Lent Term, 2015

Electromagnetism

University of Cambridge Part IB and Part II Mathematical Tripos

---

David Tong

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OBA, UK

http://www.damtp.cam.ac.uk/user/tong/em.html
d.tong@damtp.cam.ac.uk

<u>Maxwell Equations</u>

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)$$

Recommended Books and Resources

There is more or less a well established route to teaching electromagnetism. A number of good books follow this.

- David J. Griffiths, "*Introduction to Electrodynamics*"

A superb book. The explanations are clear and simple. It doesn't cover quite as much as we'll need for these lectures, but if you're looking for a book to cover the basics then this is the first one to look at.

- Edward M. Purcell and David J. Morin "*Electricity and Magnetism*"

Another excellent book to start with. It has somewhat more detail in places than Griffiths, but the beginning of the book explains both electromagnetism and vector calculus in an intertwined fashion. If you need some help with vector calculus basics, this would be a good place to turn. If not, you'll need to spend some time disentangling the two topics.

- J. David Jackson, "*Classical Electrodynamics*"

The most canonical of physics textbooks. This is probably the one book you can find on every professional physicist's shelf, whether string theorist or biophysicist. It will see you through this course and next year's course. The problems are famously hard. But it does have div, grad and curl in polar coordinates on the inside cover.

- A. Zangwill, "*Modern Electrodynamics*"

A great book. It is essentially a more modern and more friendly version of Jackson. Although, embarrassingly, Maxwell's equations on the inside cover have a typo.

- Feynman, Leighton and Sands, "*The Feynman Lectures on Physics, Volume II*"

Feynman's famous lectures on physics are something of a mixed bag. Some explanations are wonderfully original, but others can be a little too slick to be helpful. And much of the material comes across as old-fashioned. Volume two covers electromagnetism and, in my opinion, is the best of the three.

A number of excellent lecture notes, including the Feynman lectures, are available on the web. Links can be found on the course webpage:
http://www.damtp.cam.ac.uk/user/tong/em.html

Contents

Acknowledgements

These lecture notes contain material covering two courses on Electromagnetism. In Cambridge, these courses are called Part IB Electromagnetism and Part II Electrodynamics. The notes owe a debt to the previous lecturers of these courses, including Natasha Berloff, John Papaloizou and especially Anthony Challinor.

The notes assume a familiarity with Newtonian mechanics and special relativity, as covered in the *Dynamics and Relativity* notes. They also assume a knowledge of vector calculus. The notes do not cover the classical field theory (Lagrangian and Hamiltonian) section of the Part II course.

1. Introduction

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.

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, $q \in \mathbf{R}$. 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 3.5). 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

$$e = 1.602176634 \times 10^{-19}\ C$$

A much more natural unit would be to simply count charge as $q = ne$ with $n \in \mathbf{Z}$. 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.

(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).

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*,

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

defined as charge per unit volume. The total charge Q in a given region V is simply $Q = \int_V d^3x\ \rho(\mathbf{x}, t)$. 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* $\mathbf{J}(\mathbf{x}, t)$, defined as follows: for every surface S, the integral

$$I = \int_S \mathbf{J} \cdot d\mathbf{S}$$

counts the charge per unit time passing through S. (Here $d\mathbf{S}$ is the unit normal to 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

$$\mathbf{J} = \rho \mathbf{v}$$

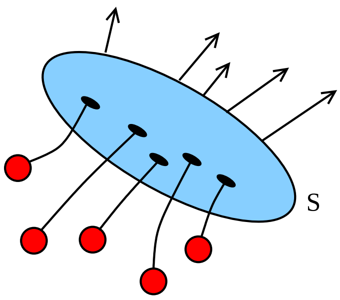


Figure 1: Current flux

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 the figure, where the underlying charged particles are shown as red balls, moving through the blue surface S.

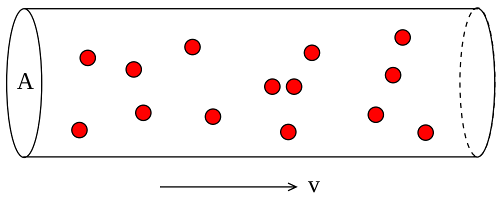


Figure 2: The wire

As a simple example, consider electrons moving along a wire. 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*,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{1.1}$$

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 V.

$$\frac{dQ}{dt} = \int_V d^3x \, \frac{\partial \rho}{\partial t} = -\int_V d^3x \, \nabla \cdot \mathbf{J} = -\int_S \mathbf{J} \cdot d\mathbf{S}$$

From our previous discussion, $\int_S \mathbf{J} \cdot d\mathbf{S}$ is the total current flowing out through the boundary S of the region V. (It is the total charge flowing *out*, rather than in, because $d\mathbf{S}$ is the outward normal to the region 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 $dQ/dt = 0$. This is the statement of (global) conservation of charge. In many applications we will take V to be all of space, $\mathbf{R}^3$, 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* $\mathbf{B}$,

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

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,

$$\mathbf{F} = q(\mathbf{E} + \dot{\mathbf{r}} \times \mathbf{B}) \tag{1.2}$$

The motion of the particle can then be determined through Newton's equation $\mathbf{F} = m\ddot{\mathbf{r}}$. We explored various solutions to this in the *Dynamics and Relativity* course. 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

$$\mathbf{f} = \rho\mathbf{E} + \mathbf{J} \times \mathbf{B} \tag{1.3}$$

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:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{1.4}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1.5}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{1.6}$$

$$\nabla \times \mathbf{B} - \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \mu_0 \mathbf{J} \tag{1.7}$$

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

$$\epsilon_0 \approx 8.85 \times 10^{-12}\ m^{-3}\, Kg^{-1}\, s^2\, C^2$$

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

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

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 3.5). Nonetheless, this can be thought of as characterising the strength of magnetic interactions (in units of Coulombs).

