor occupies the half of space $x > 0$. We start by shining the light head-on onto the surface. This mean an incident plane wave, travelling in the x-direction,

$$\mathbf{E}_i = E_0\hat{\mathbf{y}}e^{i(kx-\omega t)}, \tag{4.80}$$

where, as before, $\omega = ck$. Inside the conductor, we know that we must have $\mathbf{E} = \mathbf{0}$. But the component $\mathbf{E}\cdot\hat{\mathbf{y}}$ lies tangential to the surface and so, by continuity, must also vanish just outside at $x = 0^-$. We achieve this by adding a reflected wave, travelling in the opposite direction

$$\mathbf{E}_r = -E_0\hat{\mathbf{y}}e^{i(-kx-\omega t)}. \tag{4.81}$$

So that the combination $\mathbf{E} = \mathbf{E}_i + \mathbf{E}_r$ satisfies $E(x=0) = 0$ as it must. This is illustrated in Fig. 4.12. (Note, however, that Fig. 4.12 is a little bit misleading: the two waves are shown displaced but, in reality, both fill all of space and should be superposed on top of each other).

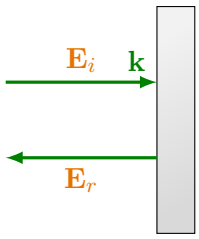


Fig. 4.12: Reflection off a conductor.

We've already seen above that the corresponding magnetic field can be determined by $\boldsymbol{\nabla}\times\mathbf{E} = -\partial\mathbf{B}/\partial t$. It is given by $\mathbf{B} = \mathbf{B}_i + \mathbf{B}_r$, with

$$\mathbf{B}_i = \frac{E_0}{c}\hat{\mathbf{z}}e^{i(kx-\omega t)} \quad \text{and} \quad \mathbf{B}_r = \frac{E_0}{c}\hat{\mathbf{z}}e^{i(-kx-\omega t)}. \tag{4.82}$$

This obeys $\mathbf{B}\cdot\hat{\mathbf{n}}$, as it should by continuity. But the tangential component doesn't vanish at the surface. Instead, we have

$$\mathbf{B}\cdot\hat{\mathbf{z}}|_{x=0} = \frac{2E_0}{c}e^{-i\omega t}. \tag{4.83}$$

Since the magnetic field vanishes inside the conductor, we have a discontinuity. But there's no mystery here. We know from our previous discussion (3.14) that this corresponds to a surface current $\mathbf{K}$ induced by the wave

$$\mathbf{K} = \frac{2E_0}{c\mu_0}\hat{\mathbf{y}}e^{-i\omega t}. \tag{4.84}$$

We see that the surface current oscillates with the frequency of the reflected wave.

4.3.3.1 Reflection at an Angle

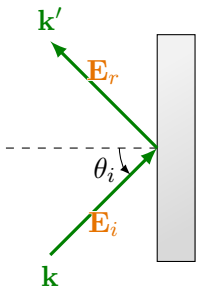


Fig. 4.13: Reflection off a conductor at an angle.

Let's now try something a little more complicated: we'll send in the original ray at an angle, θ_i, to the normal as shown in Fig. 4.13. Our incident electric field is

$$\mathbf{E}_i = E_0\hat{\mathbf{y}}e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}, \tag{4.85}$$

where

$$\mathbf{k} = k\cos\theta_i\hat{\mathbf{x}} + k\sin\theta_i\hat{\mathbf{z}}. \tag{4.86}$$

Notice that we've made a specific choice for the polarisation of the electric field: it is out of the page in Fig. 4.13, tangential to the surface. Now we have two continuity conditions to worry about. We want to add a reflected wave,

$$\mathbf{E}_r = -E_0\hat{\boldsymbol{\zeta}}e^{i(\mathbf{k}'\cdot\mathbf{x}-\omega' t)}, \tag{4.87}$$

where we've allowed for the possibility that the polarisation $\hat{\boldsymbol{\zeta}}$, the wavevector $\mathbf{k}'$ and frequency ω' are all different from the incident wave. We require two continuity conditions on the electric field

$$(\mathbf{E}_i + \mathbf{E}_r)\cdot\hat{\mathbf{n}} = 0 \quad \text{and} \quad (\mathbf{E}_i + \mathbf{E}_r)\times\hat{\mathbf{n}} = 0, \tag{4.88}$$

where, for this set-up, the normal vector is $\hat{\mathbf{n}} = -\hat{\mathbf{x}}$. This is achieved by taking $\omega' = \omega$ and $\hat{\boldsymbol{\zeta}} = \hat{\mathbf{y}}$, so that the reflected wave changes neither frequency nor polarisation. The reflected wavevector is

$$\mathbf{k}' = -k\cos\theta_i\hat{\mathbf{x}} + k\sin\theta_i\hat{\mathbf{z}}. \tag{4.89}$$

We can also check what becomes of the magnetic field. It is $\mathbf{B} = \mathbf{B}_i + \mathbf{B}_r$, with

$$\mathbf{B}_i = \frac{E_0}{c}(\hat{\mathbf{k}}\times\hat{\mathbf{y}})e^{i(\mathbf{k}\cdot\mathbf{x}-\omega' t)} \quad \text{and} \quad \mathbf{B}_r = -\frac{E_0}{c}(\hat{\mathbf{k}}'\times\hat{\mathbf{y}})e^{i(\mathbf{k}'\cdot\mathbf{x}-\omega' t)} \tag{4.90}$$

Note that, in contrast to (4.82), there is now a minus sign in the reflected $\mathbf{B}_r$, but this is simply to absorb a second minus sign coming from the appearance of $\hat{\mathbf{k}}'$ in the polarisation

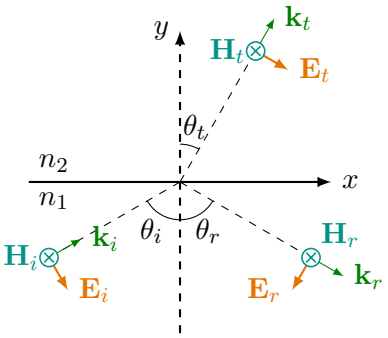


Fig. 4.14: A plane wave incident on a dielectric boundary.

vector. It is simple to check that the normal component $\mathbf{B} \cdot \hat{\mathbf{n}}$ vanishes at the interface, as it must. Meanwhile, the tangential component again gives rise to a surface current.

The main upshot of all of this discussion is relationship between $\mathbf{k}$ and $\mathbf{k}'$ which tells us something that we knew when we were five: the angle of incidence is equal to the angle of reflection. Only now we've derived this from the Maxwell equations. If this is a little underwhelming, we'll derive many more properties of waves later.

4.4 Reflection and Transmission at Interfaces

4.4.1 Reflection and Transmission at Interfaces

Generally, it is quite complicated to calculate the form of the travelling electromagnetic waves that exist when dielectric or magnetic bodies are present, and sophisticated analysis techniques have to be used. An important problem that can be analysed, however, is when a plane wave is incident on the boundary between two dielectric materials that have different refractive indices.

Consider the situation shown in Fig. 4.14. A plane wave having wave vector $\mathbf{k}_i$ is incident on a plane dielectric boundary. A transmitted wave, $\mathbf{k}_t$, and a reflected wave, $\mathbf{k}_r$, are produced, and we cannot make any assumptions about their directions of travel. The angles of incidence, reflection and transmission have been marked as θ_i, θ_r, and θ_t respectively.

In this example, we have chosen the incident wave to be polarised in the *plane of incidence* (the plane containing both the incident wavevector and the normal to the interface). Without loss of generality, choose coordinate system such that $\hat{\mathbf{z}}$ is perpendicular to the plane of incidence, so there is no z dependence for any of the fields, and so we are justified in drawing the diagram as shown

The incident, reflected, and transmitted plane waves have the functional forms

$$\begin{aligned}\mathbf{E}_i &= \mathbf{E}_{i0}\exp\left[i(\mathbf{k}_i\cdot\mathbf{x}-\omega_i t)\right],\\ \mathbf{E}_r &= \mathbf{E}_{r0}\exp\left[i(\mathbf{k}_r\cdot\mathbf{x}-\omega_r t)\right],\\ \mathbf{E}_t &= \mathbf{E}_{t0}\exp\left[i(\mathbf{k}_t\cdot\mathbf{x}-\omega_t t)\right],\end{aligned} \tag{4.91}$$

where we have not even assumed that the transmitted and reflected waves have the same temporal frequencies as the incident field.

In the following, we want to draw up relationships between ω, $\mathbf{k}$ and θ of the incident, reflected, and transmitted waves. For each wave (using $k = |\mathbf{k}|$)

$$k_x = k\sin\theta \tag{4.92}$$

and, remembering that on the x-axis, where $y = 0$, the component of the electric field *parallel* to the surface $\left|\mathbf{E}_{\parallel}\right| = E_x$ must be continuous across the boundary:

$$E_{i,x} + E_{r,x} = E_{t,x} \tag{4.93}$$

i.e.,

$$\begin{aligned}E_{i0}\exp\left[i(k_i x\sin\theta_i-\omega_i t)\right]\cos\theta_i - E_{r0}\exp\left[i(k_r x\sin\theta_r-\omega_r t)\right]\cos\theta_r\\ = E_{t0}\exp\left[i(k_t x\sin\theta_t-\omega_t t)\right]\cos\theta_t.\end{aligned} \tag{4.94}$$

Eq. (4.94) must be true for all x and t, and therefore as before we must have

$$\omega_i = \omega_r = \omega_t, \tag{4.95}$$

and

$$k_i x\sin\theta_i = k_r x\sin\theta_r = k_t x\sin\theta_t. \tag{4.96}$$

The above procedure is known as *phase matching*. In other words, the fields must have the same temporal, and x-directed spatial, frequencies.

We also know that

$$k = \frac{n\omega}{c}, \tag{4.97}$$

where n is the refractive index in that region, and therefore

$$\begin{aligned}k_1 &\equiv k_i = k_r,\\ k_2 &\equiv k_t,\end{aligned} \tag{4.98}$$

with

$$\frac{k_1}{k_2} = \frac{n_1}{n_2}. \tag{4.99}$$

We conclude that, because of Eqs. (4.98) and (4.96),

$$\theta_i = \theta_r, \tag{4.100}$$

and we find *Snell's law*, $n\sin\theta = \text{const.}$:

$$\boxed{\frac{\sin\theta_t}{\sin\theta_i} = \frac{k_1}{k_2} = \frac{n_1}{n_2},} \tag{4.101}$$

The same analysis can be applied to a linearly polarised wave with the electric field perpendicular to the plane of incidence—there are just no $\cos\theta$ factors in Eq. (4.94). The same phase-matching requirements apply, and a similar set of expressions is derived. In particular, Snell's law still holds, as would be expected.

4.4.2 Reflection and Transmission Coefficients

We have established the relationship between the angles of incidence and reflection, but we have said nothing about the *amount* of power that is reflected. This will be discussed in this Subsection.

Because the arguments of the exponentials in Eq. (4.94) are always equal,

$$\begin{aligned} E_{i0}\cos\theta_i - E_{r0}\cos\theta_r &= E_{t0}\cos\theta_t \\ \implies (E_{i0} - E_{r0})\cos\theta_i &= E_{t0}\cos\theta_t \end{aligned} \tag{4.102}$$

where the last line follows because the angles of incidence and reflection are equal. If we match the parallel component of the **H** field in an identical way, we find

$$\begin{aligned} H_{i0} + H_{r0} &= H_{t0} \\ \implies n_1 E_{i0} + n_1 E_{r0} &= n_2 E_{t0}, \end{aligned} \tag{4.103}$$

where the second line follows because we always have $E/H = Z_0/n$. It follows that

$$(E_{i0} - E_{r0})\cos\theta_i = \frac{n_1}{n_2}(E_{i0} + E_{r0})\cos\theta_t, \tag{4.104}$$

which can be rearranged to give

$$\frac{E_{r0}}{E_{i0}} = \frac{(n_2/n_1)\cos\theta_i - \cos\theta_t}{(n_2/n_1)\cos\theta_i + \cos\theta_t} = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}, \tag{4.105}$$

where Snell's law has been used for the last step.

We conclude that the reflection coefficient for a parallel-polarised plane wave is

$$r_\parallel = \frac{E_{r0}}{E_{i0}} = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}. \tag{4.106}$$

In a similar way, it can be shown that the transmission coefficient is given by

$$t_\parallel = \frac{E_{t0}}{E_{i0}} = \frac{2\cos\theta_i}{(n_2/n_1)\cos\theta_i + \cos\theta_t}. \tag{4.107}$$

An entirely equivalent procedure can be carried out when the plane of polarisation is perpendicular to the plane of incidence, giving a reflection coefficient of

$$r_\perp = \frac{E_{r0}}{E_{i0}} = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}, \tag{4.108}$$

and a transmission coefficient of

$$t_\perp = \frac{E_{t0}}{E_{i0}} = \frac{2\cos\theta_i}{\cos\theta_i + (n_2/n_1)\cos\theta_t}. \tag{4.109}$$

These four equations are known as *Fresnel's relations*.

To illustrate the usefullness of the coefficients above, consider an example where $n_1 = 1$ and $n_2 = n$. Then at normal incidence, $\theta_i = \theta_r = \theta_t = 0$, so, using small-angle formulae

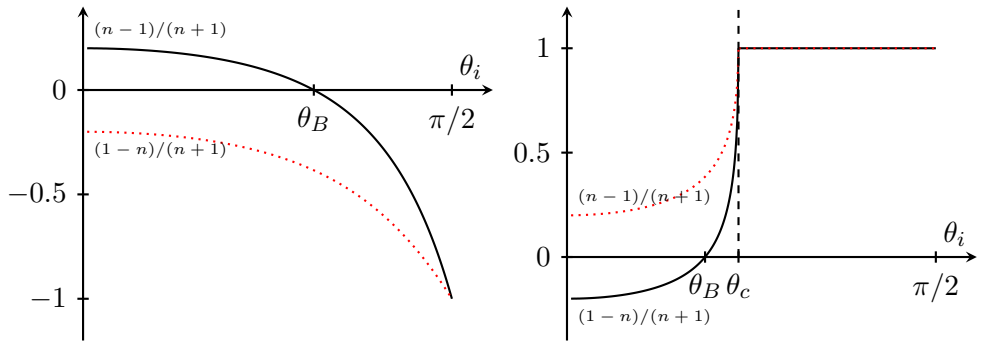


Fig. 4.15: Reflection coefficients of a dielectric boundary. Left: From air to glass ($n = 1.5$). Right: From glass to air, showing total internal reflection for $\theta_i > \theta_c$.

as $\theta_i \to 0$ in the first part of Eq. (4.105) or in Eq. (4.106), the power reflection coefficient becomes

$$R_{\parallel} = \left|r_{\parallel}\right|^2 = \left(\frac{n-1}{n+1}\right)^2 = |r_{\perp}|^2 = R_{\perp}. \tag{4.110}$$

In the case of glass, $n = 1.5$, and so $R = 4\%$.

As the angle of incidence and the polarisation are changed, the behaviour becomes quite complicated. The Fresnel equations have been used to produce the plots shown in Fig. 4.15. Notice that the reflection coefficients can take on negative values, showing that the fields can be 180° out of phase, i.e., $E_{r0} = -E_{i0}$.

An analysis of Fig. 4.15 shows that there is a particular angle of incidence where $r_{\parallel} = 0$. This angle is called the *Brewster angle*, and is given by

$$\tan\theta_{\mathrm{B}} = \frac{n_2}{n_1}, \tag{4.111}$$

where the wave is incident in the material with index n_1 and transmitted into the material with index n_2. If going from free space, the Brewster angle is simply

$$\tan\theta_{\mathrm{B}} = n_2. \tag{4.112}$$

The Brewster angle depends on whether the wave is travelling from air into glass or glass into air. Brewster-angle windows are often used to eliminate reflections from optical windows or to produce linearly polarised light from unpolarised light. Hence, they are often found in experimental optical setups.

Brewster's angle gives a simple way to create polarised light: shine unpolarised light on a dielectric at angle θ_B and the only thing that bounces back has normal polarisation. This is the way sunglasses work to block out polarised light from the Sun. It is also the way polarising filters work.

For incidence from within the material, $n_1 > n_2$, so there is some critical angle θ_c such that when $\theta_i > \theta_c$ a problem occurs:

$$\sin\theta_t = \frac{n_2}{n_1}\sin\theta_i > 1. \tag{4.113}$$

There is no (real) solution to this equation, and hence all of the incident power is reflected, regardless of the polarisation of the wave! The *critical angle* of *total internal reflection* is given by

$$\sin\theta_c = \frac{n_2}{n_1}. \tag{4.114}$$

Once the critical angle of incidence is exceeded, an *evanescent* wave is produced that decays exponentially from the surface and travels along the surface. Critical angles and polarisation properties are now being used extensively in the field of surface-plasmon physics.

4.4.3 Waves in Plasmas

Up to this point we have considered the case where the medium in which the wave is travelling does not have conduction currents, but there are numerous examples of where conduction currents are present. We will start with plasmas, and then look at metals.

A plasma is a region of space where free electrons and their parent ions are present. The mass of an electron is much less than that of an ion, and so they are more mobile than ions. In this analysis, we shall ignore the movement of ions.

We know that the equation of motion for an electron in an electromagnetic field is

$$m_e\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = -e(\mathbf{E} + \mathbf{v}\times\mathbf{B}), \tag{4.115}$$

where $\mathbf{r}$ is the electron's position, and $\mathbf{v}$ its velocity.

Suppose that we illuminate a free electron with a plane wave, then

$$\frac{E_x}{H_y} = \sqrt{\frac{\mu_0}{\epsilon_0}}, \tag{4.116}$$

i.e.,

$$\frac{E_x}{B_y} = c. \tag{4.117}$$

The cross-product term in Eq. (4.115) will have its greatest magnitude when $\mathbf{v}$ and $\mathbf{H}$ are at right angles, but when $|\mathbf{v}| \ll c$, it can be ignored, compared to the $\mathbf{E}$-field term. Let us suppose that the temporal variation of the wave is represented as a vector in the complex plane:

$$\mathbf{E} = \mathbf{E}_0\exp\left[i(kz-\omega t)\right]. \tag{4.118}$$

It then follows that

$$\mathbf{r} = \frac{e}{m_e\omega^2}\mathbf{E}_0\exp\left[i(kz-\omega t)\right], \tag{4.119}$$

which can be verified by substitution into Eq. (4.115).

Equation (4.119) indicates that the position of the electron oscillates with an amplitude that is inversely proportional to the mass and the squared frequency of the incident radiation. The inertial mass of the electron "regulates" the amplitude of the oscillation as the frequency increases.

In a plasma, we have to take into account the ions, in the sense that as the electrons move, they separate from the ions, inducing dipoles. The dipole moment of a single separating pair is

$$\mathbf{P} = -e\mathbf{r} = -\frac{e^2}{m_e\omega^2}\mathbf{E}, \tag{4.120}$$

and the dipole moment per unit volume is

$$\mathbf{P} = N\mathbf{P} = -\frac{Ne^2}{m_e\omega^2}\mathbf{E}, \tag{4.121}$$

where N is the number of electrons per unit volume, and it is assumed that they are non-interacting.

Recall from Section 2.5 that χ is defined through

$$\mathbf{P} = \epsilon_0\chi\mathbf{E}, \tag{4.122}$$

and therefore

$$\epsilon = 1 + \chi = 1 - \frac{Ne^2}{\epsilon_0 m_e\omega^2}, \tag{4.123}$$

the *relative permittivity in plasma* is usually written as

$$\epsilon = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2}, \tag{4.124}$$

where the *plasma frequency* is defined as

$$\omega_\mathrm{p}^2 = \frac{Ne^2}{\epsilon_0 m_e}. \tag{4.125}$$

The plasma frequency ω_p characterises the electromagnetic properties of a plasma. The plasma frequency can be expressed in the convenient form:

$$f_\mathrm{p} = \frac{\omega_\mathrm{p}}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{Ne^2}{\epsilon_0 m_e}} \approx 9\sqrt{N/\mathrm{m}^{-3}}\mathrm{Hz}. \tag{4.126}$$

Taking the ionosphere as a concrete example, $N \approx 10^{12}\mathrm{m}^{-3}$, and therefore $f_\mathrm{p} \approx 10\mathrm{MHz}$. Thus, at frequencies below 10MHz, electromagnetic waves are reflected off the ionosphere, enabling low-frequency communications.

The relative permittivity of a plasma ϵ is shown as a function of oscillation frequency in Fig. 4.16. The refractive index becomes a function of the oscillation frequency, too:

$$n = \sqrt{\epsilon} = \left(1 - \frac{\omega_\mathrm{p}^2}{\omega^2}\right)^{1/2}. \tag{4.127}$$

The refractive index is imaginary for frequencies below the plasma frequency, and we expect an evanescent wave.

Let us study the case where $\omega < \omega_\mathrm{p}$ in more detail. Since the refractive index is imaginary, it is convenient to define

$$n = i\beta, \tag{4.128}$$

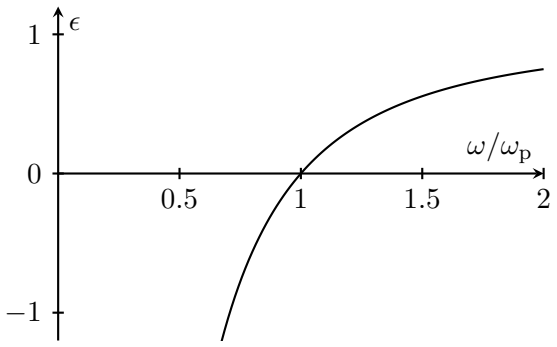


Fig. 4.16: The effective dielectric constant of a plasma.

where

$$\beta = \sqrt{|\epsilon|}. \tag{4.129}$$

Then

$$k = \frac{n\omega}{c} = \frac{i\beta\omega}{c}, \tag{4.130}$$

which can be substituted into Eq. (4.118) to give

$$\mathbf{E} = \mathbf{E}_0 \exp\left[-\frac{\omega\beta}{c}z\right] \exp\left[-i\omega t\right]. \tag{4.131}$$

The resulting $\mathbf{E}$ field no longer takes the form of a travelling wave, but decays with z, as shown in Fig. 4.17.

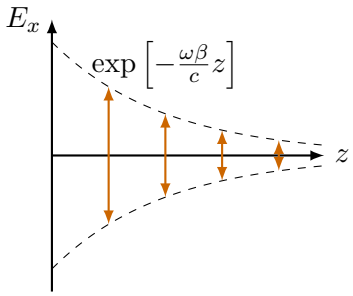


Fig. 4.17: The time varying electric field of an electromagnetic wave decays in a plasma below cut-off.

The expression in Eq. (4.131) can also be written explicitly in terms of ω_p using Eq. (4.127) for $\omega < \omega_p$,

$$\mathbf{E} = \mathbf{E}_0 \exp\left[-k_0 z\left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2}\right] \exp\left[-i\omega t\right], \tag{4.132}$$

where $k_0 = \omega/c$. The lower the frequency, the shorter the distance over which the field decays. Notice that the wave is *not* a decaying propagating wave—it does not propagate at all. A non-propagating, decaying wave of this kind is called an *evanescent wave*.

It is also useful to calculate the magnetic field, which is given by

$$H_y = \sqrt{\frac{\epsilon\epsilon_0}{\mu_0}}E_x = \frac{i\beta}{Z_0}E_x. \tag{4.133}$$

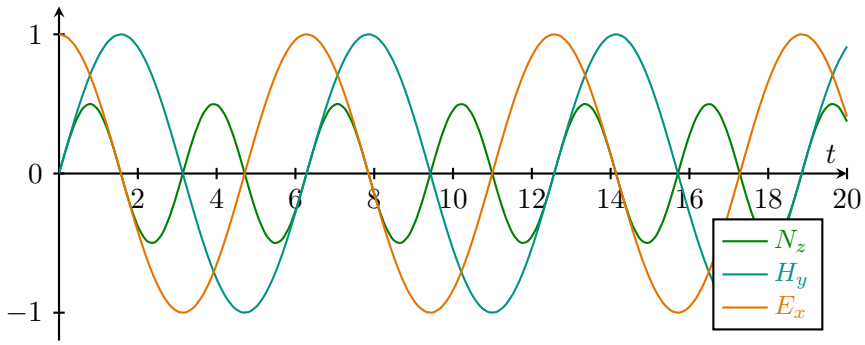


Fig. 4.18: The time variation of the electric and magnetic fields in a plasma below cut-off. The sign of the Poynting vector alternates, indicating the sloshing of energy.

The electric and magnetic fields are $\pi/2$ out of phase, as shown in Fig. 4.18. In other words, the surface impedance of the plasma is reactive.

The average and instantaneous magnitudes of the Poynting vector (see Section 4.5) become respectively

$$\left|\frac{1}{2}\,\mathrm{Re}\left[\mathbf{E}(\mathbf{r})\times\mathbf{H}(\mathbf{r})^*\right]\right| = 0 \quad \text{and} \quad \left|\frac{1}{2}\,\mathrm{Im}\left[\mathbf{E}(\mathbf{r})\times\mathbf{H}(\mathbf{r})^*\right]\right| = \frac{\beta}{2Z_0}E_x^2. \tag{4.134}$$

Thus, there is no net transfer of energy within the plasma, but energy sloshes backwards and forwards in the z-direction at any point. This is effectively caused by the energy being stored in the motion of the oscillating electrons. All of the energy in the original wave that is incident on the plasma must, on average, be reflected.

Finally, we can look at propagation velocities. When the frequency of the incident radiation is above the plasma frequency, the phase velocity becomes

$$v = \frac{\omega}{k} = \frac{c}{\sqrt{1-\omega_\mathrm{p}^2/\omega^2}}. \tag{4.135}$$

Given that v explicitly depends on ω, the waves are *dispersive*, i.e., different frequencies travel with different velocities. As a result, the amplitude envelope of a wave packet will in general change as the packet propagates. Notice that the phase velocity is greater than the speed of light.

It is straightforward to show that the group velocity is given by

$$v_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}k} = c\sqrt{1-\omega_\mathrm{p}^2/\omega^2}, \tag{4.136}$$

which is less than c, as expected. It is interesting to observe that the product of the phase and group velocities is c^2. This general behaviour is also seen in many other structures, such as waveguides.

4.4.4 Waves in Conducting Media

In a plasma, we have assumed the density of electrons (and their parent atoms) is low enough that electrons do not collide with each other during one or more cycles of the

incoming EM radiation. In a metal, electrons are separated from their parent atoms as in a plasma, but their density is much higher, and so scattering becomes significant. We will now investigate what happens if we attempt to pass an electromagnetic wave through a sheet of metal.

Remember that for all materials we have the constitutive relations

$$\mathbf{D} = \epsilon\epsilon_0\mathbf{E}, \tag{4.137}$$

$$\mathbf{B} = \mu\mu_0\mathbf{H}, \tag{4.138}$$

$$\mathbf{J} = \sigma\mathbf{E}. \tag{4.139}$$

Using the last of Maxwell's equations, we find

$$\begin{aligned}\boldsymbol{\nabla} \times \mathbf{H} &= \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \\ &= \sigma\mathbf{E} + \epsilon\epsilon_0\frac{\partial \mathbf{E}}{\partial t}.\end{aligned} \tag{4.140}$$

If we now assume solutions of the form

$$\mathbf{E} = \mathbf{E}_0 \exp[-i\omega t] \tag{4.141}$$

$$\mathbf{H} = \mathbf{H}_0 \exp[-i\omega t], \tag{4.142}$$

we find

$$\begin{aligned}\boldsymbol{\nabla} \times \mathbf{H} &= (\sigma - i\omega\epsilon\epsilon_0)\mathbf{E} \\ &= -i\omega\epsilon_0\left(\epsilon + \frac{i\sigma}{\omega\epsilon_0}\right)\mathbf{E}.\end{aligned} \tag{4.143}$$

An insulating material with dielectric constant ϵ' would give

$$\boldsymbol{\nabla} \times \mathbf{H} = -i\omega\epsilon'\epsilon_0\mathbf{E}, \tag{4.144}$$

and, therefore, for a conducting material, we can define an *effective dielectric constant*:

$$\epsilon' \equiv \left(\epsilon + \frac{i\sigma}{\omega\epsilon_0}\right). \tag{4.145}$$

The effects of the current are contained within the effective dielectric constant ϵ'. In fact, the effective dielectric constant is a complex quantity. It can now be seen that a material can be considered to behave as a dielectric if the real part of the complex dielectric constant dominates, whereas it behaves as a metal if the imaginary part of the complex dielectric constant dominates. For example, in the case of copper, $\sigma = 5 \times 10^7 \Omega^{-1}\,\mathrm{m}^{-1}$, operating up to optical frequencies ($\nu = 10^{15}$Hz), we find that the real part of the complex dielectric constant is negligible (about 5000 times smaller than the imaginary part). For metals, we therefore have

$$\epsilon' \approx \frac{i\sigma}{\omega\epsilon_0}. \tag{4.146}$$

Usually, when the dielectric constant appears in the solution of a field problem it enters by way of its square root, but now the effective dielectric constant is a complex quantity. The refractive index is therefore also a complex quantity. For a metal,

$$\begin{aligned}n &= \sqrt{\epsilon'\mu} \\ &= \sqrt{\frac{i\sigma\mu}{\omega\epsilon_0}} \\ &= \pm\frac{(1+i)}{\sqrt{2}}\sqrt{\frac{\sigma\mu}{\omega\epsilon_0}}.\end{aligned} \tag{4.147}$$

We have already shown that Maxwell's equations can be solved to give plane-wave solutions of the form

$$\mathbf{E} = \mathbf{E}_0 \exp[i(kz - \omega t)], \tag{4.148}$$
$$\mathbf{H} = \mathbf{H}_0 \exp[i(kz - \omega t)], \tag{4.149}$$

where

$$k = \frac{\omega}{c/n}. \tag{4.150}$$

In the case of a metal,

$$k = \frac{\omega}{c}(1+i)\sqrt{\frac{\sigma\mu}{2\omega\epsilon_0}} = (1+i)\sqrt{\frac{\sigma\omega\mu_0\mu}{2}} = \frac{1+i}{\delta}, \tag{4.151}$$

where the *skin depth* of the material is defined as

$$\delta \equiv \sqrt{\frac{2}{\sigma\omega\mu_0\mu}}. \tag{4.152}$$

Substituting the wave number k into Eq. (4.148) gives

$$\mathbf{E} = \mathbf{E}_0 \exp\left[-\frac{z}{\delta}\right] \exp\left[i\left(\frac{z}{\delta} - \omega t\right)\right]. \tag{4.153}$$

This new expression has a number of features: firstly, it represents a travelling wave with 'wave number' $1/\delta$, and secondly, the amplitude of the wave decays on propagation. Notice that if we were to use the other solution for $\sqrt{i}$ we would have a solution whose amplitude increases indefinitely with propagation, which is not physical. In fact, the amplitude of the field has decayed by e^{-1} after a distance δ. The attenuation in z is very severe—the wave decays by a factor of $e^{-2\pi} \approx 1/535$ in every wavelength ($\lambda/\delta = 2\pi$). This is shown in Fig. 4.19.

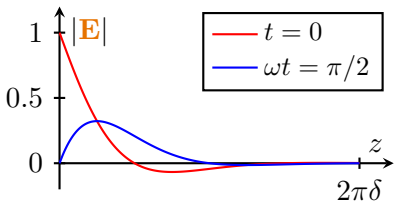


Fig. 4.19: The propagation of a field into a good conductor. The red curve corresponds to time $t = 0$, and the blue curve to $\omega t = \pi/2$.

The skin depth in metals is usually very small, and so electromagnetic waves decay very rapidly once they enter a highly conductive material. For example, in the case of copper, $\sigma = 5 \times 10^7 \Omega^{-1}\,\mathrm{m}^{-1}$, at 100MHz the skin depth is nearly 7µm, which is a tiny fraction of the free-space wavelength, 3m. Even though the attenuation in a metal is very strong, if the metal is thin enough, some transmission is possible. In other words, electromagnetic waves can penetrate into metals, but they are attenuated severely.

It is also worth looking at the relationship between the **E** and the **H** fields. If the wave is polarised in the x-direction, then

$$H_y = \sqrt{\frac{\epsilon_0\epsilon}{\mu_0\mu}}E_x = \sqrt{\frac{i\sigma}{\omega\mu\mu_0}}E_x = \sqrt{\frac{\sigma}{2\omega\mu\mu_0}}(1+i)E_x. \tag{4.154}$$

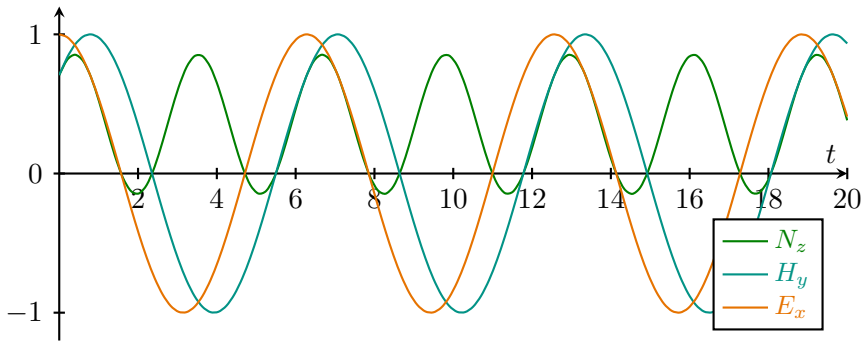


Fig. 4.20: The fields in a good conductor as a function of time

A new feature has emerged: the **E** and **H** fields are 45° out of phase. Fig. 4.20 shows the fields inside the surface of a good conductor as a function of time. Notice that H_y lags behind E_x.

Figure 4.20 also shows the instantaneous value of the Poynting vector (E_x, H_y), which points in the z-direction. The Poynting vector displays a number of features: firstly, it takes on positive and negative values, indicating that power can be travelling backwards at particular times in the cycle; secondly, on average it travels forwards, showing that power must be dissipated in the conductor. Power dissipation is caused by ohmic loss within the material.

4.4.5 The Skin Effect

The partial propagation of a field into a good conductor has particular significance for the flow of current in a wire, which we will address in this Subsection.

Consider a wire carrying a current I that oscillates at frequency ω. Power flows along the wire, and so the largest component of the Poynting vector points in the same direction as the wire. We also know, however, that because of the current flow on the surface, and the finite conductivity of the metal, there is a non-zero component of E on the surface. When the Poynting vector is calculated using this **E** field and the **H** field associated with the current flow, the result points perpendicularly into the surface of the wire, showing that energy flows *into* the surface.

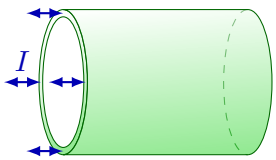


Fig. 4.21: High-frequency currents flow on the surface of a wire.

Approximating the surface of the wire as a plane surface as shown in Fig. 4.21, with $z \approx a - r$, where $a \gg \delta$ is the radius of the wire, δ is the skin depth, and the x-axis is along the wire (parallel to the current), we know that

$$E_x = E_0 \exp\left[-\frac{z}{\delta}\right] \exp\left[i\left(\frac{z}{\delta} - \omega t\right)\right], \tag{4.155}$$

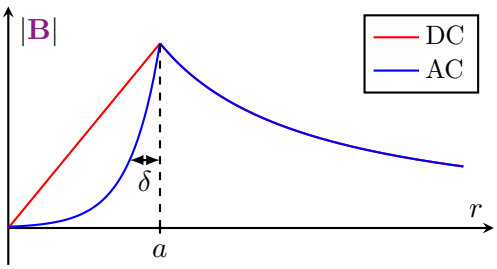


Fig. 4.22: The magnetic field inside and outside a current-carrying wire.

and therefore, using Ohm's law $\mathbf{J} = \sigma\mathbf{E}$,

$$J_x = J_0 \exp\left[-\frac{z}{\delta}\right] \exp\left[i\left(\frac{z}{\delta} - \omega t\right)\right], \tag{4.156}$$

In other words, the amplitude of the current decays away from the surface of the wire. Consequently, the resulting magnetic field B decays rapidly too, as shown in Fig. 4.22. The oscillating currents are confined to the surface of the wire, and the skin depth gets smaller as the frequency increases. The electromagnetic behaviour of a (cylindrical) wire can be analysed rigorously using cylindrical basis functions, and essentially the same result is found.

With the above toolset it is now possible to calculate the resistance of a wire at high frequencies; as the frequency increases, the resistance increases due to the current being confined to a smaller and smaller region. When the skin depth is much smaller than a, we can approximate the circular case by 'unwrapping' the shell in which the current flows. The total current in the wire is given by

$$I = \int \mathrm{d}S\, J_x \approx 2\pi a \int \mathrm{d}z\, J_x(z). \tag{4.157}$$

Taking $z = 0$ on the surface, and letting the other limit tend to ∞ for simplicity since $\delta \ll a$, we get

$$I \approx 2\pi a J_0 \exp\left[-i\omega t\right] \int_0^\infty \mathrm{d}z \exp\left[\frac{z}{\delta}(-1+i)\right], \tag{4.158}$$

which evaluates to

$$I \approx 2\pi a J_0 \exp\left[-i\omega t\right] \frac{\delta}{1-i} \frac{1+i}{1+i}, \tag{4.159}$$

or

$$I \approx \pi a J_0 \delta (1+i) \exp\left[-i\omega t\right]. \tag{4.160}$$

The actual physical current can be derived from its complex representation through

$$\begin{aligned} I(t) = \mathrm{Re}[I] &= \mathrm{Re}\left[\pi a J_0 \delta (1+i) \exp\left[-i\omega t\right]\right] \\ &= \pi a J_0 \delta (\cos\omega t + \sin\omega t), \end{aligned} \tag{4.161}$$

which has a mean square value of

$$\left\langle I(t)^2 \right\rangle = (\pi a J_0 \delta)^2. \tag{4.162}$$

The actual physical current density can also be calculated from its complex representation to give

$$J(t) = \mathrm{Re}[J_x] = J_0 \exp\left[-\frac{z}{\delta}\right] \cos\left(\frac{z}{\delta} - \omega t\right), \tag{4.163}$$

which has a mean square value of

$$\left\langle J(t)^2 \right\rangle = \frac{1}{2} J_0^2 \exp\left[-\frac{2z}{\delta}\right]. \tag{4.164}$$

The power dissipated per unit length of the wire in an elemental annulus of the cross-section, of infinitesimally small area $\mathrm{d}A = 2\pi a\,\mathrm{d}z$, is

$$\mathrm{d}P = \frac{I^2 R}{L} = J^2\,\mathrm{d}A^2 \left(\frac{L}{\sigma\,\mathrm{d}A}\right)\frac{1}{L} = \frac{J^2\,\mathrm{d}A}{\sigma}. \tag{4.165}$$

The power dissipated per unit length of the wire in the whole crosssection of the wire is therefore

$$P = \frac{J_0^2}{2\sigma} 2\pi a \int_0^\infty \mathrm{d}z \exp\left[-\frac{2z}{\delta}\right] = \frac{J_0^2 \pi a \delta}{2\sigma}. \tag{4.166}$$

Finally, we can define an effective resistance by requiring that the actual dissipated power is given when the total current and the effective resistance are used to calculate the power. The effective resistance, R, per unit length is therefore

$$R = \frac{P}{\langle I(t)^2 \rangle} = \frac{1}{2\pi a \delta \sigma}. \tag{4.167}$$

In other words, *from the point of view of power dissipation, the resistance per unit length is simply the resistance that is calculated when the current is assumed to flow uniformly in a thin shell of thickness* δ. Thus ordinary wires become very lossy at short wavelengths, and other ways of guiding electromagnetic energy must be found.

4.4.5.1 Metals vs Plasmas

The propagation of electromagnetic waves in metals is not dissimilar to the propagation of electromagnetic waves in plasmas. However, there are some differences:

- In metals, the electrons tend to be *scattered* on time-scales less than one oscillation cycle, leading to power dissipation.
- In plasmas, all of the electrons move together, and the movement is undamped. Below the plasma frequency, power is reflected from the plasma, rather than being absorbed.

At high frequencies, when the electrons move only a short distance in each cycle, collective motion occurs, and the plasma-frequency concept becomes appropriate, even for a metal.

It can be seen that, in general terms, the surface impedance of a metal will have a real part, corresponding to power absorption, and an imaginary part, corresponding to the energy stored in the motion of the electrons.

Superconductors have a particularly high kinetic inductance term due to undamped motion of Cooper pairs. The behaviour of superconducting transmission lines is modified significantly by this kinetic-inductance effect.

4.5 Transport of Energy: The Poynting Vector

Electromagnetic waves carry energy. This is an important fact: we get most of our energy from the light of the Sun. Here we'd like to understand how to calculate this energy.