The Maxwell equations (1.4), (1.5), (1.6) and (1.7) 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.

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 $\mathbf{J} = 0$. This means that we can consistently set $\mathbf{B} = 0$ and we're left with two of Maxwell's equations to solve. They are

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

and

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

If you fix the charge distribution ρ, equations (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 $V \subset \mathbf{R}^3$ of space. We'll denote the boundary of V by $S = \partial V$. We now integrate both sides of (2.1) over 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 S. We have

$$\int_V d^3x \; \nabla \cdot \mathbf{E} = \int_S \mathbf{E} \cdot d\mathbf{S} = \frac{1}{\epsilon_0} \int_V d^3x \, \rho$$

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

$$\int_S \mathbf{E} \cdot d\mathbf{S} = \frac{Q}{\epsilon_0} \tag{2.3}$$

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

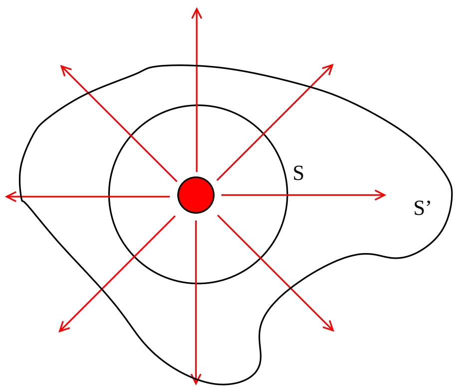


Figure 3: The flux through S and S' is the same.

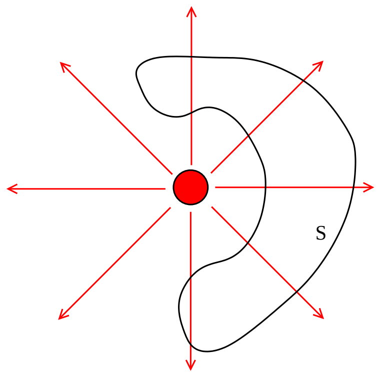


Figure 4: The flux through S vanishes.

Notice that it doesn't matter what shape the surface 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 the left-hand figure above. The choice of S is called the *Gaussian surface*; often there's a smart choice that makes a particular problem simple.

Only charges that lie inside V contribute to the flux. Any charges that lie outside will produce an electric field that penetrates through S at some point, giving negative flux, but leaves through the other side of S, depositing positive flux. The total contribution from these charges that lie outside of 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 (2.3) 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 (2.3) 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 S to be a sphere of radius r as shown in the figure. Gauss' law states

$$\int_S \mathbf{E} \cdot d\mathbf{S} = \frac{Q}{\epsilon_0}$$

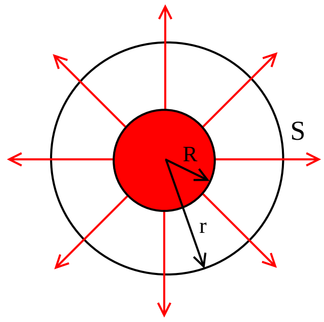


Figure 5:

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

$$\int_S \mathbf{E} \cdot d\mathbf{S} = E(r) \int_S \hat{\mathbf{r}} \cdot d\mathbf{S} = E(r)\, 4\pi r^2 = \frac{Q}{\epsilon_0}$$

where the factor of $4\pi r^2$ has arisen simply because it's the area of the Gaussian sphere. We learn that the electric field outside a spherically symmetric distribution of charge Q is

$$\mathbf{E}(\mathbf{x}) = \frac{Q}{4\pi\epsilon_0 r^2}\, \hat{\mathbf{r}} \tag{2.4}$$

That's nice. This is the familiar result that we've seen before. (See, for example, the notes on *Dynamics and Relativity*). The Lorentz force law (1.2) then tells us that a test charge q moving in the region $r > R$ experiences a force

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

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 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 $\nabla \times \mathbf{E} = 0$. 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 (2.4) does indeed solve $\nabla \times \mathbf{E} = 0$.

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 m_e and charge $-e$ given by

$$e \approx 1.6 \times 10^{-19} \text{ Coulombs} \quad \text{and} \quad m_e \approx 9.1 \times 10^{-31} \text{ Kg}$$

Regardless of the separation, we have

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

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

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

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 (2.4). What about inside? This depends on the distribution in question. The simplest is a sphere of radius R with uniform charge distribution ρ. The total charge is

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

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

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{Qr^3}{\epsilon_0 R^3}$$

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

$$\int_S \mathbf{E}\cdot d\mathbf{S} = E(r)\int_S \hat{\mathbf{r}}\cdot d\mathbf{S} = E(r)\,4\pi r^2 = \frac{Qr^3}{\epsilon_0 R^3}$$

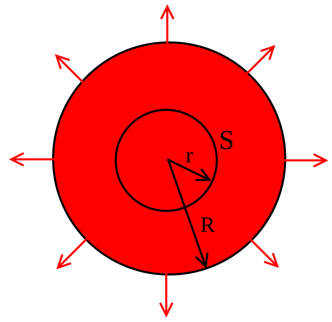


Figure 6:

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

$$\mathbf{E}(\mathbf{x}) = \frac{Qr}{4\pi\epsilon_0 R^3}\,\hat{\mathbf{r}} \qquad r < R \tag{2.5}$$

Outside the sphere we revert to the inverse-square form (2.4). At the surface of the sphere, $r = R$, the electric field is continuous but the derivative, dE/dr, is not. This is shown in the graph.