Our starting point is the expression (4.26) for the energy stored in electric and magnetic fields,

$$U = \int_{\mathcal{V}} \mathrm{d}^3x \left(\frac{\epsilon_0}{2} \mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\mathbf{B}\cdot\mathbf{B} \right). \tag{4.168}$$

The expression in brackets is the energy density. Here we have integrated this only over some finite volume $\mathcal{V}$ rather than over all of space. This is because we want to understand the way in which energy can leave this volume. We do this by calculating

$$\begin{aligned} \frac{\mathrm{d}U}{\mathrm{d}t} &= \int_{\mathcal{V}} \mathrm{d}^3x \left(\epsilon_0 \mathbf{E}\cdot\frac{\partial \mathbf{E}}{\partial t} + \frac{1}{\mu_0}\mathbf{B}\cdot\frac{\partial \mathbf{B}}{\partial t} \right) \\ &= \int_{\mathcal{V}} \mathrm{d}^3x \left(\frac{1}{\mu_0}\mathbf{E}\cdot(\boldsymbol{\nabla}\times\mathbf{B}) - \mathbf{E}\cdot\mathbf{J} - \frac{1}{\mu_0}\mathbf{B}\cdot(\boldsymbol{\nabla}\times\mathbf{E}) \right), \end{aligned} \tag{4.169}$$

where we've used the two Maxwell equations. Now we use the identity

$$\mathbf{E}\cdot(\boldsymbol{\nabla}\times\mathbf{B}) - \mathbf{B}\cdot(\boldsymbol{\nabla}\times\mathbf{E}) = -\boldsymbol{\nabla}\cdot(\mathbf{E}\times\mathbf{B}), \tag{4.170}$$

and write

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -\int_{\mathcal{V}} \mathrm{d}^3x\, \mathbf{J}\cdot\mathbf{E} - \frac{1}{\mu_0}\int_{\mathcal{S}} (\mathbf{E}\times\mathbf{B})\cdot\mathrm{d}\mathbf{S}\,, \tag{4.171}$$

where we've used the divergence theorem to write the last term. This equation is sometimes called the *Poynting theorem*.

The first term on the right-hand side is related to something that we've already seen in the context of Newtonian mechanics. The work done on a particle of charge q moving with velocity $\mathbf{v}$ for time δt in an electric field is $\delta W = q\mathbf{v}\cdot\mathbf{E}\delta t$. The integral $\int_{\mathcal{V}} \mathrm{d}^3x\, \mathbf{J}\cdot\mathbf{E}$ above is simply the generalisation of this to currents: it should be thought of as the rate of gain of energy of the particles in the region $\mathcal{V}$. Since it appears with a minus sign in (4.171), it is the rate of loss of energy of the particles.

Now we can interpret (4.171). If we write it as

$$\frac{\mathrm{d}U}{\mathrm{d}t} + \int_{\mathcal{V}} \mathrm{d}^3x\, \mathbf{J}\cdot\mathbf{E} = \frac{1}{\mu_0}\int_{\mathcal{S}} (\mathbf{E}\times\mathbf{B})\cdot\mathrm{d}\mathbf{S}\,, \tag{4.172}$$

then the left-hand side is the combined change in energy of both fields and particles in region $\mathcal{V}$. Since energy is conserved, the right-hand side must describe the energy that escapes through the surface $\mathcal{S}$ of region $\mathcal{V}$. We define the *Poynting vector* as

$$\mathbf{N} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}. \tag{4.173}$$

in free space, or in the presence of a material we nicely have

$$\mathbf{N} = \mathbf{E}\times\mathbf{H}. \tag{4.174}$$

This is a vector field. It tells us the magnitude and direction of the flow of energy in any point in space. The Poynting vector is in the direction of the wave *phase propagation*.

Let's now look at the energy carried in electromagnetic waves. Because the Poynting vector is quadratic in $\mathbf{E}$ and $\mathbf{B}$, we're not allowed to use the complex form of the waves. We need to revert to the real form. For linear polarisation, we write the solutions in the form (4.69), but with arbitrary wavevector $\mathbf{k}$,

$$\mathbf{E} = \mathbf{E}_0 \sin(\mathbf{k}\cdot\mathbf{x} - \omega t) \quad \text{and} \quad \mathbf{B} = \frac{1}{c}\left(\hat{\mathbf{k}} \times \mathbf{E}_0\right) \sin(\mathbf{k}\cdot\mathbf{x} - \omega t). \tag{4.175}$$

The Poynting vector is then

$$\mathbf{N} = \frac{E_0^2}{c\mu_0}\hat{\mathbf{k}} \sin^2(\mathbf{k}\cdot\mathbf{x} - \omega t). \tag{4.176}$$

Averaging over a period, $T = 2\pi/\omega$, we have

$$\bar{\mathbf{N}} = \frac{E_0^2}{2c\mu_0}\hat{\mathbf{k}}. \tag{4.177}$$

We learn that the electromagnetic wave does indeed transport energy in its direction of propagation $\hat{\mathbf{k}}$. It's instructive to compare this to the energy density of the field (4.26). Evaluated on the electromagnetic wave, the energy density is

$$u = \frac{\epsilon_0}{2}\mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\mathbf{B}\cdot\mathbf{B} = \epsilon_0 E_0^2 \sin^2(\mathbf{k}\cdot\mathbf{x} - \omega t). \tag{4.178}$$

Averaging over a period, $T = 2\pi/\omega$, this is

$$\bar{u} = \frac{\epsilon_0 E_0^2}{2}. \tag{4.179}$$

Then, using $c^2 = 1/\epsilon_0\mu_0$, we can write

$$\bar{\mathbf{N}} = c\bar{u}\hat{\mathbf{k}}. \tag{4.180}$$

The interpretation is simply that the energy $\bar{\mathbf{N}}$ is equal to the energy density in the wave $\bar{u}$ times the speed of the wave, c. A similar result can be obtained when instead considering a wave in a medium.

4.5.1 The Continuity Equation Revisited

Recall that, way back in Chapter 1, we introduced the continuity equation for electric charge,

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{J} = 0. \tag{4.181}$$

This equation is not special to electric charge. It must hold for any quantity that is locally conserved.

Now we have encountered another quantity that is locally conserved: energy. In the context of Newtonian mechanics, we are used to thinking of energy as a single number. Now, in field theory, it is better to think of energy density $\mathcal{E}(\mathbf{x}, t)$. This includes the energy in both fields and the energy in particles. Thinking in this way, we notice that (4.171) is simply the integrated version of a continuity equation for energy. We could equally well write it as

$$\frac{\partial \mathcal{E}}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{N} = 0. \tag{4.182}$$

We see that the Poynting vector $\mathbf{N}$ is to energy what the current $\mathbf{J}$ is to charge. We'll explore this connection further in Section 5.1.

Chapter 5

Electromagnetism and Relativity

We've seen that Maxwell's equations have wave solutions which travel at the speed of light. But there's another place in physics where the speed of light plays a prominent role: the theory of special relativity. How does electromagnetism fit with special relativity?

Historically, the Maxwell equations were discovered before the theory of special relativity. It was thought that the light waves we derived above must be oscillations of some substance which fills all of space. This was dubbed the *aether*. The idea was that Maxwell's equations only hold in the frame in which the aether is at rest; light should then travel at speed c relative to the aether.

We now know that the concept of the aether is unnecessary baggage. Instead, Maxwell's equations hold in all inertial frames and are the first equations of physics which are consistent with the laws of special relativity. Ultimately, it was by studying the Maxwell equations that Lorentz was able to determine the form of the Lorentz transformations which subsequently laid the foundation for Einstein's vision of space and time.

Our goal in this section is to view electromagnetism through the lens of relativity. We will find that observers in different frames will disagree on what they call electric fields and what they call magnetic fields. They will observe different charge densities and different currents. But all will agree that these quantities are related by the same Maxwell equations. Moreover, there is a pay-off to this. It's only when we formulate the Maxwell equations in a way which is manifestly consistent with relativity that we see their true beauty. The slightly cumbersome vector calculus equations that we've been playing with throughout these lectures will be replaced by a much more elegant and simple-looking set of equations.

5.0.1 Gauge Invariance and Relativity

5.1 More on Energy and Momentum

Chapter 6

Optics

Optics is often encountered in phenomena at optical frequencies ($400-800$nm; $430-790$THz) in which our eyes are most sensitive. Much of the relevant physics applies at other frequencies: X-rays ($\lambda < 1$nm), infrared ($\lambda = 1\text{mm}-800\text{nm}$), microwaves ($\lambda = 1\text{mm}-1\text{m}$, $\nu = 0.3-300$GHz) and radio ($\lambda = 1\text{m}-100\text{km}$, $\nu = 3\text{kHz}-0.3\text{GHz}$). At optical frequencies, in most circumstances $\mu \to 1$ and hence

$$n = \sqrt{\epsilon\mu} \to \sqrt{\epsilon}. \tag{6.1}$$

We shall start by considering general polarisation states via superpositions of the plane-wave solutions (see 4.3).

Consider superposition of two perpendicularly plane-polarised waves $\mathbf{E}_1 = a_1 E_0 \hat{\mathbf{x}} e^{i(kz-\omega t)}$ and $\mathbf{E}_2 = a_2 E_0 \hat{\mathbf{y}} e^{i(kz-\omega t)}$ as in Fig. 6.1.

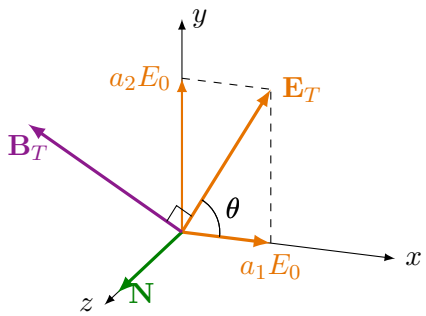


Fig. 6.1: Superposition of two perpendicularly plane-polarised waves, $\mathbf{E}_T = \mathbf{E}_1 + \mathbf{E}_2 = (a_1\hat{\mathbf{x}} + a_2\hat{\mathbf{y}})E_0 e^{i(kz-\omega t)}$.

If a_1 and a_2 are both real, then the components of $\mathbf{E}_T$ oscillate in phase. The resultant vector is at an angle of $\theta = \tan^{-1}(a_2/a_1)$ to $\hat{\mathbf{x}}$, with an amplitude $E_T = E_0\sqrt{a_1^2 + a_2^2}$.

6.1 Circular and Elliptical Polarisation

If a_1 and a_2 are *not* both real, the x and y components of $\mathbf{E}_T$ oscillate with a fixed *phase difference* $\delta \neq 0$. e.g. take $a_1 = 1$ and $a_2 = i = e^{i\pi/2}$: i.e. E_y lags behind E_x by $\delta = \pi/2$:

$$\mathbf{E}_T = E_0(\hat{\mathbf{x}}e^{i(kz-\omega t)} + \hat{\mathbf{y}}e^{i(kz-[\omega t-\pi/2])}). \tag{6.2}$$

At $z = 0$, and taking the real parts:

$$\mathbf{E}_T = E_0\hat{\mathbf{x}}\cos(\omega t) + E_0\hat{\mathbf{y}}\cos(\omega t - \pi/2) = E_0\hat{\mathbf{x}}\cos(\omega t) + E_0\hat{\mathbf{y}}\sin(\omega t). \tag{6.3}$$

At $t = 0$, and taking the real parts:

$$\mathbf{E}_T = E_0\hat{\mathbf{x}}\cos(kz) + E_0\hat{\mathbf{y}}\cos(kz + \pi/2) = E_0\hat{\mathbf{x}}\cos(kz) - E_0\hat{\mathbf{y}}\sin(kz). \tag{6.4}$$

This is defined as a Left-Hand Circularly Polarized wave. An observer towards whom the light is propagating sees $\mathbf{E}_T(z=0)$ rotate *anti-clockwise*, while the instantaneous field sweeps out a left-handed helix through space.

For $a_1 = 1$ and $a_2 = -i$, $\delta = -\pi/2$ an RCP results. An observer towards whom the light is propagating sees $\mathbf{E}_T(z=0)$ rotate *clockwise*, while the instantaneous field now sweeps out a right-handed helix through space. This is illustrated in Fig. 6.2. (Electrical engineers use the opposite convention.)

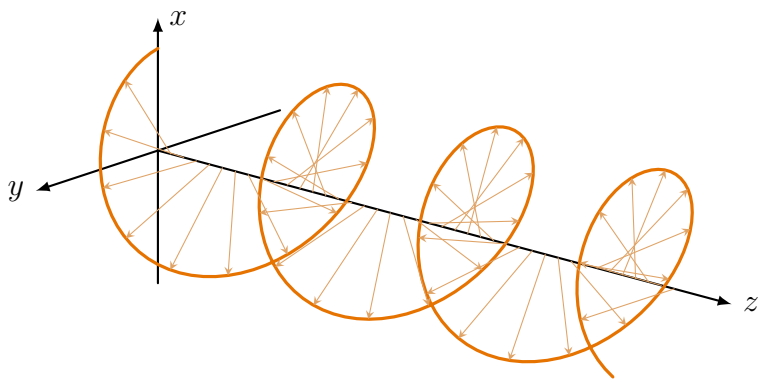


Fig. 6.2: Right-Hand Circularly Polarised Wave, $a_1 = 1$, $a_2 = -i$: $\mathbf{E}_T$ at $z = 0$ rotates clockwise with $T = 2\pi/\omega$, while the instantaneous field now sweeps out a right-handed helix through space.

For $|a_1| \neq |a_2|$ and $\delta \neq \pm\pi/2$, $\mathbf{E}$ is elliptically polarised. With $a_1 = a$, $a_2 = be^{i\delta}$ (with a, b real):

$$E_x = a\cos(\omega t) \quad \text{and} \quad E_y = b\cos(\omega t - \delta), \tag{6.5}$$

using a simple identity we can relate the components

$$\begin{aligned} \frac{E_y}{b} &= \cos(\omega t)\cos\delta + \sin(\omega t)\sin\delta \\ &= \frac{E_x}{a}\cos\delta + \sqrt{1 - \frac{E_x^2}{a^2}}\sin\delta \end{aligned} \tag{6.6}$$

giving

$$\frac{E_x^2}{a^2} + \frac{E_y^2}{b^2} - 2\cos\delta\frac{E_x}{a}\frac{E_y}{b} = \sin^2\delta \tag{6.7}$$

the equation for an ellipse with axes at an angle α to the E_x, E_y directions, shown in Fig. 6.3.

6.2 Jones Notation

The complex amplitudes a_1 and a_2 of the two x and y linearly polarised waves can be used as the basis for a useful matrix approach for handling the effects of various optical devices on the polarisation state - **Jones algebra**.

The factors of $1/\sqrt{2}$ are for normalisation [Lx, CR etc. are non-standard notation].

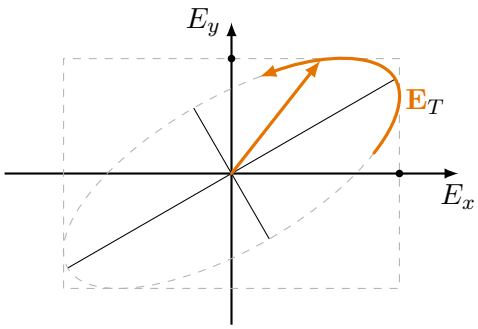


Fig. 6.3: Elliptical polarisation for $a_1 = a$, $a_2 = be^{i\delta}$: $\mathbf{E}_T$ at $z = 0$ traces an ellipse, with $\tan 2\alpha = \frac{2ab\cos\delta}{a^2-b^2}$.

Jones vector	***x*-pol**	***y*-pol**	***θ*-pol**	**RCP**	**LCP**	**General elliptical**
	$\mathbf{L}_x$	$\mathbf{L}_y$	$\mathbf{L}_\theta$	$\mathbf{C}_\mathrm{R}$	$\mathbf{C}_\mathrm{L}$	
$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -i \end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ i \end{pmatrix}$	$\begin{pmatrix} a \\ be^{i\delta} \end{pmatrix}$

Table 6.1: Jones Notation

Linear combinations, with appropriate phases, of the various polarisation states can be used to form other polarisation states. e.g.

$$\frac{1}{\sqrt{2}}(\mathbf{C}_\mathrm{R} + \mathbf{C}_\mathrm{L}) = \mathbf{L}_x. \tag{6.8}$$

The effects of various components of an optical system – polarisers, phase plates etc. can be represented by 2×2 Jones matrices, as will be seen first in Section 6.3.

6.3 Anisotropic Media

6.3.1 Dichroism

Dichroic materials absorb light linearly polarized in one direction more than light polarised in the other, cf. wire grid polariser but at molecular level. E.g. Polaroid film, a plastic containing conducting polymeric chains aligned by stretching (along $\hat{\mathbf{y}}$ say). The sheet is anisotropic, conducting along $\hat{\mathbf{y}}$ but not along $\hat{\mathbf{x}}$. Light with $\mathbf{E} \parallel \hat{\mathbf{y}}$ is absorbed, and light with $\mathbf{E} \parallel \hat{\mathbf{x}}$ is not.

So when light passes through the sheet, only the $\mathbf{E}_x$ component is transmitted. This

is included in the Jones algebra as a *Jones matrix*:

$$\underline{\underline{\mathbf{J}}}_x = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{6.9}$$

$$\underline{\underline{\mathbf{J}}}_x \mathbf{L}_x = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{L}_x \tag{6.10}$$

$$\underline{\underline{\mathbf{J}}}_x \mathbf{L}_y = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \mathbf{0}. \tag{6.11}$$

A polaroid with transmitting axis oriented at θ to $\hat{\mathbf{x}}$ is represented by the matrix

$$\underline{\underline{\mathbf{J}}}_\theta = \begin{pmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{pmatrix}. \tag{6.12}$$

For initially linearly polarised light $\mathbf{L}_x$ of intensity I_0 which then passes through polariser $\underline{\underline{\mathbf{J}}}_\theta$ the output is given by

$$\underline{\underline{\mathbf{J}}}_\theta \mathbf{L}_x = \begin{pmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \times \sqrt{I_0} = \begin{pmatrix} \cos^2\theta \\ \sin\theta\cos\theta \end{pmatrix} \times \sqrt{I_0}, \tag{6.13}$$

so the transmitted intensity is thus given by *Malus's Law,*

$$I(\theta) = I_0\Big(\cos^4\theta + \sin^2\theta\cos^2\theta\Big) = I_0\cos^2\theta. \tag{6.14}$$

For $\theta = \pi/2$, note as expected $I(\pi/2) = 0$.

Note that for light passing through a sequence of optical elements A,B,C the overall Jones matrix is

$$\underline{\underline{\mathbf{J}}} = \underline{\underline{\mathbf{J}}}_\mathrm{C} \cdot \underline{\underline{\mathbf{J}}}_\mathrm{B} \cdot \underline{\underline{\mathbf{J}}}_\mathrm{A}, \tag{6.15}$$

since $\underline{\underline{\mathbf{J}}}_\mathrm{A}$ is applied to the Jones vector first. So for *crossed polarisers*,

$$\underline{\underline{\mathbf{J}}} = \underline{\underline{\mathbf{J}}}_\theta \cdot \underline{\underline{\mathbf{J}}}_{\pi/2+\theta} = \underline{\underline{\mathbf{0}}}. \tag{6.16}$$

6.3.2 Birefringence

Optical anisotropy also occurs naturally – e.g. in materials which are **birefringent** because they are structurally anisotropic – e.g. calcite.

It has been assumed so far that the medium is *isotropic*

$$\mathbf{P} = \epsilon_0\chi\mathbf{E}, \quad \mathbf{D} = \epsilon\epsilon_0\mathbf{E}, \quad \mathbf{M} = \chi_\mathrm{m}\mathbf{H}, \quad \mathbf{B} = \mu\mu_0\mathbf{H}, \tag{6.17}$$

irrespective of the directions of $\mathbf{E}$ and $\mathbf{H}$. i.e. that χ, ϵ, and χ_m are **scalars**.

For materials with an anisotropic crystal structure this is not so, and the relative permittivity (and therefore also the refractive index) is a *tensor of the second rank*, conveniently written in matrix form

$$\mathbf{D} = \epsilon_0\underline{\underline{\boldsymbol{\epsilon}}} \cdot \mathbf{E}, \quad D_i = \epsilon_0\sum_j \epsilon_{ij}E_j, \quad \begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \epsilon_0 \begin{pmatrix} \epsilon_{xx} & \epsilon_{xy} & \epsilon_{xz} \\ \epsilon_{yx} & \epsilon_{yy} & \epsilon_{yz} \\ \epsilon_{zx} & \epsilon_{zy} & \epsilon_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}. \tag{6.18}$$

So in general $\mathbf{D} \nparallel \mathbf{E}$.