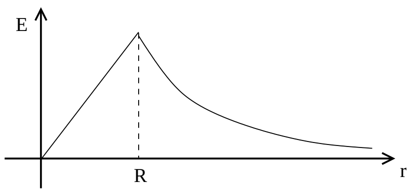


Figure 7:

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. We have

$$\int_S \mathbf{E}\cdot d\mathbf{S} = \frac{\eta L}{\epsilon_0}$$

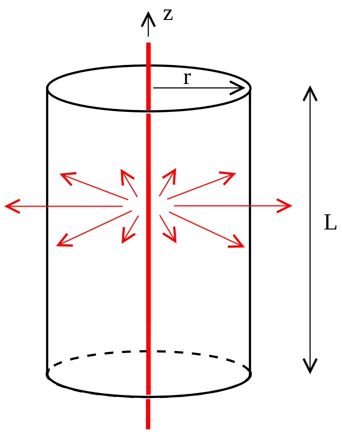


Figure 8:

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 $\mathbf{E} = E(r)\hat{\mathbf{r}}$. The symmetry means that the two end caps of the Gaussian surface don't contribute to the integral because their normal points in 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,

$$\int_S \mathbf{E} \cdot d\mathbf{S} = E(r)\, 2\pi r L = \frac{\eta L}{\epsilon_0}$$

So that the electric field is

$$\mathbf{E}(r) = \frac{\eta}{2\pi\epsilon_0 r}\hat{\mathbf{r}} \tag{2.6}$$

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

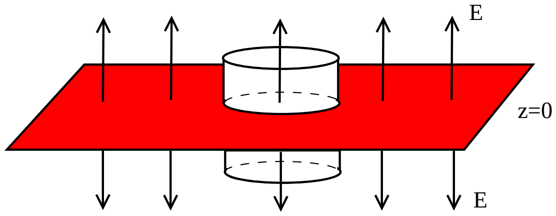


Figure 9:

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

$$E = E(z)\hat{\mathbf{z}}$$

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

$$\int_S \mathbf{E} \cdot d\mathbf{S} = E(z)A - E(-z)A = 2E(z)A = \frac{\sigma A}{\epsilon_0}$$

The electric field above an infinite plane of charge is therefore

$$E(z) = \frac{\sigma}{2\epsilon_0} \tag{2.7}$$

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.

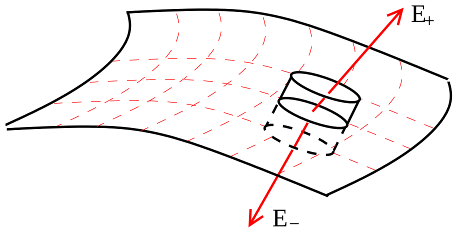


Figure 10: The normal component of the electric field is discontinuous

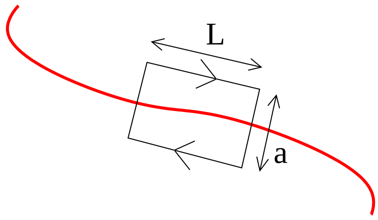


Figure 11: The tangential component of the electric field is continuous.

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

$$E(z \to 0^+) - E(z \to 0^-) = \frac{\sigma}{\epsilon_0} \tag{2.8}$$

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

$$\hat{\mathbf{n}} \cdot \mathbf{E}|_+ - \hat{\mathbf{n}} \cdot \mathbf{E}|_- = \frac{\sigma}{\epsilon_0} \tag{2.9}$$

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 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

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = \int \nabla \times \mathbf{E} \cdot d\mathbf{S}$$

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

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

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

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 the figure. To compute the electric field we need only add the fields for arising from two planes, each of which takes the form (2.7). We find that the electric field between the two planes is

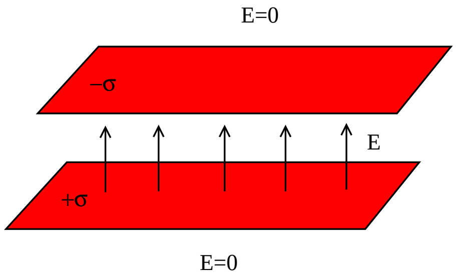


Figure 12:

$$\mathbf{E} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{z}} \qquad 0 < z < a \tag{2.10}$$

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

A Plane Slab

We can rederive the discontinuity (2.9) 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

$$2AE(z) = \frac{2zA\rho}{\epsilon_0} \quad \Rightarrow \quad E(z) = \frac{\rho z}{\epsilon_0}$$

Meanwhile, for $z > d$ we get our earlier result (2.7). 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 (2.8).

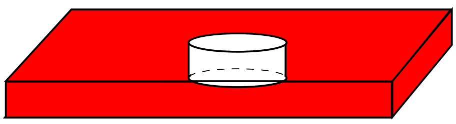


Figure 13: The Gaussian surface for a plane slab

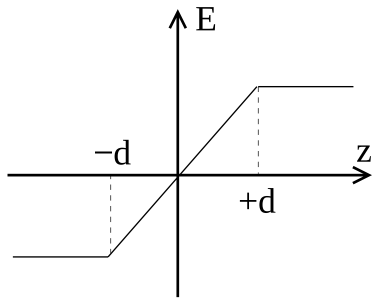


Figure 14: The resulting electric field

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

$$Q = 4\pi R^2 \sigma$$

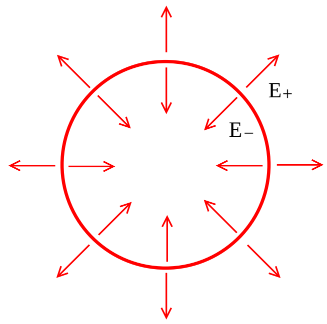


Figure 15:

We already know that outside the shell, $r > R$, the electric field takes the standard inverse-square form (2.4). What about inside? Well, since any surface with $r < R$ doesn't surround a charge, Gauss' law tells us that we necessarily have $\mathbf{E} = 0$ inside. That means that there is a discontinuity at the surface $r = R$,

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

in accord with the expectation (2.9).

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, $\nabla \times \mathbf{E} = 0$.