An important property of the dielectric tensor $\underline{\underline{\boldsymbol{\epsilon}}}$ can be deduced from energy flow considerations,

$$\mathbf{N} = \mathbf{E} \times \mathbf{H} \quad \text{and} \quad u = \frac{1}{2}\mathbf{E}\cdot\mathbf{D} + \frac{1}{2}\mathbf{B}\cdot\mathbf{H} \tag{6.19}$$

and the vector identity

$$\boldsymbol{\nabla}\cdot(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot(\boldsymbol{\nabla}\times\mathbf{a}) - \mathbf{a}\cdot(\boldsymbol{\nabla}\times\mathbf{b}), \tag{6.20}$$

by considering the first time derivative of u,

$$\begin{aligned}
\frac{\mathrm{d}u}{\mathrm{d}t} &= \frac{1}{2}\left(\dot{\mathbf{E}}\cdot\mathbf{D} + \mathbf{E}\cdot\dot{\mathbf{D}} + \dot{\mathbf{B}}\cdot\mathbf{H} + \mathbf{B}\cdot\dot{\mathbf{H}}\right) && (6.21)\\
&= -\boldsymbol{\nabla}\cdot\mathbf{N} - \mathbf{J}\cdot\mathbf{E} \qquad \text{(from energy conservation)}\\
&= -[\mathbf{H}\cdot(\boldsymbol{\nabla}\times\mathbf{E}) - \mathbf{E}\cdot(\boldsymbol{\nabla}\times\mathbf{H})] - \mathbf{J}\cdot\mathbf{E}\\
&= \mathbf{H}\cdot\dot{\mathbf{B}} + \mathbf{E}\cdot(\mathbf{J}+\dot{\mathbf{D}}) - \mathbf{J}\cdot\mathbf{E} \qquad \text{(using Maxwell's relations)}\\
&= \mathbf{H}\cdot\dot{\mathbf{B}} + \mathbf{E}\cdot\dot{\mathbf{D}}. && (6.22)
\end{aligned}$$

So comparing the above lines (6.21) and (6.22),

$$\left(\dot{\mathbf{E}}\cdot\mathbf{D} - \mathbf{E}\cdot\dot{\mathbf{D}}\right) + \left(\dot{\mathbf{H}}\cdot\mathbf{B} - \mathbf{H}\cdot\dot{\mathbf{B}}\right) = \mathbf{0}. \tag{6.23}$$

The dielectric and magnetic responses can (usually) be taken to be independent, so each of these bracketed terms must be zero, giving

$$\mathbf{E}\cdot\dot{\mathbf{D}} = \mathbf{D}\cdot\dot{\mathbf{E}} \tag{6.24}$$

[ϵ is not a scalar, so this is not as obvious as it might look].

When $\mathbf{E}$ and $\mathbf{D}$ might be complex, in any calculation of a physical quantity like $\dot{u}$ it is necessary to take the real parts first.

$$\mathrm{Re}\{\mathbf{E}\}\cdot\mathrm{Re}\left\{\dot{\mathbf{D}}\right\} = \mathrm{Re}\{\mathbf{D}\}\cdot\mathrm{Re}\left\{\dot{\mathbf{E}}\right\}. \tag{6.25}$$

Without loss of generality, it can be assumed that $\mathbf{E}$ and $\mathbf{D}$ are varying time-harmonically. For any complex numbers a and b varying time-harmonically, the time-averaged product of the real parts is

$$\left(\mathrm{Re}\{a\}\,\mathrm{Re}\{b\}\right) = \frac{1}{2}\,\mathrm{Re}\{a^*b\} \tag{6.26}$$

so,

$$\begin{aligned}
\mathrm{Re}\left\{\mathbf{E}^*\cdot\dot{\mathbf{D}}\right\} &= \mathrm{Re}\left\{\mathbf{D}^*\cdot\dot{\mathbf{E}}\right\} = \mathrm{Re}\left\{\dot{\mathbf{E}}\cdot\mathbf{D}^*\right\}\\
\mathrm{Re}\left\{\sum_{i,j} E_i^*\epsilon_{ij}\dot{E}_j\right\} &= \mathrm{Re}\left\{\sum_{i,j}\dot{E}_i\epsilon_{ij}^*E_j^*\right\} \overset{i\rightleftharpoons j}{=} \mathrm{Re}\left\{\sum_{i,j}\dot{E}_j\epsilon_{ji}^*E_i^*\right\}\\
&\implies \quad \epsilon_{ij} = \epsilon_{ji}^*. && (6.27)
\end{aligned}$$

An important property of the dielectric tensor is that is must be Hermitian

$$\boxed{\epsilon_{ij} = \epsilon_{ji}^*,} \tag{6.28}$$

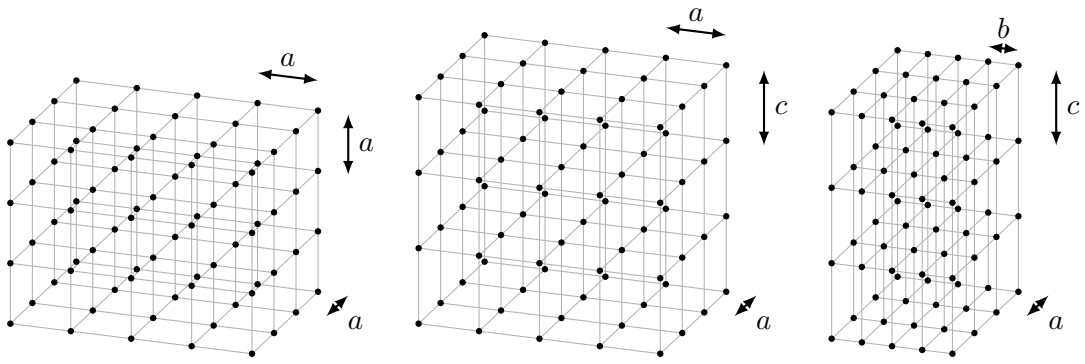
from this it follows that $\underline{\underline{\boldsymbol{\epsilon}}}$ must be diagonalisable.

For lossless media (and in the absence of "optical activity" (see Section 6.7)) the ϵ_{ij} are real, and $\underline{\underline{\boldsymbol{\epsilon}}}$ is symmetric.

So for such a material there is a set of orthogonal axes $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$ – the **principal axes** – such that:

$$\underline{\underline{\boldsymbol{\epsilon}}} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} = \begin{pmatrix} n_1^2 & 0 & 0 \\ 0 & n_2^2 & 0 \\ 0 & 0 & n_3^2 \end{pmatrix}, \tag{6.29}$$

where n_1, n_2 and n_3 are the *principal refractive indices*. If $n_1 \neq n_2 \neq n_3$, the material is *biaxial*. If two of n_1, n_2, n_3 are equal the material is *uniaxial*.



$$\underline{\underline{\boldsymbol{\epsilon}}} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_1 \end{pmatrix} \qquad \underline{\underline{\boldsymbol{\epsilon}}} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} \qquad \underline{\underline{\boldsymbol{\epsilon}}} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix}$$

Fig. 6.4: Simple crystal structures representing isotropic (cubic), uniaxial (tetragonal) and biaxial (orthorhombic) structures, from left to right

Calcite ($CaCO_3$) is an example of a uniaxial material. The crystal plane perpendicular to the optical axis has three-fold symmetry. The refractive index depends on the whether the direction of the electric field is in the plane of the triangular CO_3 clusters or perpendicular to them.

Fig. 6.5: Crystal structure of calcite ($CaCO_3$) showing the triangular CO_3 clusters oriented with the plane perpendicular to the optical axis. The right view shows the atomic arrangement looking down along the optical axis

For uniaxial systems like calcite it is conventional to take $n_1 = n_2 \neq n_3$

$$\underline{\underline{\boldsymbol{\epsilon}}} = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} = \begin{pmatrix} n_\mathrm{o}^2 & 0 & 0 \\ 0 & n_\mathrm{o}^2 & 0 \\ 0 & 0 & n_\mathrm{e}^2 \end{pmatrix}, \tag{6.30}$$

where "o" denotes the "ordinary" directions, and "e" the "extraordinary" direction – the **optic axis**.

Birefringence for a uniaxial material is given by $\Delta n = n_\mathrm{e} - n_\mathrm{o}$, and can be positive or negative. Calcite has negative birefringence, with $\Delta n = n_\mathrm{e} - n_\mathrm{o} = -0.172$.

If **E** lies along one of the principal axes of a uniaxial or biaxial medium, **D** $\parallel$ **E**. If **E** is perpendicular to the optic axis of a uniaxial medium, **D** $\parallel$ **E**.

6.4 Linearly Polarised EM Waves in Anisotropic Materials

From Eq. (4.76),

$$\mathbf{B} \perp \mathbf{k} \text{ and } \mathbf{E}, \quad \mathbf{D} \perp \mathbf{k} \text{ and } \mathbf{H}. \tag{6.31}$$

For a non-magnetic (or magnetically isotropic) material μ is a scalar so **B** $\parallel$ **H**. Then **D**, **H** ($\parallel$ **B**) and **k** are mutually orthogonal.

6.4.1 Special Symmetry Cases

If **D** lies along a principal axis (say $\hat{\mathbf{e}}_1$), **E** $\parallel$ **D**. Then $\mathbf{E} \times \mathbf{H} = \mathbf{N} \parallel \mathbf{k}$. The wave equation

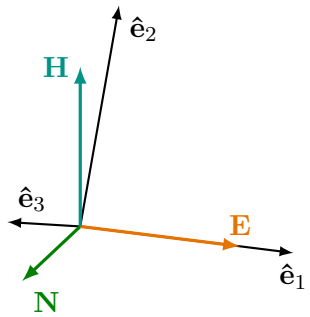


Fig. 6.6: For **D** $\parallel \hat{\mathbf{e}}_1$ the wave propagates with velocity c/n_1.

is then identical to that for an isotropic medium with an ϵ corresponding to that for the axis along which **D** and **E** are directed.

6.4.2 Uniaxial Materials

For a uniaxial material,

$$\mathbf{D} = \epsilon_0 \underline{\underline{\boldsymbol{\epsilon}}} \cdot \mathbf{E} = \epsilon_0 \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}. \tag{6.32}$$

So even if **D** $\nparallel \hat{\mathbf{e}}_1$ or $\hat{\mathbf{e}}_2$, but lies in the $\hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2$ plane for which $\epsilon_1 = \epsilon_2 = n_\mathrm{o}^2$, **D** $\parallel$ **E**.

So for **D** $\perp \hat{\mathbf{z}}$, the optic axis, **E** $\parallel$ **D** whatever its direction in this plane, and the wave velocity is c/n_o. But if **D** is not in the $\hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2$ plane, it cannot be assumed that **E** $\parallel$ **D**.

The Poynting vector $\mathbf{N} = \mathbf{E} \times \mathbf{H}$ is therefore *not necessarily parallel* to $\mathbf{k}$, thus the *phase and the energy may propagate in different directions.*

6.4.2.1 Geometric Approach

Recall that the energy density is proportional to $\mathbf{D} \cdot \mathbf{E}$. Formally, for some fixed energy density and in an appropriate system of units $\epsilon_1 \mathbf{D} \cdot \mathbf{E} = 1$ or equivalently $\mathbf{D} \cdot \epsilon^{-1} \cdot \mathbf{D} = 1$. This defines an ellipsoid, the so-called *optical indicatrix*,

$$\frac{D_x^2}{\epsilon_1} + \frac{D_y^2}{\epsilon_2} + \frac{D_z^2}{\epsilon_3} = 1. \tag{6.33}$$

For a given $\mathbf{D}$ the corresponding $\mathbf{E}$ can be shown to be normal to the ellipsoid surface at the tip of $\mathbf{D}$. The refractive index of wave with polarisation vector D can then be found as follows.

Recall $\mathbf{k} \times \mathbf{k} \times \mathbf{E} = -\mu_0 \omega^2 \mathbf{D}$. When $\mathbf{E}$ makes an angle α to the plane perpendicular to $\mathbf{k}$, $|\mathbf{k} \times \mathbf{k} \times \mathbf{E}| = k^2 E \cos\alpha$, and $\epsilon_0 E D \cos\alpha = \epsilon_1 \mathbf{E} \cdot \mathbf{D} = 1$

$$n^2 = c^2 \frac{k^2}{\omega^2} = \frac{c^2 \mu_0 D}{E \cos\alpha} = \frac{D^2}{\epsilon_0 E D \cos\alpha} = D^2, \tag{6.34}$$

i.e., the length of the radius vector of the ellipsoid in each particular direction equals the refractive index for a wave with polarisation vector $\mathbf{D}$ in that direction.

It is the polarisation direction, not the propagation direction that determines the wave velocity.

Fig. 6.7: Illustration of the optical indicatrix. Here $n_\mathrm{e} > n_\mathrm{o}$ (i.e. the material has positive birefringence) so that $v_\mathrm{e} < v_\mathrm{o}$.

This can equivalently be represented in terms of the speed of Huygen's wavelets emanating from a point in the crystal and traveling in a particular propagation direction.

1. For wavelets with $\mathbf{D} \perp \hat{\mathbf{z}}$, the optic axis, $v_\mathrm{o} = c/n_\mathrm{o}$, independent of their direction (as shown in Fig. 6.8).

   These are "ordinary" wavelets, and form spherical wavefronts.

2. For the linear polarisations orthogonal to 1, $\mathbf{D}$ lies in the $\mathbf{k}_\mathrm{w} - \hat{\mathbf{e}}_3$ plane and in general $\mathbf{E} \nparallel \mathbf{D}$.

   The wavelet speed is $v_\mathrm{e} = c/n_b$, where the effective refractive index n_b is):

$$\frac{(n_b \sin\theta)^2}{n_\mathrm{e}^2} + \frac{(n_b \cos\theta)^2}{n_\mathrm{o}^2} = 1, \tag{6.35}$$

Fig. 6.8: The wavelet's speed depends on the propagation direction and polarisation of the wave. Here $n_\mathrm{e} > n_\mathrm{o}$ (i.e. the material has negative birefringence) so that $v_\mathrm{e} > v_\mathrm{o}$.

where θ is the angle between the wavelet direction and $\hat{\mathbf{e}}_3$.

These are "extra-ordinary" wavelets, and form ellipsoidal wavefronts since the system has cylindrical symmetry around $\hat{\mathbf{z}}$.

6.4.3 Double Refraction

Consider linearly-polarised light normally incident on a surface $\mathcal{S}$ of a *uniaxial* crystal (e.g. calcite): $\mathbf{k}_{\text{inc}} \parallel \hat{\mathbf{n}}_{\mathcal{S}}$, the surface normal. Take the optic axis $\hat{\mathbf{e}}_3$ to be at an angle θ to $\hat{\mathbf{n}}_{\mathcal{S}}$ in the plane of figure 6.9.

Fig. 6.9: Light linearly polarised perpendicular to the plane of the diagram

Inside the crystal, $\mathbf{k}$ is the wavevector for the *transmitted* ray formed from the superposition of many *Huygen's wavelets* propagating in all directions $\mathbf{k}_{\text{w}}$. $\mathbf{k} \parallel \hat{\mathbf{n}}_{\mathcal{S}}$, since $\mathbf{k}_{\parallel}$ $(= 0)$ is conserved at any interface, and $\mathbf{k}$ makes an angle θ with $\hat{\mathbf{e}}_3$.

Case (a) (Fig. 6.9) $\mathbf{D} \perp \hat{\mathbf{e}}_3$, $\mathbf{D}$ lies in the $\hat{\mathbf{e}}_1 - \hat{\mathbf{e}}_2$ plane again, so $\mathbf{E} \parallel \mathbf{D}$ whatever its direction in this plane.

The wavelets for the Huygens construction have speed c/n_{o}, independent of direction, and are therefore *spherical*

$$\mathbf{E} \times \mathbf{H} = \mathbf{N} \parallel \mathbf{k} \tag{6.36}$$

- This is the "ordinary" ray;
- At non-normal incidence (in the plane of Fig. 6.9 so that $\mathbf{D}$ remains perpendicular to the optic axis) the the ordinary ray would refract in the ordinary way corresponding to a medium with refractive index n_{o}.

Case (b) (Fig. 6.10) For the linear polarization orthogonal to (a), $\mathbf{D}$ lies in the plane including $\hat{\mathbf{e}}_3$, and in general $\mathbf{E} \nparallel \mathbf{D}$. The wavelet speed is c/n_b, where n_b is given by Eq. (6.35), and the Huygens wavelets are *ellipsoidal*.

Fig. 6.10: Light linearly polarised perpendicular to the plane of the diagram.

The tangent planes to the superposition of these ellipsoidal wavelets give the overall wavefronts for the propagating ray, and the direction of $\mathbf{k}$ for this ray remains normal to $\mathcal{S}$.

$\mathbf{D}$ is necessarily perpendicular to $\mathbf{k}$ as shown, BUT in general $\mathbf{E} \nparallel \mathbf{D}$, so

$$\mathbf{E} \times \mathbf{H} = \mathbf{N} \nparallel \mathbf{k}, \tag{6.37}$$

- The phase again propagates along the surface normal nbS;
- The energy now propagates at an angle to the normal;
- This ray – "*extraordinary*" ray – is therefore *laterally shifted* when it emerges from the crystal.

So an object viewed through a uniaxial crystal produces two images, one for the ordinary rays and one for the extraordinary rays – *double refraction*. The images obviously have different polarization properties. Double refraction in calcite was an early pointer to the polarization properties of light.

Fig. 6.11: Double refraction in calcite

6.5 Optical Elements: Waveplates (or Retarders)

Consider the simplified case when **D** and **E** ∥ a principal axis (see Subsection 6.4.1), this can be exploited in the construction of *waveplates* (*retarders*) for normal incidence on a plate as shown in Fig. 6.12. A plane-polarized EM wave $e^{i(kz-\omega t)}$ travels along $\hat{\mathbf{z}}$ at different speeds c/n_f or c/n_s depending on whether **E** ∥ $\hat{\mathbf{x}}$ or $\hat{\mathbf{y}}$.

For $\mathbf{L}_x$, $e^{ik(z=0)} \to e^{ik_\mathrm{f}(z=d)}$ as it traverses the plate, where $k_\mathrm{f} = \omega n_\mathrm{f}/c$. So the plate applies phase terms depending on the different *optical thicknesses*:

$$\left.\begin{matrix} e^{i\omega n_\mathrm{f} d/c} \\ e^{i\omega n_\mathrm{s} d/c} \end{matrix}\right\} \quad \text{to} \quad \left\{\begin{matrix} \mathbf{L}_x \\ \mathbf{L}_y \end{matrix}\right. . \tag{6.38}$$

The Jones matrix for the plate is

$$\begin{pmatrix} e^{i\omega n_\mathrm{f} d/c} & 0 \\ 0 & e^{i\omega n_\mathrm{s} d/c} \end{pmatrix} \propto \begin{pmatrix} e^{i\omega (n_\mathrm{f}-n_\mathrm{s}) d/c} & 0 \\ 0 & 1 \end{pmatrix} \propto \begin{pmatrix} e^{-i\Delta\phi/2} & 0 \\ 0 & e^{i\Delta\phi/2} \end{pmatrix}, \tag{6.39}$$

where $\Delta\phi = \omega(n_\mathrm{f} - n_\mathrm{s})d/c$ is the phase difference induced by the plate for waves polarised along the fast and slow axes.

$$\left\{\begin{matrix} \Delta\phi = \pi/2 \\ \Delta\phi = \pi \end{matrix}\right\} \to \left\{\begin{matrix} \lambda/4 \\ \lambda/2 \end{matrix}\right\} \text{ in vaccuum – a } \left\{\begin{matrix} \text{quarter-wave plate} \\ \text{half-wave plate} \end{matrix}\right\} \tag{6.40}$$

So for a $\lambda/4$ plate with fast axis along $\hat{\mathbf{x}}$ the Jones matrix is:

$$\underline{\underline{\mathbf{J}}}_{\lambda/4,x} = \begin{pmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{pmatrix} = e^{-\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \tag{6.41}$$

For the fast axis along $\hat{\mathbf{y}}$:

$$\underline{\underline{\mathbf{J}}}_{\lambda/4,y} = \begin{pmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{pmatrix} = e^{\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix} \tag{6.42}$$

(Note the prefactors have no net effect on the polarisation state.)

Fig. 6.12: Caption

So $\lambda/2$ plates:

$$\underline{\underline{\mathbf{J}}}_{\lambda/2,x} = \begin{pmatrix} e^{-i\pi/2} & 0 \\ 0 & e^{i\pi/2} \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = e^{-i\pi/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{6.43}$$

and,

$$\underline{\underline{\mathbf{J}}}_{\lambda/2,y} = \begin{pmatrix} e^{i\pi/2} & 0 \\ 0 & e^{-i\pi/2} \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = e^{i\pi/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{6.44}$$

Quarter- and half-wave plates can, with linear polarisers, be used to manipulate and analyse the polarisation state of light in detail.

Suppose a plane polarized wave is incident on a wave plate (fast axis along $\hat{\mathbf{x}}$) with **E**-vector at angle θ to $\hat{\mathbf{x}}$. The incident wave has Jones vector $(\cos\theta, \sin\theta)$, so the transmitted wave is

$$\begin{pmatrix} e^{-i\Delta\phi/2} & 0 \\ 0 & e^{i\Delta\phi/2} \end{pmatrix}\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} = \begin{pmatrix} e^{-i\Delta\phi/2}\cos\theta \\ e^{i\Delta\phi/2}\sin\theta \end{pmatrix} \tag{6.45}$$

(i) If $\Delta\phi = \pi/2$ – a *quarter-wave plate*: $\rightarrow (\cos\theta, i\sin\theta)$. (Dropping prefactor.)

From Section 6.1 this is seen to be *elliptically polarised* light with $\alpha = 0$. i.e. axes of the ellipse lie along $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ and have lengths $\cos\theta$ and $\sin\theta$.

If $\theta = 45°$ this is circularly polarised – LCP.