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

$$\mathbf{E} = -\nabla \phi \tag{2.11}$$

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*

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad \Rightarrow \quad \nabla^2 \phi = -\frac{\rho}{\epsilon_0} \tag{2.12}$$

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

$$\nabla^2 \phi = 0 \tag{2.13}$$

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 Section 5.3.1. For now, we'll eliminate this redundancy by requiring that $\phi(\mathbf{r}) \to 0$ 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. (See Section 2.2 of the *Dynamics and Relativity* lecture notes). Specifically, a test particle of mass m, position $\mathbf{r}(t)$ and charge q moving in a background electric field has conserved energy

$$E = \frac{1}{2} m \dot{\mathbf{r}} \cdot \dot{\mathbf{r}} + q\phi(\mathbf{r})$$

- 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_2 for another charge distribution ρ_2, then the potential for $\rho_1 + \rho_2$ is simply $\phi_1 + \phi_2$. 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 $\nabla \times \mathbf{E} = 0$ is equivalent to writing $\mathbf{E} = -\nabla\phi$. This is true when space is $\mathbf{R}^3$ or, in fact, if we take space to be any open ball in $\mathbf{R}^3$. But if our background space has a suitably complicated topology then there are solutions to $\nabla \times \mathbf{E} = 0$ which cannot be written in the form $\mathbf{E} = -\nabla\phi$. 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 $\mathbf{R}^4$ and, correspondingly, any spatial hypersurface to be $\mathbf{R}^3$.

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

$$\nabla^2\phi = -\frac{Q}{\epsilon_0}\delta^3(\mathbf{x}) \tag{2.14}$$

You've solved problems of this kind in your *Methods* course. The solution is essentially the Green's function for the Laplacian ∇^2, an interpretation that we'll return to in Section 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

$$\phi = \frac{\alpha}{r}$$

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

$$\nabla r = \hat{\mathbf{r}} \tag{2.15}$$

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 $\nabla(1/r) = -\hat{\mathbf{r}}/r^2 = -\mathbf{x}/r^3$. This gives us

$$\nabla\phi = -\frac{\alpha}{r^3}\mathbf{x} \quad \Rightarrow \quad \nabla^2\phi = -\alpha\left(\frac{\nabla\cdot\mathbf{x}}{r^3} - \frac{3\,\mathbf{x}\cdot\mathbf{x}}{r^5}\right)$$

But $\nabla\cdot\mathbf{x} = 3$ 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 normalization 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 (2.14) over some region V which includes the origin. Integrating the charge density gives

$$\rho(\mathbf{x}) = Q\,\delta^3(\mathbf{x}) \quad \Rightarrow \quad \int_V d^3x\,\rho = Q$$

So, using Gauss' law (2.3), we require

$$\int_S \nabla\phi\cdot d\mathbf{S} = -\frac{Q}{\epsilon_0}$$

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 S to be a sphere of radius r, tells us that we must have $\alpha = Q/4\pi\epsilon_0$, or

$$\phi = \frac{Q}{4\pi\epsilon_0 r} \tag{2.16}$$

Taking the gradient of this using (2.15) gives us Coulomb's law

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

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 $\int d^3x$ and sometimes as $\int d^3r$. The advantage of the $\mathbf{r}$ 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,

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

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

$$f(\mathbf{r}+\mathbf{d}) \approx f(\mathbf{r}) + \mathbf{d}\cdot\nabla f(\mathbf{r}) + \frac{1}{2}(\mathbf{d}\cdot\nabla)^2 f(\mathbf{r}) + \ldots \tag{2.17}$$

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

$$\begin{aligned}\frac{1}{|\mathbf{r}+\mathbf{d}|} &\approx \frac{1}{r} + \mathbf{d}\cdot\nabla\frac{1}{r} + \frac{1}{2}\left(\mathbf{d}\cdot\nabla\right)^2\frac{1}{r} + \ldots \\ &= \frac{1}{r} - \frac{\mathbf{d}\cdot\mathbf{r}}{r^3} - \frac{1}{2}\left(\frac{\mathbf{d}\cdot\mathbf{d}}{r^3} - \frac{3(\mathbf{d}\cdot\mathbf{r})^2}{r^5}\right) + \ldots\end{aligned}$$

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

$$\phi \approx \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{r} - \frac{1}{r} - \mathbf{d} \cdot \nabla \frac{1}{r} + \ldots\right) = \frac{Q}{4\pi\epsilon_0}\frac{\mathbf{d} \cdot \mathbf{r}}{r^3} + \ldots \tag{2.18}$$

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

$$\mathbf{p} = Q\mathbf{d}$$

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

$$\mathbf{E} = -\nabla\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{3(\mathbf{p} \cdot \hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{p}}{r^3}\right) + \ldots \tag{2.19}$$

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 $\ldots$ terms in (2.18) and (2.19) disappear since they contain higher powers of d. Often when people talk about the "dipole", they implicitly mean taking this limit.

2.2.3 General Charge Distributions

Our derivation of the potential due to a point charge (2.16), 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

$$\nabla^2 G(\mathbf{r}; \mathbf{r}') = \delta^3(\mathbf{r} - \mathbf{r}')$$

which, from our discussion of the point charge, we now know to be

$$G(\mathbf{r}; \mathbf{r}') = -\frac{1}{4\pi}\frac{1}{|\mathbf{r} - \mathbf{r}'|} \tag{2.20}$$

We can now apply our usual Green's function methods to the general Poisson equation (2.12). In what follows, we'll take $\rho(\mathbf{r}) \neq 0$ only in some compact region, V, of space. The solution to the Poisson equation is given by

$$\phi(\mathbf{r}) = -\frac{1}{\epsilon_0}\int_V d^3r'\; G(\mathbf{r};\mathbf{r}')\,\rho(\mathbf{r}') = \frac{1}{4\pi\epsilon_0}\int_V d^3r'\; \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \tag{2.21}$$

(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

$$\begin{aligned}\mathbf{E}(\mathbf{r}) = -\nabla\phi(\mathbf{r}) &= -\frac{1}{4\pi\epsilon_0}\int_V d^3r'\;\rho(\mathbf{r}')\,\nabla\frac{1}{|\mathbf{r}-\mathbf{r}'|} \\ &= \frac{1}{4\pi\epsilon_0}\int_V d^3r'\;\rho(\mathbf{r}')\,\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}\end{aligned}$$

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...

Long Distance Behaviour

Suppose now that you want to know what the electric field looks like far from the region V. This means that we're interested in the electric field at $\mathbf{r}$ with $|\mathbf{r}| \gg |\mathbf{r}'|$ for all $\mathbf{r}' \in V$. We can apply the same Taylor expansion (2.17), now replacing $\mathbf{d}$ with $-\mathbf{r}'$ for each $\mathbf{r}'$ in the charged region. This means we can write

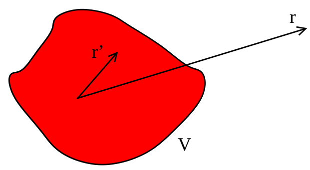


Figure 16:

$$\begin{aligned}\frac{1}{|\mathbf{r}-\mathbf{r}'|} &= \frac{1}{r} - \mathbf{r}'\cdot\nabla\frac{1}{r} + \frac{1}{2}(\mathbf{r}'\cdot\nabla)^2\frac{1}{r} + \ldots \\ &= \frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \frac{1}{2}\left(\frac{3(\mathbf{r}\cdot\mathbf{r}')^2}{r^5} - \frac{\mathbf{r}'\cdot\mathbf{r}'}{r^3}\right) + \ldots\end{aligned} \tag{2.22}$$

and our potential becomes

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int_V d^3r'\,\rho(\mathbf{r}')\left(\frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \ldots\right)$$

The leading term is just

$$\phi(\mathbf{r}) = \frac{Q}{4\pi\epsilon_0 r} + \ldots$$

where $Q = \int_V d^3r'\, \rho(\mathbf{r}')$ is the total charge contained within 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,

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{r} + \frac{\mathbf{p}\cdot\hat{\mathbf{r}}}{r^2} + \ldots\right)$$

where

$$\mathbf{p} = \int_V d^3r'\ \mathbf{r}'\rho(\mathbf{r}')$$

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

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

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

$$\mathbb{Q}_{ij} = \int_V d^3r'\ \rho(\mathbf{r}')\left(3r'_i r'_j - \delta_{ij} r'^2\right)$$

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.

A Comment on Infinite Charge Distributions

In the above, we assumed for simplicity that the charge distribution was restricted to some compact region of space, 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, we computed the electric field in (2.6). It goes as

$$E(r) = \frac{\rho}{2\pi r}\hat{\mathbf{r}}$$

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

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

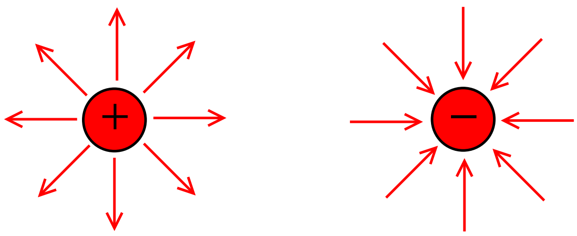
Because of the log function, we necessarily have $\phi(r) \to \infty$ as $r \to \infty$. Instead, we need to pick an arbitrary, but finite distance, r_0 at which the potential vanishes.

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).

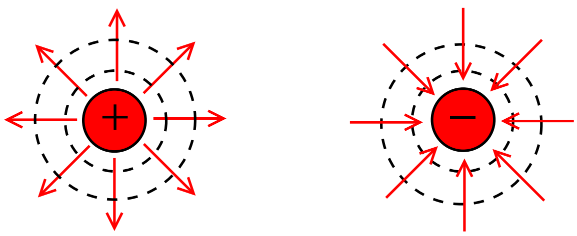
Field lines are continuous. They begin and end only at charges. They can never cross.

The field lines for positive and negative point charges are:

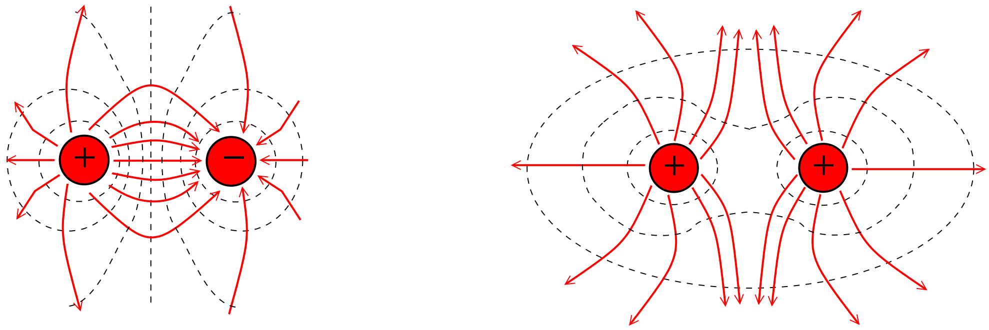


By convention, the positive charges act as sources for the lines, with the arrows emerging. The negative charges act as sinks, with the arrows approaching.

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 $\mathbf{E} = -\nabla\phi$ ensures that the equipotentials cut the field lines at right angles. We usually draw them as dotted lines:



Meanwhile, we can (very) roughly sketch the field lines and equipotentials for the dipole (on the left) and for a pair of charges of the same sign (on the right):



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.