If $\theta = 0$, $90°$ this is linearly polarised.

(ii) If $\Delta\phi = \pi$ – a *half-wave plate*: $\rightarrow (\cos\theta, -\sin\theta)$.

Plane polarised light with the **E**-vector directed at $-\theta$ to $\hat{\mathbf{x}}$. The direction of plane polarisation is *rotated*.

If also $\theta = 45°$, the plane of polarisation becomes perpendicular to the original.

(iii) if $\theta = 0$ or $\pi/2$, the incident plane polarised wave is unaffected whatever the value of the plate thickness d, since the incident **E** is parallel to one of the principal axes of the plate.

Summarising,

***x*-polariser**	***y*-polariser**	***θ*-polariser**	**λ/4 plate**	**λ/2 plate**
$\underline{\underline{\mathbf{J}}}_x$	$\underline{\underline{\mathbf{J}}}_y$	$\underline{\underline{\mathbf{J}}}_\theta$	$\underline{\underline{\mathbf{J}}}_{\lambda/4,x}$	$\underline{\underline{\mathbf{J}}}_{\lambda/2,x}$
$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

Table 6.2: Caption

6.6 Induced Birefringence

Optical anisotropy can be induced in otherwise isotropic materials

6.6.1 Photoelasticity

Photoelasticity (or stress birefringence) is the birefringence induced when an otherwise isotropic material is subjected to *stress*. The corresponding distortion of the material, at molecular level, changes the dielectric response, producing an anisotropic permittivity tensor. A transparent *isotropic* object placed between crossed linear polarisers would not change the initial polarisation, so no light should be transmitted.

But if stressed to induce birefringence, linearly polarized light passing through the material has its polarisation state affected in complex ways depending on the stress field, and some light is transmitted, allowing patterns of stress in transparent mechanical structures to be visualised.

6.6.2 The Kerr and Pockels Effects

In an applied electric field $\mathbf{E}_0$ an otherwise isotropic material can become uniaxially birefringent, with the optic axis along $\mathbf{E}_0$. In liquids and gases, this can be understood as arising from the alignment of anisotropic molecules by the field.

Since in an otherwise isotropic liquid or gas the optical properties cannot be sensitive to the sign of the field the change in the refractive index must be quadratic in the electric field to lowest order: $\Delta n = \lambda_0 K E_0^2$. K is the Kerr constant. The Kerr effect is an example of a non-linear optical phenomenon.

In solids a similar effect, the so-called Pockels effect, is associated with the lowering of the crystal symmetry by the induced *macroscopic* dielectric polarisation. Crystals that do not have a centre of inversion symmetry could distinguish between positive and negative fields. Therefore, a linear electric field dependence is possible for the Pockels effect.

Suitable materials can therefore be used to make *voltage-controlled wave-plates*.

Fig. 6.13: Schematic of a Kerr cell, after Hecht Fig. 8.56. The applied electric fields can switch the properties of the dielectric at high frequency, forming the basis of fast optical modulators.

6.7 Optical Activity

6.7.1 Chiral Materials

Some materials, while isotropic, have molecules with a *chiral* structure – a *handedness* built in at molecular level. The defining characteristic is that the *mirror image* of the molecular structure cannot be superposed on the original. An important biological example would be an α-amino acid of the form $H_2NCHRCOOH$.

Fig. 6.14: Left: the molecule and its mirror image cannot coincide. Right: Right and Left handed helices. For natural materials (e.g. dextrose, quartz) one type, right or left handed is usually prevalent.

This *chirality* is in-built, *even in liquid form with no molecular organisation or crystal structure.*

A chiral material is *optically active*, or *circularly birefringent*, and responds differently to LCP and RCP waves. These then are the natural polarisation states to use – the characteristic waves for the medium – with two refractive indices, n_{L} and n_{R} (just as for birefringence and plane polarisation).

A chiral wave plate introduces different phases for LCP and RCP waves, just as a birefringent plate does for $\mathbf{L}_x$ and $\mathbf{L}_y$.

Fig. 6.15: A slab of chiral material different optical thicknesses $n_{\mathrm{L}}d$ and $n_{\mathrm{R}}d$, for LCP and RCP light.

A plane polarised wave can be written as the sum of two counter-rotating circularly polarised waves,

$$\mathbf{L}_x = \begin{pmatrix}1\\0\end{pmatrix} = \frac{1}{\sqrt{2}}\left[\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-i\end{pmatrix} + \frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix}\right] = \frac{\mathbf{C}_{\mathrm{L}} + \mathbf{C}_{\mathrm{R}}}{\sqrt{2}}. \tag{6.46}$$

A chiral plate of thickness d applies a phase term of $e^{i\omega n_{\mathrm{R,L}}d/c}$ to $\mathbf{C}_{\mathrm{R,L}}$ (cf. Eq. (6.39)).

Using $\Delta\phi = \omega(n_{\mathrm{L}} - n_{\mathrm{R}})d/c$ as the relative phase, the $\mathbf{L}_x$ wave above becomes

$$\frac{1}{\sqrt{2}}\left[\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-i\end{pmatrix}e^{-i\Delta\pi/2} + \frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix}e^{i\Delta\pi/2}\right] = \begin{pmatrix}\cos(\Delta\phi/2)\\-\sin(\Delta\phi/1)\end{pmatrix}, \tag{6.47}$$

another *plane polarised* wave, but with its plane *rotated clockwise* by

$$\Delta\phi/2 = \frac{(n_{\mathrm{L}} - n_{\mathrm{R}})d}{2c} = \frac{\pi(n_{\mathrm{L}} - n_{\mathrm{R}})d}{\lambda}, \tag{6.48}$$

where λ is the wavelength in air.

The rotation per unit length, the *specific rotatory power*, is:

$$\alpha = \frac{\pi(n_{\mathrm{L}} - n_{\mathrm{R}})}{\lambda} = \frac{1}{2}(k_{\mathrm{L}} - k_{\mathrm{R}}) = \frac{\omega}{2c}(n_{\mathrm{L}} - n_{\mathrm{R}}). \tag{6.49}$$

If the plane of polarisation has rotated *clockwise* ($\alpha > 0$, $n_{\mathrm{L}} > n_{\mathrm{R}}$), the medium is said

Fig. 6.16: The orientation of plane-polarised light is continuously rotated as the light passes through an optically active dextrorotatory medium.

to be *dextrorotatory*, or *d-rotatory*; if anticlockwise ($\alpha < 0$, $n_{\mathrm{L}} < n_{\mathrm{R}}$), *levorotatory*, or *l-rotatory.*

Fig. 6.17: The Faraday geometry

6.7.2 The Faraday Effect

Birefringence can be introduced by applying an electric field to an isotropic material. Chirality can be introduced to a non-chiral system by an applied magnetic field which alters the response of the electrons in the system to the optical fields. Applied field $\mathbf{B}_0$ is directional implies medium is not isotropic and so geometry matters. The basic points can be illustrated with a simple example: EM waves with $\mathbf{k} \parallel \hat{\mathbf{z}}$ in a plasma, with $\mathbf{B}_0 \parallel \hat{\mathbf{z}}$, the *Faraday geometry*.

For each electron the equation of motion is (neglecting any scattering and the negligible effect of the magnetic field of the EM wave)

$$m\ddot{\mathbf{r}} = -e(\mathbf{E} + \dot{\mathbf{r}} \times \mathbf{B}_0), \tag{6.50}$$

where $\mathbf{E} = (E_x, E_y)e^{-i\omega t}$ is the electric field of the incident EM wave. $\mathbf{r}$ will also vary harmonically as $\mathbf{r} = (x, y) = (x_0, y_0)e^{-i\omega t}$ then,

$$m\ddot{x} = -eE_x - eB_0\dot{y} \implies -\omega^2 x_0 = -\frac{e}{m}E_x + i\omega\frac{eB_0}{m}y_0 \tag{6.51}$$

$$m\ddot{y} = -eE_y + eB_0\dot{x} \implies -\omega^2 y_0 = -\frac{e}{m}E_y - i\omega\frac{eB_0}{m}x_0, \tag{6.52}$$

so, with $eB_0/m = \omega_c$, the *cyclotron frequency*,

$$-\omega^2(x_0 + iy_0) = -\frac{e}{m}(E_x + iE_y) + \omega\omega_c(x_0 + iy_0) \tag{6.53}$$

$$-\omega^2(x_0 - iy_0) = -\frac{e}{m}(E_x - iE_y) - \omega\omega_c(x_0 - iy_0) \tag{6.54}$$

If $|E_x| = |E_y| = E$ then $|x_0| = |y_0| = a$ say, and Eq. (6.54) corresponds to an oscillator driven by a LCP EM wave $\mathbf{C}_\mathrm{L} = E(1, i)$, with a LCP circular displacement $a(1, i)$ as the response.

So (with n the electron number density) there is a *circular polarisation*:

$$\mathbf{P}_\mathrm{L} = -ena\begin{pmatrix}1\\ i\end{pmatrix} = \epsilon_0\chi_\mathrm{L}E\begin{pmatrix}1\\ i\end{pmatrix}, \tag{6.55}$$

where the effective susceptibility for $\mathbf{C}_\mathrm{L}$ is

$$\chi_\mathrm{L} = -\left(\frac{ne^2}{\epsilon_0 m}\right)\frac{1}{\omega^2 - \omega\omega_c} = -\frac{\omega_p^2}{\omega^2 - \omega\omega_c} \tag{6.56}$$

with ω_p the plasma frequency. Similarly for $\mathbf{C}_\mathrm{R}$ waves from Eq. (6.53)

$$\chi_\mathrm{R} = -\frac{\omega_p^2}{\omega^2 + \omega\omega_c}. \tag{6.57}$$

So,

$$\epsilon_{\frac{\mathrm{L}}{\mathrm{R}}}(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega \mp \omega_c)}. \tag{6.58}$$

If $B_0 = 0$, $\omega_c = 0$ and the familiar result for a simple plasma is recovered.

So n_L and n_R are different, just as for optical activity, and the plane of polarisation of a plane polarised wave is steadily rotated as it passes along $\hat{\mathbf{z}}$.

From 6.48, the angle of rotation is

$$\theta = \frac{\Delta\phi}{2} = \frac{\omega(n_L - n_R)d}{2c}$$
$$= \frac{\omega d}{2c}\left[\left(1 - \frac{\omega_p^2}{\omega(\omega - \omega_c)}\right) - \left(1 - \frac{\omega_p^2}{\omega(\omega + \omega_c)}\right)\right] \tag{6.59}$$
$$\approx -\frac{\omega_p^2 \omega_c d}{2c\omega^2\sqrt{1 - \frac{\omega_p^2}{\omega^2}}} \qquad \text{(if } B_0\text{, and hence } \omega_c\text{, is small)} \tag{6.60}$$

The *Verdet coefficient* V is defined from

$$\theta = V B_0 d, \tag{6.61}$$

so for the plasma in a weak field B_0,

$$V = -\frac{e\omega_p^2}{2mc\omega^2\sqrt{1 - \frac{\omega_p^2}{\omega^2}}} \tag{6.62}$$

6.7.2.1 The Dielectric Tensor

Solving Eqs. (6.53) and (6.54) for x_0 and y_0,

$$x_0 = \frac{\frac{e}{m}E_x - \frac{i\omega_c}{\omega}\frac{e}{m}E_y}{(\omega^2 - \omega_c^2)} \tag{6.63}$$
$$y_0 = \frac{\frac{e}{m}E_y + \frac{i\omega_c}{\omega}\frac{e}{m}E_x}{(\omega^2 - \omega_c^2)} \tag{6.64}$$

with a corresponding polarisation density

$$\mathbf{P} = -ne\begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} e^{-i\omega t} = \epsilon_0 \underline{\underline{\boldsymbol{\chi}}} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} e^{-i\omega t}. \tag{6.65}$$

$\mathbf{B}_0 \parallel \hat{\mathbf{z}}$ so the z-motion is unaffected by the magnetic field and the z susceptibility is as for an unmagnetised plasma.

So,

$$\underline{\underline{\boldsymbol{\epsilon}}} = \underline{\underline{\mathbf{1}}} + \underline{\underline{\boldsymbol{\chi}}} = \begin{pmatrix} 1 - \frac{\omega_p^2}{\omega^2 - \omega_c^2} & \frac{i\omega_c\omega_0^2}{\omega(\omega^2 - \omega_c^2)} & 0 \\ -\frac{i\omega_c\omega_0^2}{\omega(\omega^2 - \omega_c^2)} & 1 - \frac{\omega_p^2}{\omega^2 - \omega_c^2} & 0 \\ 0 & 0 & 1 - \frac{\omega_p^2}{\omega^2} \end{pmatrix} \tag{6.66}$$

is the dielectric matrix for a plasma in the presence of $\mathbf{B}_0 \parallel \hat{\mathbf{z}}$ in the absence of damping.

The *off-diagonal terms* reflect the *magnetically-induced chirality* of the system.

The dielectric matrix is Hermitian, as required from Eq. (6.28).

If $\mathbf{B}_0 \to \mathbf{0}$ and $\omega_c \to 0$, the familiar result for an isotropic plasma emerges.

6.8 Interference and Partial Polarisation

6.8.1 Interference of Polarised Waves

Acoustics, QM, etc., interference is between scalar waves, but with EM wave interference must consider their vector nature. Consider the superposition of two waves along $\hat{\mathbf{z}}$ perpendicularly plane polarised with arbitrary phase difference δ (c.f. elliptically polarised wave 6.1). The net Poynting vector is

$$\begin{aligned}\mathbf{N} &= [\hat{\mathbf{x}}E_{1x}\cos\omega t + \hat{\mathbf{y}}E_{2y}\cos(\omega t+\delta)] \\ &\quad \times [\hat{\mathbf{x}}H_{2x}\cos(\omega t+\delta) + \hat{\mathbf{y}}H_{1y}\cos\omega t] \\ &= \hat{\mathbf{z}}\Big[E_{1x}H_{1y}\cos^2\omega t - E_{2y}H_{2x}\cos^2(\omega t+\delta)\Big],\end{aligned} \tag{6.67}$$

identical to the result taking the two plane polarised waves *independently*. (Remember that H_{2x} is negative.)

The *intensity at a point* produced by *two* superposed *perpendicularly plane polarised* waves is the *sum of the intensities* of the two waves. Perpendicularly plane polarised waves *do not interfere*.

6.8.2 Unpolarised and Partially Polarised Light

Light from most natural sources has the *direction of the* **E***-vector changing randomly* with time and space (as well as variations in its amplitude and phase as discussed later for coherence in Section 6.9).

It is *unpolarised*, and time-variations of the x and y components of **E** are *uncorrelated*.

Plane polarised beams produced by passing the same beam of unpolarised light through x and y polarisers are therefore *mutually incoherent*. There is no well-defined phase difference – so they *cannot interfere*.

A beam which includes both polarised and unpolarised light (with intensities I_{pol} and I_{unpol}) is *partially polarised*, with a *degree of polarisation*:

$$V = \frac{I_{\text{pol}}}{I_{\text{pol}} + I_{\text{unpol}}}. \tag{6.68}$$

6.8.3 The Fresnel-Arago Laws

Summarising the conclusions of Subsections 6.8.1 and 6.8.2 we have the **Fresnel-Arago laws**:

1. Two beams, plane polarised parallel, interfere (if coherent);
2. Two beams, plane polarised perpendicularly, cannot interfere (even if perfectly coherent);
3. Two plane polarised beams cannot interfere (even if polarised parallel) if they are *derived* from perpendicularly polarised components of *unpolarised* light since these must be *mutually incoherent*.

6.9 Coherence

For simplicity, in this discussion we consider scalar waves. Interference phenomena, diffraction etc., rely on the *well defined phase* between wavelets (determined by optical path lengths, etc.) which are eventually summed for the overall wave amplitude. This can be exact only for *purely monochromatic* waves – a single, well-defined frequency. However, Fourier Theory means that only *infinitely long* (in time and space) waves can be purely monochromatic.

No such waves actually exist. Real sources and waves (even lasers) are at best *quasi-monochromatic*

Fig. 6.18: A quasi-monochromatic waveform deviates in amplitude and phase from the pure reference wave

The formal theory of *coherence* was developed by Zernike in 1938 building on earlier work of Michelson and Fizeau. It describes non-monochromatic wavefields *quantitatively* and formed the basis of many modern experimental techniques in radio and optical astronomy and spectroscopy for materials characterisation.

Fig. 6.19: A highly schematic representation of a partially coherent wavefield. **Phase registration** is lost over a distance $c\tau_c - \tau_c$ is the (*temporal*) *coherence length* along the direction of propagation – and over a distance x_c – the (*spatial*) *coherence width* – perpendicular to the direction of propagation.

6.9.1 The Power Spectrum

Using time as the variable, the Fourier Transform relates the *time domain* t to the *frequency domain* ω.

A time-dependent function $f(t)$ can be described in terms of its *temporal harmonics* $F(\omega)e^{i\omega t}$ (Note the sign convention is not that used for waves earlier)

$$\boxed{f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}\,\mathrm{d}\omega\,, \quad F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,\mathrm{d}t} \tag{6.69}$$

The power in the frequency range ω to $\omega + \mathrm{d}\omega$, *the Power Spectrum*, is:

$$\boxed{P(\omega)\,\mathrm{d}\omega \sim |F(\omega)|^2\,\mathrm{d}\omega} \tag{6.70}$$

6.9.1.1 Lasers

For a pure harmonic wave of frequency ω_0, $f(t) \sim \cos(\omega t + \alpha)$

$$\begin{aligned} F(\omega) &= \mathcal{F}[f(t)] \\ &= \frac{1}{2}\int \left(e^{i(\omega_0 t+\alpha)} + e^{-i(\omega_0 t+\alpha)}\right)e^{-i\omega t}\,\mathrm{d}t \qquad (6.71)\\ &\propto e^{i\alpha}\delta(\omega-\omega_0) + e^{-i\alpha}(\omega+\omega_0). \qquad (6.72)\end{aligned}$$

So the power spectrum for a pure harmonic wave is a pair of δ-functions at $\pm\omega_0$.

This is the case (almost) for a *laser*, where stimulated emission produces an (almost) perfectly harmonic wave - an (almost) ideal *line source*.

6.9.1.2 Spectral Lines

E.g. gas discharge lamps, astrophysics.

Lifetime Broadening Unstimulated emission from an isolated, stationary atom can be represented semi-classically as a decaying harmonic wave, beginning at $t = 0$ and characterised by a decay time τ_s or a scattering frequency $\omega_\mathrm{s} = 1/\tau_\mathrm{s}$

$$F(\omega) = \int_0^{\infty} e^{i\omega_0 t}e^{-\omega_\mathrm{s} t}e^{-i\omega t}\,\mathrm{d}t = \frac{1}{\omega-\omega_0-i\omega_\mathrm{s}} \tag{6.73}$$

$$P(\omega) \sim |F(\omega)|^2 = \frac{1}{(\omega-\omega_0)^2+\omega_\mathrm{s}^2} \tag{6.74}$$

$P(\omega)$ is now a *Lorentzian peak* centred on ω_0 and with a linewidth (FWHM) of $2\omega_\mathrm{s}$, determined by the decay time τ_s.

Fig. 6.20: The Lorentzian lineshape ($\equiv$ the response of a damped oscillator).

Thermal Broadening Radiation from atoms moving along the line of sight (say the x-axis) with velocity v_x will be *Doppler-shifted* in frequency

$$\omega - \omega_0 = \frac{\omega_0 v_x}{c}. \tag{6.75}$$

From kinetic theory, the distribution of atomic/molecular velocities along $\hat{\mathbf{x}}$ is *Gaussian*

$$f(v_x)\,\mathrm{d}v_x \sim \exp\left(-\frac{mv_x^2}{2k_BT}\right)\mathrm{d}v_x\,, \tag{6.76}$$

so that the observed frequency spectrum will also be a Gaussian:

$$P(\omega) = C\exp\left(-\frac{m(\omega-\omega_0)^2c^2}{2\omega_0^2k_BT}\right) = C\exp\left(-\frac{(\omega-\omega_0)^2}{2\sigma^2}\right). \tag{6.77}$$

So, even neglecting the natural linewidth, if the atoms form a gas there will be Doppler broadening and the observed spectrum is again not a δ-function but a narrow line in ω-space of linewidth (FWHM),

$$\mathrm{FWHM} = 2.36\,\sigma = 2.36\,\omega_0\left(\frac{k_BT}{mc^2}\right)^{1/2}. \tag{6.78}$$

Not surprisingly, this depends on the *temperature* T of the gas.

Pressure Broadening In a gas, an individual atom is subject to collisions with other atoms, which at the very least perturb the phase correlation of the emitted wave before and after each collision.

Fig. 6.21: Amplitude profile for atom subjected to collisions

The mean time τ_1 between collisions is

$$\tau_1 = \frac{1}{4N\overline{v}A} = \frac{b\sqrt{T}}{p}, \tag{6.79}$$

where N is the number density of atoms of collision cross-section A and with mean velocity $\overline{v}$, and p is the pressure.