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 $\mathbf{r}_\star$ that is stable for a particle of charge $q > 0$. By "empty space", we mean that $\rho(\mathbf{r}) = 0$ in a neighbourhood of $\mathbf{r}_\star$. 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 $\mathbf{r}_\star$; never away. We could then surround $\mathbf{r}_\star$ by a small surface S and compute

$$\int_S \mathbf{E} \cdot d\mathbf{S} < 0$$

But, by Gauss' law, the right-hand side must be the charge contained within 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.

Let's start by recalling a fact from our first course on classical mechanics[1]. Suppose we have some test charge q moving in a background electrostatic potential ϕ. We'll denote the potential energy of the particle as $U(\mathbf{r})$. (We used the notation $V(\mathbf{r})$ in the *Dynamics and Relativity* course but we'll need to reserve V for the voltage later). The potential $U(\mathbf{r})$ of the particle can be thought of as the work done bringing the particle in from infinity;

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

where we've assumed our standard normalization 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

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

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.

[1]See Section 2.2 of the lecture notes on *Dynamics and Relativity*.

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

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

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

$$U = \sum_{i=1}^{N} W_i = \frac{1}{4\pi\epsilon_0}\sum_{i<j}\frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.23}$$

where $\sum_{i<j}$ means that we sum over each pair of particles once. In fact, you probably could have just written down (2.23) 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

$$U = \frac{1}{2}\frac{1}{4\pi\epsilon_0}\sum_{i}\sum_{j\neq i}\frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{2.24}$$

where now we sum over each pair twice.

There is a slicker way of writing (2.24). The potential at $\mathbf{r}_i$ due to all the other charges q_j, $j \neq i$ is

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

which means that we can write the potential energy as

$$U = \frac{1}{2}\sum_{i=1}^{N} q_i\phi(\mathbf{r}_i) \tag{2.25}$$

This is the potential energy for a set of point charges. But there is an obvious generalization 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

$$U = \frac{1}{2}\int d^3r\ \rho(\mathbf{r})\,\phi(\mathbf{r}) \tag{2.26}$$

where we can quite happily take the integral over all of $\mathbf{R}^3$, 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

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

But the first term is a total derivative. And since we're taking the integral over all of space and $\phi(\mathbf{r}) \to 0$ as $r \to \infty$, this term just vanishes. In the second term we can replace $\nabla\phi = -\mathbf{E}$. 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,

$$U = \frac{\epsilon_0}{2}\int d^3r\ \mathbf{E}\cdot\mathbf{E} \tag{2.27}$$

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 (2.27) 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 (2.27) 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

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}} \tag{2.28}$$

So, by (2.27), 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 (2.25) to $\rho\phi$ in (2.26). 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 (2.27) 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 (2.27) 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 (2.28) 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

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

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

$$m_e c^2 > \frac{e^2}{8\pi\epsilon_0}\frac{1}{a} \quad\quad \Rightarrow \quad\quad a > \frac{e^2}{8\pi\epsilon_0}\frac{1}{m_e c^2} \tag{2.29}$$

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.

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

$$a = \frac{\hbar}{m_e c}$$

We see that the inequality (2.29) translates into an inequality on a bunch of fundamental constants. For the whole story to hang together, we require

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

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

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

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

$$\alpha \approx \frac{1}{137}$$

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

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\frac{\mathbf{p}_1 \cdot \mathbf{r}}{r^3}$$

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 (2.23),

$$\begin{aligned} U = Q\Big(\phi(\mathbf{r}) - \phi(\mathbf{r} - \mathbf{d})\Big) &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{p}_1 \cdot \mathbf{r}}{r^3} - \frac{\mathbf{p}_1 \cdot (\mathbf{r} - \mathbf{d})}{|\mathbf{r} - \mathbf{d}|^3}\right) \\ &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{p}_1 \cdot \mathbf{r}}{r^3} - \mathbf{p}_1 \cdot (\mathbf{r} - \mathbf{d})\left(\frac{1}{r^3} + \frac{3\mathbf{d} \cdot \mathbf{r}}{r^5} + \ldots\right)\right) \\ &= \frac{Q}{4\pi\epsilon_0}\left(\frac{\mathbf{p}_1 \cdot \mathbf{d}}{r^3} - \frac{3(\mathbf{p}_1 \cdot \mathbf{r})(\mathbf{d} \cdot \mathbf{r})}{r^5}\right) \end{aligned}$$

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 $\mathbf{r}$,

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

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

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\nabla\left(\frac{3(\mathbf{p}_1\cdot\mathbf{r})(\mathbf{p}_2\cdot\mathbf{r})}{r^5} - \frac{\mathbf{p}_1\cdot\mathbf{p}_2}{r^3}\right) \tag{2.30}$$

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 $\nabla\cdot\mathbf{E} = \rho/\epsilon_0$, 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 $\mathbf{E} = -\nabla\phi$ 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 (2.9), together with the fact that $\mathbf{E} = 0$ inside, the electric field just outside the conductor must be

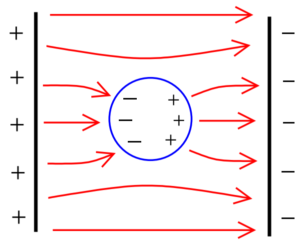
$$\mathbf{E} = \frac{\sigma}{\epsilon_0}\hat{\mathbf{n}} \tag{2.31}$$

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.

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 Section 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 this:



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

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 Section 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 $-Q$.

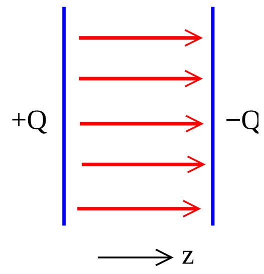


Figure 17:

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 Section 2.1.4. The electric field necessarily vanishes inside the conductor while, between the plates we have the result (2.10),

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

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

$$C = \frac{Q}{V}$$

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 $E = -d\phi/dz$ is constant, we must have

$$\phi = -Ez + c \qquad \Rightarrow \qquad V = \phi(0) - \phi(d) = Ed = \frac{Qd}{A\epsilon_0}$$

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

$$C = \frac{A\epsilon_0}{d}$$

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 see another example below.

Capacitors are usually employed as a method to store electrical energy. We can see how much. Using our result (2.27), we have

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

This is the energy stored in a parallel plate capacitor.

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

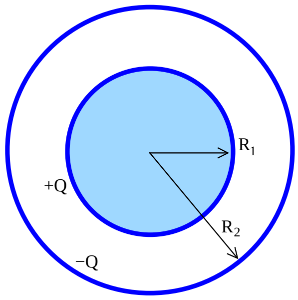


Figure 18:

$$\mathbf{E} = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}} \qquad\qquad R_1 < r < R_2$$

Correspondingly, the potential is

$$\phi = \frac{Q}{4\pi\epsilon_0 r} \qquad\qquad R_1 < r < R_2$$

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

$$\nabla^2\phi = 0$$

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 $\nabla\phi \cdot \hat{\mathbf{n}}$ is fixed perpendicular to a given surface S_i

Notice that, for each 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 the following theorem.

Theorem: With either Dirichlet or Neumann boundary conditions chosen on each surface S_i, the Laplace equation has a unique solution.

Proof: Suppose that there are two solutions, ϕ_1 and ϕ_2 with the same specified boundary conditions. Let's define $f = \phi_1 - \phi_2$. We can look at the following expression

$$\int_V d^3r\ \nabla \cdot (f\nabla f) = \int_V d^3r\ \nabla f \cdot \nabla f \tag{2.32}$$

where the $\nabla^2 f$ term vanishes by the Laplace equation. But, by the divergence theorem, we know that

$$\int_V d^3r\ \nabla \cdot (f\nabla f) = \sum_i \int_{S_i} f\,\nabla f \cdot d\mathbf{S}$$

However, if we've picked Dirichlet boundary conditions then $f = 0$ on the boundary, while Neumann boundary conditions ensure that $\nabla f = 0$ on the boundary. This means that the integral vanishes and, from (2.32), we must have $\nabla f = 0$ throughout space. But if we have imposed Dirichlet boundary conditions somewhere, then $f = 0$ on that boundary and so $f = 0$ everywhere. Alternatively, if we have Neumann boundary conditions on all surfaces than $\nabla f = 0$ 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. $\square$

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*.

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 $x = -\mathbf{d}$. This is called the *image charge*. The potential for this pair of charges is just the potential

$$\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{q}{\sqrt{(x-d)^2+y^2+z^2}} - \frac{q}{\sqrt{(x+d)^2+y^2+z^2}}\right) \tag{2.33}$$

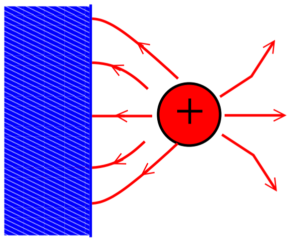


Figure 19: A particle near a conducting plane...

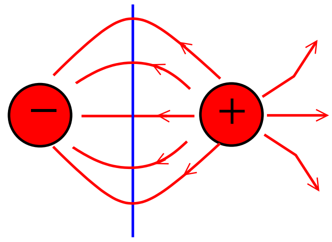


Figure 20: ...looks like a dipole

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 the figures. 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 $\phi = 0$ for $x < 0$.

With the solution (2.33) 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 (2.33). If we focus on the electric field in the x direction, it is

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

Meanwhile, $E_x = 0$ for $x < 0$. The discontinuity of E_x at the surface of the conductor determines the induced surface charge (2.31). It is

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

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,

$$q_{\text{induced}} = \int dydz\; \sigma = -q$$

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

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

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

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 $q' = -qR/d$, placed at $x = R^2/d$ and, by symmetry, $y = z = 0$. A cartoon of this is shown in the figure.

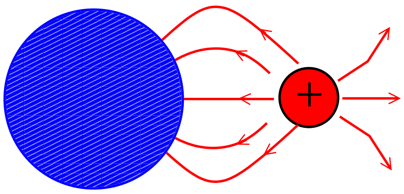


Figure 21: A particle near a conducting sphere...

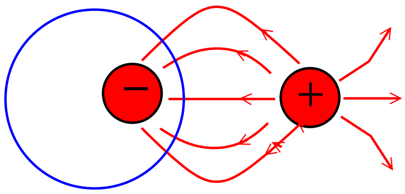


Figure 22: ...looks like a slightly different dipole

The resulting potential is

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

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 electric field is perpendicular to the surface. This is now our solution.

2.4.4 Many many more problems

There are many more 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,

$$\mathbf{E}_0 = E_0\hat{\mathbf{z}} \qquad \Rightarrow \qquad \phi_0 = -E_0 z = -E_0 r\cos\theta$$

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 (2.18). We'll place the dipole at the origin, and orient it along the $\mathbf{z}$ axis like so:

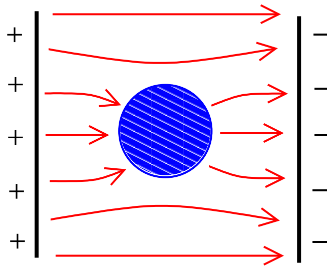


Figure 23: A conducting sphere between charged plates...