Without detailed mathematics, this also produces (irrespective of the natural lifetime τ) another Lorenztian frequency profile

$$P(\omega) = |\Psi(\omega)|^2 \sim \frac{1}{(\omega-\omega_0)^2 + 1/\tau_1^2}. \tag{6.80}$$

At fixed T (i.e. fixed $\overline{v}$) the collision rate obvious increases with p, and so does the observed linewidth, *pressure broadening*. So usually try to run lamps at low pressure to give sharpest lines.

Overall So for a gas discharge lamp, there are several line broadening mechanisms, with either Lorentzian or Gaussian profiles, and the overall lineshape is some mixture of the two depending on the gas conditions.

For a gas discharge lamp, the output is the superposition of large numbers of independent photons from individual similar atoms.

This is essentially harmonic with frequency ω_0 say, but with an amplitude and phase which have some random fluctuations, *quasi-monochromatic light*. ω_0 is the underlying

Fig. 6.22: A quasi-monochromatic waveform and the corresponding frequency spectrum

harmonic wave, and the linewidth $\Delta\omega$ can be related to some overall broadening equivalent to a lifetime τ_s.

At the extreme, a large number of atoms of different emission frequencies, or oscillators in the surface of an incandescent black body, result in *white light* with a very broad power spectrum covering the visible. Totally irregular profile implies many Fourier frequencies

Fig. 6.23: The amplitude and frequency spectrum for a *white light* source

and so a wide power spectrum results in zero coherence.

6.9.2 Coherence and Interference

If one illuminated a Michelson interferometer or a double slit experiment with incoherent, white light the waveform at one point in time/space is not correlated / not coherent with that at another so there are generally *no interference effects*.

But obviously for a strongly correlated, strongly *coherent*, waveform (a laser) interference effects are very clear.

But what about diffraction and interference using an intermediate waveform, quasi-monochromatic, *partially coherent*, light?

It turns out that interference ideas provide a useful quantified description for the *degree of correlation*, or *degree of coherence*, of the partially coherent wavefield arising from a light source.

6.9.3 A Partially Coherent Wavefield

Coherence means that a recipe exists that allows the prediction of the amplitude and phase of the wavefield at some location $\mathbf{r}_2$ and time t_2 from the knowledge of the amplitude and phase of the wavefield at location $\mathbf{r}_1$ and time t_1.

Fig. 6.24: A highly schematic representation of a partially coherent wavefield. **Phase registration** is lost over a distance $l_c = c\tau_c$ (where τ_c is the *temporal coherence length*) along the direction of propagation, and over a distance w_c, the *spatial coherence width*, perpendicular to the direction of propagation.

6.9.4 The "Optical Stethoscope"

Fig. 6.25: The optical stethoscope (after Lipson et al.) is an imaginary device for investigating the time and spatial variation of wavefields and their *temporal* (a) and *spatial coherence* (b). Two identical optical fibres sample the wavefield at points A_1 and A_2 and transfer the amplitudes of the wavefield to the closely spaced points B_1 and B_2 which act as point sources to generate an interference pattern on a screen P. The fibres are lossless and introduce identical phase shifts which can be ignored.

The *time average* for any function $g(t)$ is defined as

$$\langle g(t)\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(t)\,\mathrm{d}t\,. \tag{6.81}$$

If one brings the ends of the two fibres, B_1 and B_2, close together into a point the resultant intensity at this point is

$$I \sim \langle (A_1+A_2)(A_1+A_2)^*\rangle = \left\langle |A_1|^2\right\rangle + \left\langle |A_2|^2\right\rangle + \langle A_1A_2^*\rangle + \langle A_2A_1^*\rangle\,, \tag{6.82}$$

here the cross-terms determine any *interference effects*.

Suppose the stethoscope samples the wavefield f at two different points at different times,

$$f_1 \text{ at } \mathbf{r}_1 \text{ at } t, \quad f_2 \text{ at } \mathbf{r}_2 \text{ at } t-\tau \tag{6.83}$$

then the (complex) *mutual coherence function* Γ is defined as

$$\boxed{\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) = \langle f_1(\mathbf{r}_1,t)f_2^*(\mathbf{r}_2,t-\tau)\rangle\,.} \tag{6.84}$$

Intensity-normalising gives the complex *degree of mutual coherence* as

$$\boxed{\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) = \frac{\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)}{\sqrt{I_1I_2}},} \tag{6.85}$$

where I_1 and I_2 are the mean intensities at $\mathbf{r}_1$ and $\mathbf{r}_2$,

$$I_1 = \langle f_1(\mathbf{r}_1,t)f_1^*(\mathbf{r}_1,t)\rangle = \Gamma(\mathbf{r}_1,\mathbf{r}_1,0) \tag{6.86}$$

$$I_2 = \langle f_2(\mathbf{r}_2,t)f_2^*(\mathbf{r}_2,t)\rangle = \Gamma(\mathbf{r}_2,\mathbf{r}_2,0) \tag{6.87}$$

But what does $\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ mean physically? How is it useful?

$\gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ determines how effectively the disturbances (wavelets) originating from $\mathbf{r}_1$ and $\mathbf{r}_2$ can interfere, if I_1 and I_2 are equal and $|\gamma| \sim$, then the resultant intensity I can vary down to zero, giving good fringe contrast, as will become clear.

We can consider two limiting cases as in Fig. (6.25) (a) and (b):

(a) The system examines how the wavefield differs *along the direction of propagation*. It compares the wavefield $A_1 = f(t)$ at one time with the wavefield $A2 = f(t-\tau)$ at some earlier time, at the "same point" on the wavefront: $\tau \neq 0$, $\mathbf{r}_1 = \mathbf{r}_2$.

The configuration is an example of *amplitude division*. i.e. it principally examines the *time dependence* of the field, and therefore it is **Temporal** (or longitudinal) **Coherence** (see Subsection 6.9.5)

(b) The system examines how the wavefield differs *across the wavefront*. It compares the wavefield A_1 at one point $\mathbf{r}_1$ in space with the wavefield A_2 at some other point $\mathbf{r}_2$, *at the same time*: $\tau = 0$, $\mathbf{r}_1 \neq \mathbf{r}_2$.

The configuration is an example of *wavefront division*. i.e. it principally examines the *space dependence* of the field, and it is **Spatial** (or transverse, or lateral) **Coherence** (see Subsection 6.9.6)

6.9.5 Temporal Coherence

A practical arrangement for the "stethoscope" might be The *time delay* between the two

Fig. 6.26: A set-up for examining temporal/longitudinal coherence

rays $\tau = 2d/c$ can be altered *spatially* by moving the retro-reflector. This is *amplitude division*, so $\mathbf{r}_1 \equiv \mathbf{r}_2$ and the spatial coordinates will be omitted for convenience in the following. The divided amplitudes are taken to be equal: $A_1 = f(t)$, $A_2 = f(t-\tau)$.

The output intensity depends on the *spatially introduced time interval* τ

$$\begin{aligned} I(\tau) &= \langle (A_1 + A_2)(A_1^* + A_2^*) \rangle \\ &= \langle [f(t) + f(t-\tau)][f^*(t) f^*(t-\tau)] \rangle \\ &= \langle f(t) f^*(t) \rangle + \langle f(t-\tau) f^*(t-\tau) \rangle + \langle f(t) f^*(t-\tau) \rangle + \langle f(t-\tau) f^*(t) \rangle \\ &= 2I_0 + \Gamma(\tau) + \Gamma^*(\tau) \end{aligned} \tag{6.88}$$

6.9.6 Spatial Coherence

Chapter 7

Special Relativity

Chapter 8

Radiation and Relativistic Electrodynamics

Appendix A

Appendix

A.1 Green's Functions

A.1.1 Second-Order Linear Ordinary Differential Equations

The general *second-order linear* ordinary differential equation (ODE) for $y(x)$ can, wlog, be written as

$$y'' + p(x)y' + q(x)y = f(x) \quad \text{or} \quad \mathcal{L}y(x) = f(x), \tag{A.1}$$

where $\mathcal{L}$ is the differential operator

$$\mathcal{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + p(x)\frac{\mathrm{d}}{\mathrm{d}x} + q(x), \tag{A.2}$$

If $f(x) = 0$ the equation is said to be *homogeneous* (unforced), otherwise it is said to be *inhomogeneous* (forced).

A.1.1.1 Homogeneous Second-Order Linear ODEs

If f $= 0$ then any two solutions of

$$y'' + p(x)y' + q(x)y = 0, \tag{A.3}$$

can be superposed to give a third, i.e. if y_1 and y_2 are two solutions then for $\alpha, \beta \in \mathbb{R}$ another solution is

$$y = \alpha y_1 + \beta y_2. \tag{A.4}$$

Further, suppose that y_1 and y_2 are two *linearly independent* solutions, where by linearly independent we mean that

$$\alpha y_1(x) + \beta y_2(x) \equiv 0 \quad \Longrightarrow \quad \alpha = \beta = 0. \tag{A.5}$$

Then since (A.3) is second order, the *general solution* of (A.3) will be of the form (A.4). $y_1(x)$ and $y_2(x)$ are often referred to as *complementary functions*, while the parameters α and β can be viewed as the two integration constants. This means that in order to find the general solution of a second order linear homogeneous ODE we need to find two linearly-independent solutions.

If y_1 and y_2 are linearly dependent then $y_2 = \gamma y_1$ for some $\gamma \in \mathbb{R}$, in which case (A.4) becomes

$$y = (\alpha + \beta\gamma)y_1, \tag{A.6}$$

and we have, in effect, a solution with only one integration constant $\sigma = (\alpha + \beta\gamma)$.

A.1.1.2 Inhomogeneous Second-Order Linear ODEs

If $y_0(x)$ is *any* solution of the real inhomogeneous equation (A.1), i.e. if

$$\mathcal{L}y_0 \equiv y_0'' + p(x)y_0' + q(x)y_0 = f(x), \tag{A.7}$$

then the general solution of (A.1) has the form

$$y(x) = y_0(x) + \alpha y_1(x) + \beta y_2(x), \tag{A.8}$$

since

$$\begin{aligned}\mathcal{L} &= \mathcal{L}y_0 + \alpha\mathcal{L}y_1 + \beta\mathcal{L}y_2 \\ &= f + 0 + 0 = f.\end{aligned} \tag{A.9}$$

Here $y_1(x)$ and $y_2(x)$ are complementary functions, while $y_0(x)$ is referred to as a *particular solution*, or a *particular integral*.

A.1.1.3 The Wronskian

If y_1 and y_2 are linearly dependent (i.e. $y_2 = \gamma y_1$ for some γ), then so are y_1' and y_2' (since, from differentiating, $y_2' = \gamma y_1'$). Hence y1 and y2 are linearly dependent only if the equation

$$\begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0, \tag{A.10}$$

has a non-zero solution for α and β. Conversely, if this equation has a solution then y_1 and y_2 are linearly dependent. It follows that non-zero functions y_1 and y_2 are linearly independent if and only if

$$\begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \quad \Longrightarrow \quad \alpha = \beta = 0. \tag{A.11}$$

Define the Wronskian, $\mathbb{W}(x)$, of the two solutions to be the function

$$\mathbb{W}[y_1, y_2] = y_1 y_2' - y_2 y_1'. \tag{A.12}$$

Since $\mathrm{A}\mathbf{x} = \mathbf{0}$ only has a zero solution if and only if $\det \mathrm{A} \neq 0$, we conclude that y_1 and y_2 are linearly independent if and only if

$$\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1' = \mathbb{W} \neq 0, \tag{A.13}$$

i.e. the Wronskian is non-zero.

A.1.1.4 Initial-Value and Boundary-Value Problems

Two *boundary conditions* (BCs) must be specified to determine fully the solution of a second-order ODE. A boundary condition is usually an equation relating the values of y

and y' at one point. Without loss of generality we can assume that the BCs do not involve y'' and higher derivatives, since the ODE allows y'' and higher derivatives to be expressed in terms of y and y'. The general form of a *linear* BC at a point $x = a$ is

$$Ay(a) + By'(a) = E, \tag{A.14}$$

where A, B and E are constants, and A and B are not both zero. If $E = 0$ the BC is said to be homogeneous.

If both BCs are specified at the same point we have an *initial-value problem*, e.g. solve

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = F(t) \quad \text{for} \quad t \geq 0, \quad \text{subject to} \quad x = \frac{\mathrm{d}x}{\mathrm{d}t} = 0 \quad \text{at} \quad t = 0. \tag{A.15}$$

If the BCs are specified at different points we have a *two-point boundary-value problem*, e.g. solve

$$y''(x) + y(x) = f(x) \quad \text{for} \quad a \leq x \leq b, \quad \text{subject to} \quad y(a) = y(b) = 0. \tag{A.16}$$

A.1.2 Differential Equations Containing Delta Functions

If a differential equation involves a step function or delta function, this generally implies a lack of smoothness in the solution. The equation can be solved separately on either side of the discontinuity and the two parts of the solution connected by applying the appropriate matching conditions. Consider, as an example, the linear second-order ODE

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + y = \delta(x). \tag{A.17}$$

If x represents time, this equation could represent the behaviour of a simple harmonic oscillator in response to an impulsive force. In each of the regions $x < 0$ and $x > 0$ separately, the right-hand side vanishes and the general solution is a linear combination of $\cos x$ and $\sin x$. We may write

$$y = \begin{cases} \alpha_- \cos x + \beta_- \sin x, & x < 0 \\ \alpha_+ \cos x + \beta_+ \sin x, & x > 0 \end{cases}. \tag{A.18}$$

Since the general solution of a second-order ODE should contain only two arbitrary constants, it should be possible to relate α_+ and β_+ to α_- and β_-.

What is the nature of the non-smoothness in y? Integrate (A.17) from $x = -\epsilon$ to $x = \epsilon$ to obtain

$$\int_{-\epsilon}^{\epsilon} \frac{\mathrm{d}^2y}{\mathrm{d}x^2}\,\mathrm{d}x + \int_{-\epsilon}^{\epsilon} y(x)\,\mathrm{d}x = \int_{-\epsilon}^{\epsilon} \delta(x)\,\mathrm{d}x\,, \tag{A.19}$$

i.e.

$$y'(\epsilon) - y'(-\epsilon) + \int_{-\epsilon}^{\epsilon} y(x)\,\mathrm{d}x = 1. \tag{A.20}$$

Now let $\epsilon \to 0$. If we assume that y is bounded, then the integral term makes no contribution and we get

$$\left[\frac{\mathrm{d}y}{\mathrm{d}x}\right] \equiv \lim_{\epsilon\to 0}\left[\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{x=-\epsilon}^{x=\epsilon} = 1. \tag{A.21}$$

Since there is only a finite jump in the derivative of y, we may further conclude that y is continuous, in which case the jump conditions are

$$[y] = 0, \quad \left[\frac{\mathrm{d}y}{\mathrm{d}x}\right] = 1 \quad \text{at} \quad x = 0. \tag{A.22}$$

Applying these conditions, we obtain

$$\alpha_+ - \alpha_- = 0 \quad \text{and} \quad \beta_+ - \beta_- = 1. \tag{A.23}$$

Hence the general solution is

$$y = \begin{cases} \alpha_- \cos x + \beta_- \sin x, & x < 0 \\ \alpha_- \cos x + (\beta_- + 1) \sin x, & x > 0 \end{cases}. \tag{A.24}$$

In particular, if the oscillator is at rest before the impulse occurs, then $\alpha_- = \beta_- = 0$ and the solution is $y = H(x) \sin x$.

A.1.3 Green's Functions

A.1.3.1 The Green's Function for Two-Point Homogeneous Boundary-Value Problems

Suppose that we wish to solve (A.1), i.e

$$\mathcal{L}y(x) = f(x), \tag{A.25}$$

where $\mathcal{L}$ is the general second-order linear differential operator in x, i.e.

$$\mathcal{L} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + p(x)\frac{\mathrm{d}}{\mathrm{d}x} + q(x), \tag{A.26}$$

with p and q being continuous functions. To fix ideas we will assume that the solution should satisfy *homogeneous* boundary conditions at $x = a$ and $x = b$, i.e

$$\begin{aligned} Ay(a) + By'(a) &= 0, \\ Cy(b) + Dy'(b) &= 0. \end{aligned} \tag{A.27}$$

where A, B, C and D are constants such that A and B are not both zero, and C and D are not both zero.

Next, suppose that we can find a solution $G(x;\zeta)$ that is the response of the system to forcing at a point ζ, i.e. $G(x;\zeta)$ is the solution to

$$\mathcal{L}G(x;\zeta) = \delta(x - \zeta), \tag{A.28}$$

subject to the boundary conditions (A.27)

$$AG(a;\zeta) + BG_x(a;\zeta) = 0 \quad \text{and} \quad CG(b;\zeta) + DG_x(b;\zeta) = 0, \tag{A.29}$$

where

$$\mathcal{L} = \frac{\partial^2}{\partial x^2} + p(x)\frac{\partial}{\partial x} + q(x), \tag{A.30}$$

$$G_x(x;\zeta) = \frac{\partial G}{\partial x}, \tag{A.31}$$

and we have used $\partial/\partial x$ rather than $\mathrm{d}/\mathrm{d}x$ since G is a function of both x and ζ. Then we claim that the solution of the original problem (A.25) is

$$y(x) = \int_a^b G(x;\zeta) f(\zeta)\,\mathrm{d}\zeta\,. \tag{A.32}$$

To see this we first note that (A.32) satisfies the boundary conditions (A.27), since from (A.29)

$$Ay(a) + By'(a) = \int_a^b (AG(a;\zeta) + BG_x(a;\zeta)) f(\zeta)\,\mathrm{d}\zeta = 0, \tag{A.33}$$

$$Cy(b) + Dy'(b) = \int_a^b (CG(b;\zeta) + DG_x(b;\zeta)) f(\zeta)\,\mathrm{d}\zeta = 0. \tag{A.34}$$

Further, (A.32) also satisfies the inhomogeneous equation (A.25) because

$$\begin{aligned} \mathcal{L}y(x) &= \int_a^b \mathcal{L}G(x;\zeta) f(\zeta)\,\mathrm{d}\zeta \\ &= \int_a^b \delta(x-\zeta) f(\zeta)\,\mathrm{d}\zeta \\ &= f(x). \end{aligned} \tag{A.35}$$

The function $G(x;\zeta)$ is called the *Green's function* of $\mathcal{L}$ for the given homogeneous boundary conditions.

A.1.3.2 Two Properties of Green's Functions

In the next subsection we will construct a Green's function. However, first we need to derive two properties of $G(x;\zeta)$. Suppose that we integrate equation (A.28) from $\zeta - \epsilon$ to $\zeta + \epsilon$ for $\epsilon > 0$ and consider the limit $\epsilon \to 0$ (cf. (A.19)). The right hand side is equal to 1, since the definition of the generalised function $\delta(x)$ is such that for all nice[1] functions $f(x)$

$$\int_{-\infty}^{\infty} \delta(x-\xi) f(x)\,\mathrm{d}x = f(\xi), \tag{A.37}$$

[1]By *nice* we mean, for instance, that $f(x)$ is everywhere differentiable any number of times, and that

$$\int_{-\infty}^{\infty} \left|\frac{\mathrm{d}^n f}{\mathrm{d}x^n}\right|^2 \mathrm{d}x < \infty \quad \text{for all integers} \quad n \geq 0. \tag{A.36}$$

and hence

$$\begin{aligned}
1 &= \lim_{\epsilon\to 0}\int_{\zeta-\epsilon}^{\zeta+\epsilon} \mathcal{L}G\,\mathrm{d}x \\
&= \lim_{\epsilon\to 0}\int_{\zeta-\epsilon}^{\zeta+\epsilon} \left(\frac{\partial^2 G}{\partial x^2} + p\frac{\partial G}{\partial x} + qG\right)\mathrm{d}x \\
&= \lim_{\epsilon\to 0}\int_{\zeta-\epsilon}^{\zeta+\epsilon} \frac{\partial}{\partial x}\left(\frac{\partial G}{\partial x} + pG\right)\mathrm{d}x + \lim_{\epsilon\to 0}\int_{\zeta-\epsilon}^{\zeta+\epsilon}\left(-\frac{\mathrm{d}p}{\mathrm{d}x}G + qG\right)\mathrm{d}x \\
&= \lim_{\epsilon\to 0}\left[\frac{\partial G}{\partial x} + pG\right]_{x=\zeta-\epsilon}^{x=\zeta+\epsilon} - \lim_{\epsilon\to 0}\int_{\zeta-\epsilon}^{\zeta+\epsilon}\left(\frac{\mathrm{d}p}{\mathrm{d}x} - q\right)G\,\mathrm{d}x\,. \qquad \text{(A.38)}
\end{aligned}$$

How can this equation be satisfied? Taking the lead from (A.22), suppose that $G(x;\zeta)$ is bounded near $x=\zeta$, then since p and q are continuous, (A.38) reduces to

$$\lim_{\epsilon\to 0}\left[\frac{\partial G}{\partial x} + pG\right]_{x=\zeta-\epsilon}^{x=\zeta+\epsilon} = 1. \qquad \text{(A.39)}$$

This implies that the jump in the derivative of G is bounded (cf. the unit jump in the Heaviside step function $H(x)$ at $x=0$). In turn, this means that G must be continuous. We conclude that

$$\lim_{\epsilon\to 0}\Big[G(x;\zeta)\Big]_{x=\zeta-\epsilon}^{x=\zeta+\epsilon} = 0 \quad \text{and} \quad \lim_{\epsilon\to 0}\left[\frac{\partial G}{\partial x}\right]_{x=\zeta-\epsilon}^{x=\zeta+\epsilon} = 1. \qquad \text{(A.40)}$$

i.e. G is continuous and there is a unit jump in the derivative of G at $x=\zeta$. Note that a function can be continuous and its derivative discontinuous, but not vice versa.

A.1.3.3 Construction of the Green's Function

$G(x;\zeta)$ can be constructed by the following procedure. First we note that when $x\neq\zeta$, G satisfies the homogeneous equation, and hence G should be the sum of two linearly independent solutions, say y_1 and y_2, of the homogeneous equation. So let

$$G(x;\zeta) = \begin{cases} \alpha_-(\zeta)y_1(x) + \beta_-(\zeta)y_2(x) & \text{for } a\le x<\zeta \\ \alpha_+(\zeta)y_1(x) + \beta_+(\zeta)y_2(x) & \text{for } \zeta\le x\le b \end{cases}. \qquad \text{(A.41)}$$

By construction this satisfies (A.28) for $x\neq\zeta$. Next we obtain equations relating $\alpha_\pm(\zeta)$ and $\beta_\pm(\zeta)$ by requiring at $x=\zeta$ that G is continuous and $\frac{\partial G}{\partial x}$ has a unit discontinuity. It follows from (A.40) that

$$\begin{aligned}
[\alpha_+(\zeta)y_1(\zeta) + \beta_+(\zeta)y_2(\zeta)] - [\alpha_-(\zeta)y_1(\zeta) + \beta_-(\zeta)y_2(\zeta)] &= 0, \\
[\alpha_+(\zeta)y_1'(\zeta) + \beta_+(\zeta)y_2'(\zeta)] - [\alpha_-(\zeta)y_1'(\zeta) + \beta_-(\zeta)y_2'(\zeta)] &= 1,
\end{aligned} \qquad \text{(A.42)}$$

i.e., grouping the y_1 and y_2 terms,

$$\begin{aligned}
y_1(\zeta)[\alpha_+(\zeta) - \alpha_-(\zeta)] + y_2(\zeta)[\beta_+(\zeta) - \beta_-(\zeta)] &= 0, \\
y_1'(\zeta)[\alpha_+(\zeta) - \alpha_-(\zeta)] + y_2'(\zeta)[\beta_+(\zeta) - \beta_-(\zeta)] &= 1,
\end{aligned} \qquad \text{(A.43)}$$

i.e.,

$$\begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}\begin{pmatrix} \alpha_+ - \alpha_- \\ \beta_+ - \beta_- \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \qquad \text{(A.44)}$$

A solution exists to this equation if, see (A.13),

$$\mathbb{W} = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} \neq 0, \tag{A.45}$$

i.e. if y_1 and y_2 are linearly independent; if so then

$$\alpha_+ - \alpha_- = -\frac{y_2(\zeta)}{\mathbb{W}(\zeta)} \quad \text{and} \quad \beta_+ - \beta_- = \frac{y_1(\zeta)}{\mathbb{W}(\zeta)}. \tag{A.46}$$

Finally we impose the boundary conditions. For instance, suppose that the solution y is required to satisfy (cf. (A.15))

$$y(a) = y(b) = 0. \tag{A.47}$$

Then the appropriate boundary conditions for G would be

$$G(a;\zeta) = G(b;\zeta) = 0, \tag{A.48}$$

i.e. $A = C = 1$ and $B = D = 0$ in (A.29). It follows from (A.41) that we would require

$$\begin{aligned} \alpha_-(\zeta)y_1(a) + \beta_-(\zeta)y_2(a) &= 0, \\ \alpha_+(\zeta)y_1(b) + \beta_+(\zeta)y_2(b) &= 0. \end{aligned} \tag{A.49}$$

$\alpha_\pm$, $\beta_\pm$ could then be determined from the four equations in (A.46) and (A.49).

More generally, for the homogeneous boundary conditions (A.27), i.e.

$$Ay(a) + By'(a) = 0 \quad \text{and} \quad Cy(b) + Dy'(b) = 0, \tag{A.50}$$

the appropriate boundary conditions for G are

$$\begin{aligned} AG(a;\zeta) + BG_x(a;\zeta) &= 0 \\ CG(b;\zeta) + DG_x(b;\zeta) &= 0. \end{aligned} \tag{A.51}$$

For simplicity construct complementary functions y_1 and y_2 so that they satisfy the boundary condition at a and b respectively, i.e. choose y_1 and y_2 so that

$$Ay(a)_1 + By_1'(a) = 0 \quad \text{and} \quad Cyy_2(b) + Dyy_2'(b) = 0. \tag{A.52}$$

Then

$$\alpha_+ = \beta_- = 0, \tag{A.53}$$

and the solution (A.41) simplifies to

$$G(x;\zeta) = \begin{cases} \alpha_-(\zeta)y_1(x) & \text{for } a \leq x < \zeta \\ \beta_+(\zeta)y_2(x) & \text{for } \zeta \leq x \leq b \end{cases}, \tag{A.54}$$

and thence from (A.46)

$$\alpha_- = \frac{y_2(\zeta)}{\mathbb{W}(\zeta)} \quad \text{and} \quad \beta_+ = \frac{y_1(\zeta)}{\mathbb{W}(\zeta)}. \tag{A.55}$$

It follows from (A.41) that

$$G(x;\zeta) = \begin{cases} \dfrac{y_2(\zeta)y_1(x)}{\mathbb{W}(\zeta)} & \text{for } a \le x < \zeta \\ \dfrac{y_1(\zeta)y_2(x)}{\mathbb{W}(\zeta)} & \text{for } \zeta \le x \le b \end{cases}. \tag{A.56}$$

This method fails if the Wronskian $\mathbb{W}[y_1, y_2]$ vanishes. This happens if y_1 is proportional to y_2, i.e. if there is a complementary function that happens to satisfy the homogeneous boundary conditions both at $x = a$ and $x = b$. In this case the equation $\mathcal{L}y = f$ may not have a solution satisfying the boundary conditions; if it does, the solution will not be unique (cf. resonance).

A.1.3.4 The Green's Function for Homogeneous Initial-Value Problems

Suppose that instead of the two-point boundary conditions (A.27), we require that

$$y(a) = y'(a) = 0. \tag{A.57}$$

We then require, by analogy with (A.29), that

$$G(a;\zeta) = G_x(a;\zeta) = 0. \tag{A.58}$$

Choose the complementary functions so that $y_1(a) = 0$ and $y_2'(a) = 0$ (which can be shown to be linearly independent and always possible), then (A.41) simplifies to

$$G(x;\zeta) = \begin{cases} 0 & \text{for } a \le x < \zeta \\ \alpha_+(\zeta)y_1(x) + \beta_+(\zeta)y_2(x) & \text{for } \zeta \le x \le b \end{cases}, \tag{A.59}$$

i.e. $\alpha_- = \beta_- = 0$. The conditions that G be continuous and $\partial G/\partial x$ has a unit discontinuity then give that

$$\begin{aligned} \alpha_+(\zeta)y_1(\zeta) + \beta_+(\zeta)y_2(\zeta) &= 0, \\ \alpha_+(\zeta)y_1'(\zeta) + \beta_+(\zeta)y_2'(\zeta) &= 1. \end{aligned} \tag{A.60}$$

Or in matrix form

$$\begin{pmatrix} y_1(\zeta) & y_2(\zeta) \\ y_1'(\zeta) & y_2'(\zeta) \end{pmatrix} \begin{pmatrix} \alpha_+(\zeta) \\ \beta_+(\zeta) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{A.61}$$

with solution

$$\begin{pmatrix} \alpha_+(\zeta) \\ \beta_+(\zeta) \end{pmatrix} = \frac{1}{\mathbb{W}(\zeta)} \begin{pmatrix} y_2'(\zeta) & -y_2(\zeta) \\ -y_1'(\zeta) & y_1(\zeta) \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{\mathbb{W}(\zeta)} \begin{pmatrix} -y_2(\zeta) \\ y_1(\zeta) \end{pmatrix}. \tag{A.62}$$

The Green's function is therefore

$$G(x;\zeta) = \begin{cases} 0 & \text{for } a \le x < \zeta \\ \dfrac{y_1(\zeta)y_2(x) - y_1(x)y_2(\zeta)}{\mathbb{W}(\zeta)} & \text{for } \zeta \le x \le b \end{cases}. \tag{A.63}$$

A.1.3.5 Inhomogeneous Boundary Conditions

So far we only considered problems with homogeneous boundary conditions. One can also use Green's functions to solve problems with inhomogeneous boundary conditions. The trick is to solve the homogeneous equation $\mathcal{L}y_{\rm ibc} = 0$ for a function $y_{\rm ibc}$ which satisfies the *inhomogeneous* boundary conditions. Then solve the inhomogeneous equation $\mathcal{L}y_{\rm hbc} = f$, perhaps using the Green's function method discussed in this chapter, imposing *homogeneous* boundary conditions on $y_{\rm hbc}$. Then linearity means that $y_{\rm ibc} + y_{\rm hbc}$ satisfies the inhomogeneous equation with inhomogeneous boundary conditions.

A.2 Laplace and Poisson's Equations

Poisson's equation,

$$\nabla^2\Psi = \rho(\mathbf{x}), \tag{A.64}$$

is a second-order partial differential equation which arises in many different physical situations. (Watch out for different sign conventions here!) The *source term* $\rho(\mathbf{x})$ is often zero everywhere, or everywhere except in some specific regions or at some particular points. In this case we get Laplace's equation,

$$\nabla^2\Psi = 0. \tag{A.65}$$

A.2.1 Separation of Variables for Laplace's Equation

Recall that, because Laplace's equation $\nabla^2\Psi = 0$ is linear in Ψ, the superposition of any two (or more) solutions is another solution. The general solution can be written as a linear combination of some set of basis solutions - the number of solutions is infinite and so the space of solutions can be viewed as an infinite-dimensional vector space.

Separation of variables in some orthogonal system of coordinates provides a method to find a useful basis set of solutions. In Cartesian coordinates the Laplacian is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{A.66}$$

You have previously used separation of variables in Cartesian coordinates to solve Laplace's equation, considering solutions of the factorised form,

$$\Psi(x, y, z) = X(x)Y(y)Z(z). \tag{A.67}$$

The general solution can then be written as a linear superposition of these solutions.

For any given problem, choosing a basis set appropriately, according to the symmetry of the problem, can often lead to the solution in a simpler form, e.g. only a few of the basis set may be needed. This involves choosing an appropriate coordinate system. For example, for a spherically-symmetric source in infinite space we expect spherical polar

coordinates to be most useful, whereas for the flow of air around a very long cylinder, cylindrical polar coordinates may be more appropriate.

We'll consider plane polar coordinates (equivalent to cylindrical polars with no z-dependence) and spherical polar coordinates with cylindrical symmetry

A.2.1.1 Plane Polar Coordinates

Recalling from last term the expression for ∇^2 in plane polar coordinates (r, ϕ), where $x = r\cos\phi$ and $y = r\sin\phi$, acting on a scalar field $\Psi(r, \phi)$, Laplace's equation is,

$$\nabla^2\Psi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Psi}{\partial\phi^2} = 0. \tag{A.68}$$

The same equation arises in cylindrical polar coordinates (r, ϕ, z) when $\partial\Psi/\partial z = 0$.

If we consider separable solutions of the form $\Psi(r, \phi) = R(r)\Phi(\phi)$ then,

$$\frac{\Phi}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}R}{\mathrm{d}r}\right) + \frac{R}{r^2}\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi^2} = 0. \tag{A.69}$$

Rearranging we get,

$$\frac{r}{R}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}R}{\mathrm{d}r}\right) = -\frac{1}{\Phi}\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi^2}. \tag{A.70}$$

The LHS is a function of r only and the RHS of ϕ only, and so both must be a constant λ. The equation for $\Phi(\phi)$ gives,

$$\Phi'' = -\lambda\Phi, \tag{A.71}$$

and so

$$\Phi = \begin{cases} A + B\phi & \lambda = 0 \\ A\cos\sqrt{\lambda}\phi + B\sin\sqrt{\lambda}\phi & \lambda \neq 0 \end{cases}. \tag{A.72}$$

In many cases Ψ corresponds to some physical quantity (e.g. the concentration of solute or the temperature) and must be periodic: $\Psi(r, \phi) = \Psi(r, \phi + 2\pi)$. However, in other situations (e.g. electrostatics) Ψ is not a physical quantity just a potential and $\boldsymbol{\nabla}\Psi$ must be periodic. We will allow for the more general case by requiring $\Phi'(\phi) = \Phi'(\phi + 2\pi)$. This gives

$$2\pi\sqrt{\lambda} = 2\pi n \quad \Longrightarrow \quad \lambda = n^2, \quad n \in \mathbb{Z}. \tag{A.73}$$

Hence

$$\Phi = \begin{cases} A + B\phi & n = 0 \\ A\cos n\phi + B\sin n\phi & n \neq 0 \end{cases}. \tag{A.74}$$

Returning to the equation for $R(r)$,

$$\frac{r}{R}\frac{\mathrm{d}}{\mathrm{d}r}\left(r\frac{\mathrm{d}R}{\mathrm{d}r}\right) = n^2 \quad \Longrightarrow \quad r^2R'' + rR' - n^2R = 0. \tag{A.75}$$

By direct verification (or making a substitution $u = \ln r$) it can be shown that the solution of this second-order ODE is,

$$R = \begin{cases} C + D\ln r & n = 0 \\ Cr^n + Dr^{-n} & n \neq 0 \end{cases}. \tag{A.76}$$

Combining R and Φ we get,

$$\Psi = R\Phi = \begin{cases} (C + D\ln r)(A + B\phi) & n = 0 \\ (Cr^n + Dr^{-n})(A\cos n\phi + B\sin n\phi) & n \neq 0 \end{cases} \tag{A.77}$$

We exclude the $\phi \ln r$ combination because it does not satisfy the periodicity requirement on $\boldsymbol{\nabla}\Psi$. The general solution is therefore (relabelling the arbitrary constants),

$$\boxed{\Psi = A_0 + B_0\phi + C_0 \ln r + \sum_{n=1}^{\infty}(A_n r^n + C_n r^{-n})\cos n\phi + \sum_{n=1}^{\infty}(B_n r^n + D_n r^{-n})\sin n\phi.} \tag{A.78}$$

With some further relabelling this can be rewritten more compactly as

$$\Psi = A_0 + B_0\phi + C_0 \ln r + \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} r^n (A_n \cos n\phi + B_n \sin n\phi). \tag{A.79}$$

A.2.1.2 Spherical Polar Coordinates (Axisymmetric Case)

In spherical polars (r, θ, ϕ), where $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$ and $z = r\cos\theta$. When $\Psi(r, \theta, \phi)$ is axisymmetric (i.e. independent of ϕ, $\partial\Psi/\partial\phi = 0$), Laplace's equation is,

$$\nabla^2\Psi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Psi}{\partial\theta}\right) = 0. \tag{A.80}$$

Looking for separable solutions of the form $\Psi(r, \theta) = R(r)T(\theta)$ then,

$$\frac{1}{R}\left(r^2R'\right)' = -\frac{1}{T\sin\theta}(T'\sin\theta)'. \tag{A.81}$$

The LHS is a function of r only and the RHS of θ only, and so both must be a constant λ.

The equation for $T(\theta)$ gives

$$(T'\sin\theta)' = -\lambda T\sin\theta. \tag{A.82}$$

Setting $u = \cos\theta$, we have $\mathrm{d}/\mathrm{d}\theta = -\sin\theta\,\mathrm{d}/\mathrm{d}u$ and so,

$$-\sin\theta\frac{\mathrm{d}}{\mathrm{d}u}\left(-\sin^2\theta\frac{\mathrm{d}T}{\mathrm{d}u}\right) = -\lambda T\sin\theta, \tag{A.83}$$

giving

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\left(1-u^2\right)\frac{\mathrm{d}T}{\mathrm{d}u}\right) + \lambda T = 0 \tag{A.84}$$

We have already encountered this equation a number of times: for well-behaved solutions at $u = \pm 1\,(\theta = 0, \pi)$ we require $\lambda = \ell(\ell+1)$ where $\ell = 0, 1, 2, \cdots$ (the series terminates if we look for a series solution). The resulting polynomials, Legendre polynomials $P_\ell(u)$, are normalised so that $P_\ell(1) = 1$.

Returning to the equation for $R(r)$,

$$\left(r^2R'\right)' = \lambda R \quad \Longrightarrow \quad r^2R'' + 2rR' - \ell(\ell+1)R = 0. \tag{A.85}$$

This has solution (find by trying $R = r^k$ or using substitution $v = \ln r$),

$$R = Ar^\ell + Br^{-\ell-1}. \tag{A.86}$$

The general solution to Laplace's equation in spherical polar coordinates in the axisymmetric case is therefore (redefining arbitrary constants),

$$\boxed{\Psi(r,\theta) = \sum_{\ell=0}^{\infty} \left(A_\ell r^\ell + B_\ell r^{-\ell-1}\right) P_\ell(\cos\theta).} \tag{A.87}$$

In the non-axisymmetric case, a similar analysis would give an extra equation involving ϕ and the Legendre polynomials would be replaced by associated Legendre polynomials (solutions of the associated Legendre equation)

A.2.2 The Green's Function and the Fundamental Solution

Suppose we are solving Poisson's equation with Dirichlet boundary conditions. Recall from Appendix A.1 that the *Green's function* associated with the problem, $G(\mathbf{r}, \mathbf{r}')$, satisfies,

$$\nabla^2_{\mathbf{r}} G(\mathbf{r}, \mathbf{r}') = \delta^{(3)}(\mathbf{r} - \mathbf{r}') \qquad \mathbf{r} \text{ in } \mathcal{V}. \tag{A.88}$$

$$G = 0 \qquad \mathbf{r} \text{ on } \mathcal{S}. \tag{A.89}$$

Here $\delta^{(3)}(\mathbf{r} - \mathbf{r}') = \delta(x - x')\delta(y - y')\delta(z - z')$ is the 3D Dirac delta function satisfying,

$$\int_{\mathcal{V}} f(\mathbf{r})\delta^{(3)}(\mathbf{r} - \mathbf{r}')\,\mathrm{d}V = f(\mathbf{r}'), \tag{A.90}$$

if $\mathbf{r}'$ is in $\mathcal{V}$ and otherwise the integral is zero.

If $\mathcal{V}$ is all of space (the limit of a sphere with radius $\to \infty$), the Green's function is known as the *fundamental solution.*

It's possible to prove that a real Green's function is symmetric, $G(\mathbf{r}, \mathbf{r}') = G(\mathbf{r}', \mathbf{r})$. We saw this for a 1D Green's function written as a sum over eigenfunctions and can check it is true in the examples we consider. We can think of G as the potential due to a point charge at $\mathbf{r}'$; the potential at $\mathbf{r}$ due to a source at $\mathbf{r}'$ is the same as the potential at $\mathbf{r}'$ due to a source at $\mathbf{r}$.

If we are solving Poisson's equation with Neumann boundary conditions, instead of $G = 0$ on $\mathcal{S}$ we require,

$$\frac{\partial G}{\partial n} = \frac{1}{A} \quad \text{on } \mathcal{S}, \tag{A.91}$$

where $A = \oint_{\mathcal{S}} \mathrm{d}S$ is the surface area. When $A \to \infty$ the condition becomes $\partial G/\partial n = 0$ on $\mathcal{S}$.

A.2.2.1 The Fundamental Solution in 3D

We first consider $\mathbf{r}' = \mathbf{0}$ (a point source at the origin): $\nabla^2 G = \delta^{(3)}(\mathbf{r})$ and $G \to 0$ as $|\mathbf{r}| \to \infty$. The problem has spherical symmetry and so we'll assume that G is a function of r only - we know any solution we find is *the* unique solution.

Using the expression for ∇^2 in spherical polars,

$$\left(r^2 G'\right)' = 0 \quad \Longrightarrow \quad G = \frac{C}{r} + A \quad (r \neq 0), \tag{A.92}$$

for some constants A, C; we require $A = 0$ from the boundary condition at infinity. To determine C we integrate $\nabla^2 G$ over a sphere of radius ϵ centred on the origin. Using the divergence theorem

$$\begin{aligned}\int_{r<\epsilon} \nabla^2 G \,\mathrm{d}V &= \oint_{r=\epsilon} \boldsymbol{\nabla} G \cdot \mathrm{d}\mathbf{S} = \oint_{r=\epsilon} \boldsymbol{\nabla} G \cdot \hat{\mathbf{n}} \,\mathrm{d}S \\ &= \oint_{r=\epsilon} \frac{\partial G}{\partial r} \,\mathrm{d}S \\ &= -\frac{C}{\epsilon^2} \oint_{r=\epsilon} \mathrm{d}S \\ &= -4\pi C,\end{aligned} \tag{A.93}$$

noting that we can take $\partial G/\partial r$ outside the integral as it is constant over the surface $r = \epsilon$.

Because we can take ϵ as small as we like, we deduce that

$$\nabla^2 G = -4\pi C \delta^{(3)}(\mathbf{r}), \tag{A.94}$$

and so $C = -1/4\pi$ giving $G = -1/4\pi|\mathbf{r}|$. Shifting the origin to $\mathbf{r}'$ we obtain,

$$\boxed{G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi|\mathbf{r} - \mathbf{r}'|}.} \tag{A.95}$$

A.2.2.2 The Fundamental Solution in 2D

Again, we consider $\mathbf{r}' = 0$: $\nabla^2 G = \delta^{(2)}(\mathbf{r})$. We will find that $G \to 0$ as $|\mathbf{r}| \to \infty$ is impossible; instead we must require that e.g. G vanishes on some circle of radius R or $|\boldsymbol{\nabla} G| \to 0$ as $|\mathbf{r}| \to \infty$. The problem has circular symmetry and so we assume that G is a function of r only.

Using the expression for ∇^2 in circular polars,

$$\left(rG'\right)' = 0 \quad \Longrightarrow \quad G = C \ln r + A \quad (r \neq 0), \tag{A.96}$$

for some constants A, C. Integrating $\nabla^2 G$ over a circle of radius ϵ centred on the origin

and using the divergence theorem (2D version),

$$\begin{aligned}\int_{r<\epsilon} \nabla^2 G \,\mathrm{d}A &= \oint_{r=\epsilon} \boldsymbol{\nabla} G \cdot \mathrm{d}\mathbf{l} = \oint_{r=\epsilon} \boldsymbol{\nabla} G \cdot \hat{\mathbf{n}} \,\mathrm{d}l \\ &= \oint_{r=\epsilon} \frac{\partial G}{\partial r} \,\mathrm{d}l \\ &= \frac{C}{\epsilon} \oint_{r=\epsilon} \mathrm{d}l \\ &= 2\pi C. \end{aligned} \tag{A.97}$$

Because we can take ϵ as small as we like, we deduce that

$$\nabla^2 G = 2\pi C \delta^{(2)}(\mathbf{r}), \tag{A.98}$$

and so $C = 1/2\pi$ giving $G = \frac{1}{2\pi} \ln|\mathbf{r}| + A$. (Note that $G'(r) \to 0$ as $r \to \infty$ but $G(r) \to \infty$). Shifting the origin to $\mathbf{r}'$ we obtain,

$$\boxed{G(\mathbf{r}, \mathbf{r}') = \frac{1}{2\pi} \ln|\mathbf{r} - \mathbf{r}'| + \text{const.}} \tag{A.99}$$

A.2.3 The Method of Images

So far we have found G when $\mathcal{V}$ is all of space, the fundamental solution. We can use the *method of images* to find G in some other simple geometries. If we find a solution that satisfies Poisson's equation and the boundary conditions, from the uniqueness of solutions it must be *the* solution (up to a constant if we have purely Neumann boundary conditions). We'll see how this works with a number of examples.

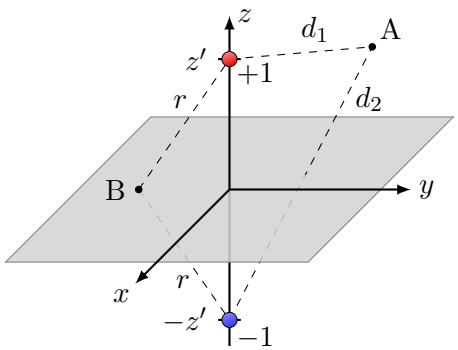


Fig. A.1: Image source location, at the image point $\mathbf{r}'' = (x', y', -z')$, for the fundamental Green's function in the half-space of $\mathbb{R}^3$ with $z > 0$. At point B at $z = 0$, the boundary conditions are satisfied. We have that $d_1 = |\mathbf{r} - \mathbf{r}'|$ and $d_2 = |\mathbf{r} - \mathbf{r}''|$.

3D Half-Space What is the Green's function for a domain $\mathbb{D}$ with *Dirichlet boundary conditions*, where $\mathbb{D}$ is the half-space of $\mathbb{R}^3$ with $z > 0$? The Green's function satisfies,

$$\nabla^2 G = \delta^{(3)}(\mathbf{r} - \mathbf{r}') \qquad \mathbf{r} \in \mathbb{D}, \tag{A.100}$$

$$G = 0 \qquad \text{on } z = 0, \tag{A.101}$$

$$G \to 0 \qquad \text{as } |\mathbf{r}| \to \infty, \mathbf{r} \in \mathbb{D}. \tag{A.102}$$

Uniqueness of solutions allows us to solve this using a trick: remove the boundary at $z = 0$, consider all of space and add a point source of *opposite sign*, an *image source*, at the image point $\mathbf{r}'' = (x', y', -z')$, as shown in Fig. A.1

The Green's function satisfies,

$$\nabla^2 G = \delta^{(3)}(\mathbf{r} - \mathbf{r}') - \delta^{(3)}(\mathbf{r} - \mathbf{r}'') \tag{A.103}$$

which, by superposition of fundamental solutions, gives,

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi|\mathbf{r} - \mathbf{r}'|} + \frac{1}{4\pi|\mathbf{r} - \mathbf{r}''|}. \tag{A.104}$$

This satisfies the boundary condition at $z = 0$ because when $z = 0$, $|\mathbf{r} - \mathbf{r}'| = \sqrt{(x - x')^2 + (y - y')^2 + z'^2} = |\mathbf{r} - \mathbf{r}''|$, and it satisfies the other two requirements when $\mathbf{r} \in \mathbb{D}$. Therefore, by uniqueness,

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \frac{1}{|\mathbf{r} - \mathbf{r}''|}\right) \tag{A.105}$$

If instead we impose *Neumann boundary conditions* at $z = 0$, i.e. $\partial G/\partial n = -\partial G/\partial z = 0$ at $z = 0$, but still require $G \to 0$ as $|\mathbf{r}| \to \infty$, $\mathbf{r} \in \mathbb{D}$, then we need a point charge of the *same sign* at the image point and the Green's function is,

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{1}{|\mathbf{r} - \mathbf{r}''|}\right) \tag{A.106}$$

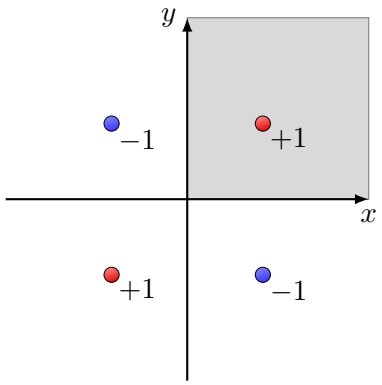


Fig. A.2: Image source locations for a domain $\mathbb{D}$, the quarter plane of $\mathbb{R}^2$ with $x > 0$, $y > 0$, with Dirichlet boundary conditions. The domain $\mathbb{D}$ is shaded in grey.

2D Quarter-Plane What is the Green's function $G(\mathbf{r}, \mathbf{r}_0)$ for a domain $\mathbb{D}$, the quarter plane of $\mathbb{R}^2$ with $x > 0$, $y > 0$, with Dirichlet boundary conditions? We require $G = 0$ on $x = 0$ and on $y = 0$, and $\nabla^2 G = \delta(\mathbf{r} - \mathbf{r}_0)$ for $r \in \mathbb{D}$. It turns out we need three image charges as shown in Fig. A.2, $\mathbf{r}_1$ and $\mathbf{r}_2$ with strength -1 and $\mathbf{r}_3$ with strength $+1$. This gives,

$$\begin{aligned} G(\mathbf{r}, \mathbf{r}_0) &= +\frac{1}{2\pi}\ln|\mathbf{r} - \mathbf{r}_0| - \frac{1}{2\pi}\ln|\mathbf{r} - \mathbf{r}_1| - \frac{1}{2\pi}\ln|\mathbf{r} - \mathbf{r}_2| + \frac{1}{2\pi}\ln|\mathbf{r} - \mathbf{r}_3| + C \\ &= \frac{1}{2\pi}\ln\frac{|\mathbf{r} - \mathbf{r}_0||\mathbf{r} - \mathbf{r}_3|}{|\mathbf{r} - \mathbf{r}_1||\mathbf{r} - \mathbf{r}_2|}, \end{aligned} \tag{A.107}$$

where the constant C is zero from the boundary conditions at $x = 0$ and $y = 0$.

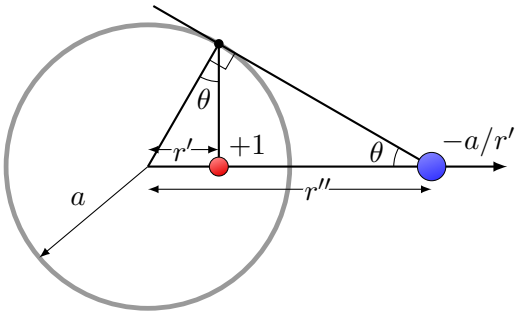


Fig. A.3: Image source locations for a domain $\mathbb{D}$, the sphere in $\mathbb{R}^3$ with $r < a$, with Dirichlet boundary conditions. For a source charge at r', the image source has strength $-a/r'$ at a distance $r'' = a^2/r'$.

Images in a Sphere What is the Green's function (Dirichlet boundary conditions) for a domain $\mathbb{D}$ which is $r < a$ in $\mathbb{R}^3$? The Green's function satisfies,

$$\nabla^2 G = \delta^{(3)}(\mathbf{r} - \mathbf{r}') \qquad r < a, \tag{A.108}$$

$$G = 0 \qquad r = a. \tag{A.109}$$

The image source has strength $-a/r'$ and the image point is the inverse point, it follows directly from the illustrated geometry in Fig. A.3 that

$$\cos\theta = \frac{a}{r''} = \frac{r'}{a} \quad \Longrightarrow \quad \mathbf{r}'' = \frac{a^2}{r'^2}\mathbf{r}'. \tag{A.110}$$

The Green's function is therefore,

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \frac{a/r'}{|\mathbf{r} - \mathbf{r}''|}\right). \tag{A.111}$$

This satisfies the boundary condition at $|\mathbf{r}| = a$, and the same result holds if the domain is instead $r > a$.

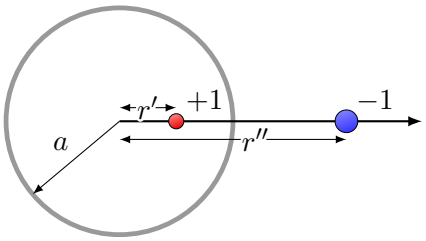


Fig. A.4: Image source locations for a domain $\mathbb{D}$, the circle in $\mathbb{R}^2$ with $r < a$, with Dirichlet boundary conditions. For a source charge at r', the image source has strength -1 at a distance $r'' = a^2/r'$.

Images in a Circle What is the Green's function (Dirichlet boundary conditions) for a domain $\mathbb{D}$ which is $r < a$ in $\mathbb{R}^2$? This is the 2D equivalent of the above. The image point is the inverse point again, $\mathbf{r}'' = a^2/r'^2\mathbf{r}'$, but now the image source just has strength -1. The Green's function is,

$$G(\mathbf{r}, \mathbf{r}') = \frac{1}{2\pi} \ln \frac{|\mathbf{r} - \mathbf{r}'|}{|\mathbf{r} - \mathbf{r}''|} + C, \tag{A.112}$$

where the constant C is chosen to ensure that $G = 0$ on the circle $r = a$.

A.2.4 The Integral Solution of Poisson's Equation

So far we've found solutions of Poisson's equation with point sources - Green's functions can be used to find the solutions with an arbitrary source distribution. We'll need *Green's identity*: for any smooth functions Φ and Ψ defined in a volume $\mathcal{V}$ with surface $\mathcal{S}$,

$$\boxed{\int_{\mathcal{V}} \left(\Phi\nabla^2\Psi - \Psi\nabla^2\Phi\right) \mathrm{d}V = \oint_{\mathcal{S}} (\Phi\boldsymbol{\nabla}\Psi - \Psi\boldsymbol{\nabla}\Phi)\cdot\hat{\mathbf{n}}\,\mathrm{d}S = \oint_{\mathcal{S}} \left(\Phi\frac{\partial\Psi}{\partial n} - \Psi\frac{\partial\Phi}{\partial n}\right)\mathrm{d}S\,,} \tag{A.113}$$

which can easily be shown using the divergence theorem on $\mathbf{F} \equiv \Phi\boldsymbol{\nabla}\Psi - \Psi\boldsymbol{\nabla}\Phi$ and the vector identity $\boldsymbol{\nabla}\cdot(\Phi\boldsymbol{\nabla}\Psi) = \boldsymbol{\nabla}\Phi\cdot\boldsymbol{\nabla}\Psi + \Phi\nabla^2\Psi$.

There's also a 2D version for a plane surface $\mathcal{S}$ bounded by a curve $\mathcal{C}$,

$$\int_{\mathcal{S}} \left(\Phi\nabla^2\Psi - \Psi\nabla^2\Phi\right)\mathrm{d}A = \oint_{\mathcal{C}} \left(\Phi\frac{\partial\Psi}{\partial n} - \Psi\frac{\partial\Phi}{\partial n}\right)\mathrm{d}l\,. \tag{A.114}$$

Consider Poisson's equation with Dirichlet boundary conditions: $\nabla^2\Phi = \rho(\mathbf{r})$ in $\mathcal{V}$ and $\Phi(\mathbf{r}) = f(\mathbf{r})$ on $\mathcal{S}$. Applying Green's identity with $\Psi = G$

$$\begin{aligned}\int_{\mathcal{V}} \left(\Phi\nabla^2 G - G\nabla^2\Phi\right)\mathrm{d}V &= \oint_{\mathcal{S}} (\Phi\boldsymbol{\nabla}G - G\boldsymbol{\nabla}\Phi)\cdot\hat{\mathbf{n}}\,\mathrm{d}S \\ \implies \int_{\mathcal{V}} \Phi^{(3)}(\mathbf{r}-\mathbf{r}')\,\mathrm{d}V &= \int_{\mathcal{V}} G\rho(\mathbf{r})\,\mathrm{d}V + \oint_{\mathcal{S}} f\frac{\partial G}{\partial n}\,\mathrm{d}S\,,\end{aligned} \tag{A.115}$$

and hence we arrive at

$$\boxed{\Phi(\mathbf{r}') = \int_{\mathcal{V}} \rho(\mathbf{r})G(\mathbf{r},\mathbf{r}')\,\mathrm{d}V + \oint_{\mathcal{S}} f(\mathbf{r})\frac{\partial G}{\partial n}\,\mathrm{d}S\,,} \tag{A.116}$$

the *integral solution of Poisson's equation* with Dirichlet boundary conditions. This expression can also be used to solve Laplace's equation by setting $\rho(\mathbf{r}) = 0$.

If we want $\mathcal{V}$ to be all space, we can use the fundamental solution for G but we need to ensure that the surface integral $\to 0$. (Consider a sphere of radius R and the limit $R \to \infty$.) In this case,

$$\Phi(\mathbf{r}') = \int_{\mathbb{R}^3} \rho(\mathbf{r})G(\mathbf{r},\mathbf{r}')\,\mathrm{d}V\,. \tag{A.117}$$

For example, consider a charge distribution $\rho_{\mathrm{q}}(\mathbf{r})$ that decays rapidly far from the origin. Then we have,

$$\Phi(\mathbf{r}') = \int_{\mathbb{R}^3} \frac{\rho_{\mathrm{q}}(\mathbf{r})}{4\pi\epsilon_0|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}V\,. \tag{A.118}$$

This can be understood as the superposition of many infinitesimal charge elements.

An integral solution of Poisson's equation can also be derived for Neumann boundary conditions. ($\partial\Phi/\partial n = f(\mathbf{r})$ on $\mathcal{S}$ and $\partial G/\partial n = 1/A$ on $\mathcal{S}$ where A is the surface area.) From Green's identity we have,

$$\Phi(\mathbf{r}') = \int_{\mathcal{V}} \rho(\mathbf{r})G(\mathbf{r},\mathbf{r}')\,\mathrm{d}V + \frac{1}{A}\oint_{\mathcal{S}} \Phi(\mathbf{r})\,\mathrm{d}S - \oint_{\mathcal{S}} f(\mathbf{r})G(\mathbf{r},\mathbf{r}')\,\mathrm{d}S\,. \tag{A.119}$$

If $\mathcal{V}$ is all of space, $A \to \infty$ and then, as long as the surface integral over Φ is finite,

$$\Phi(\mathbf{r}') = \int_{\mathcal{V}} \rho(\mathbf{r}) G(\mathbf{r}, \mathbf{r}') \, \mathrm{d}V - \oint_{\mathcal{S}} f(\mathbf{r}) G(\mathbf{r}, \mathbf{r}') \, \mathrm{d}S \,. \tag{A.120}$$

A.3 Uniqueness of Solutions of Poisson's Equation

To get a unique solution of Poisson's equation, $\nabla^2 \Phi = \rho(\mathbf{r})$, in a volume $\mathcal{V}$ with boundary surface $\mathcal{S}$ (possibly comprised of a number of connected surfaces $\mathcal{S}_i$), we need to impose boundary conditions.

- Dirichlet Boundary Conditions: The value of Φ is fixed on a given surface S_i.
- Neumann Boundary Conditions: The value of $\boldsymbol{\nabla}\Phi \cdot \hat{\mathbf{n}}$ is fixed perpendicular to a given surface $\mathcal{S}_i$.

Notice that, for each $\mathcal{S}_i$, we need to decide which of the two boundary conditions we want. We don't get to chose both of them. We then have that with either Dirichlet or Neumann boundary conditions chosen on each surface $\mathcal{S}_i$, the Laplace equation has a unique solution.

Suppose that there are two solutions, $\Phi_1(\mathbf{r})$ and $\Phi_2(\mathbf{r})$ with the same specified boundary conditions. Let's define $\Psi \equiv \Phi_1 - \Phi_2$. We can see that

$$\nabla^2 \Psi = \nabla^2 \Phi_1 - \nabla^2 \Phi_2 = \rho - \rho = 0 \quad \text{in } \mathcal{V}, \tag{A.121}$$

so $\nabla^2 \Psi$ vanishes by the Laplace equation.

Now consider evaluating the identity

$$\begin{aligned} \boldsymbol{\nabla} \cdot (\Psi \boldsymbol{\nabla} \Psi) &= \boldsymbol{\nabla}\Psi \cdot \boldsymbol{\nabla}\Psi + \Psi \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla}\Psi) \\ &= |\boldsymbol{\nabla}\Psi|^2 + \Psi \nabla^2 \Psi. \end{aligned} \tag{A.122}$$

We can now look at the following expression

$$\int_{\mathcal{V}} \mathrm{d}^3 r \, \boldsymbol{\nabla} \cdot (\Psi \boldsymbol{\nabla} \Psi) = \int_{\mathcal{V}} \mathrm{d}^3 r \, |\boldsymbol{\nabla}\Psi|^2 + \Psi \nabla^2 \Psi, \tag{A.123}$$

where we know from (A.121) that the $\nabla^2 \Psi$ must vanish. By the divergence theorem, we also know that

$$\int_{\mathcal{V}} \mathrm{d}^3 r \, \boldsymbol{\nabla} \cdot (\Psi \boldsymbol{\nabla} \Psi) = \sum_i \int_{\mathcal{S}_i} \Psi \boldsymbol{\nabla}\Psi \cdot \mathrm{d}\mathbf{S} \,. \tag{A.124}$$

However, if we've picked Dirichlet boundary conditions then $\Psi = 0$ on the boundary, while Neumann boundary conditions ensure that $\boldsymbol{\nabla}\Psi = 0$ on the boundary. This means that the integral vanishes and, from (A.123), we must have $\boldsymbol{\nabla}\Psi = 0$ throughout space. But if we have imposed Dirichlet boundary conditions somewhere, then $\Psi = 0$ on that boundary and so $\Psi = 0$ everywhere. Alternatively, if we have Neumann boundary conditions on all surfaces than $\boldsymbol{\nabla}\Psi = 0$ everywhere and the two solutions Φ_1 and Φ_2 can differ only by a constant. But, as discussed in Section 2.2, this constant has no physical meaning.