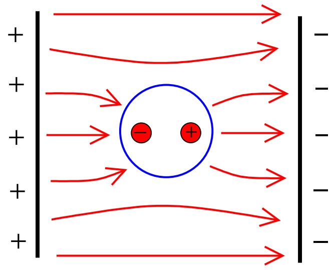


Figure 24: ...looks like a dipole between the plates

The resulting potential is

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

Since we've added a dipole term, we can be sure that this still solves the Laplace equation outside the conductor. 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 the conductor. It is

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

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

2.4.5 A History of Electrostatics

Perhaps the simplest demonstration of the attractive properties of electric charge comes from rubbing a balloon on your head and sticking it to the wall. This phenomenon was known, at least in spirit, to the ancient Greeks and is credited to Thales of Miletus around 600 BC. Although, in the absence of any ancient balloons, he had to make do with polishing pieces of amber and watching it attract small objects.

A systematic, scientific approach to electrostatics starts with William Gilbert, physicist, physician and one-time bursar of St Johns College, Cambridge. (Rumour has it that he'd rather have been at Oxford.) His most important work, *De Magnete*, published in 1600 showed, among other things, that many materials, not just amber, could be electrified. With due deference, he referred to these as "electrics", derived from the Greek "$\eta\lambda\epsilon\kappa\tau\rho o\nu$" (electron) meaning "amber". These are materials that we now call "insulators".

There was slow progress over the next 150 years, much of it devoted to building machines which could store electricity. A notable breakthrough came from the experiments of the little-known English scientist Stephen Grey, who was the first to appreciate that the difficulty in electrifying certain objects is because they are conductors, with any charge quickly flowing through them and away. Grey spent most of his life as an amateur astronomer, although his amateur status appears to be in large part because he fell foul of Isaac Newton who barred his entry into more professional scientific circles. He performed his experiments on conductors in the 1720s, late in life when the lack of any income left him destitute and pensioned to Chaterhouse (which was, perhaps, the world's fanciest poorhouse). Upon Newton's death, the scientific community clamoured to make amends. Grey was awarded the Royal Society's first Copley medal. Then, presumably because they felt guilty, he was also awarded the second. Grey's experiments were later reproduced by the French chemist Charles François de Cisternay DuFay, who came to the wonderful conclusion that all objects can be electrified by rubbing apart from "metals, liquids and animals". He does not, to my knowledge, state how much rubbing of animals he tried before giving up. He was also the first to notice that static electricity can give rise to both attractive and repulsive forces.

By the 1750s, there were many experiments on electricity, but little theory to explain them. Most ideas rested on a fluid description of electricity, but arguments raged over whether a single fluid or two fluids were responsible. The idea that there were both positive and negative charges, then thought of as a surplus and deficit of fluid, was introduced independently by the botanist William Watson and the US founding father

Benjamin Franklin. Franklin is arguably the first to suggest that charge is conserved although his statement wasn't quite as concise as the continuity equation:

> It is now discovered and demonstrated, both here and in Europe, that the Electrical Fire is a real Element, or Species of Matter, not created by the Friction, but collected only.
>
> *Benjamin Franklin, 1747*

Still, it's nice to know that charge is conserved both in the US and in Europe.

A quantitative understanding of the theory of electrostatics came only in the 1760s. A number of people suggested that the electrostatic force follows an inverse-square law, prominent among them Joseph Priestly who is better known for the discovery of Oxygen and, of at least equal importance, the invention of soda water. In 1769, the Scottish physicist John Robison announced that he had measured the force to fall off as $1/r^{2.06}$. This was before the invention of error bars and he seems to receive little credit. Around the same time, the English scientist Henry Cavendish, discover of Hydrogen and weigher of the Earth, performed a number of experiments to demonstrate the inverse-square law but, as with his many of his other electromagnetic discoveries, he chose not to publish. It was left to French physicist Charles Augustin de Coulomb to clean up, publishing the results of his definitive experiments in 1785 on the force that now carries his name.

In its final form, Coulomb's law becomes transmuted into Gauss' law. For once, this was done by the person after whom it's named. Gauss derived this result in 1835, although it wasn't published until 1867.

3. Magnetostatics

Charges give rise to electric fields. Current give rise to magnetic fields. In this section, 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:

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

and

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

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

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

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$$

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

$$\nabla \cdot \mathbf{J} = 0$$

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, $\nabla \cdot (\nabla \times \mathbf{B}) = 0$.

3.1 Ampère's Law

The first equation of magnetostatics,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{3.3}$$

is known as *Ampère'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 S with boundary $C = \partial S$. Integrating (3.3) over the surface, we can use Stokes' theorem to turn the integral of $\nabla \times \mathbf{B}$ into a line integral over the boundary C,

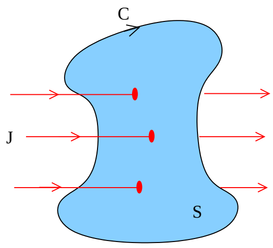


Figure 25:

$$\int_S \nabla \times \mathbf{B} \cdot d\mathbf{S} = \oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 \int_S \mathbf{J} \cdot d\mathbf{S}$$

Recall that there's an implicit orientation in these equations. The surface S comes with a normal vector $\hat{\mathbf{n}}$ which points away from 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 S is the same thing as the total current I that passes through S. Ampère's law in integral form then reads

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 I \tag{3.4}$$

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 (3.4) 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 S to lie in the $x - y$ plane, centered on the wire. For the line integral in (3.4) 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 below). Any magnetic field of this form automatically satisfies the second Maxwell equation $\nabla\cdot\mathbf{B} = 0$. We need only worry about Ampère's law which tells us

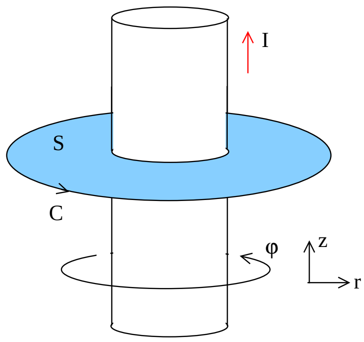


Figure 26:

$$\oint_C \mathbf{B}\cdot d\mathbf{r} = B(r)\int_0^{2\pi} r\, d\varphi = 2\pi r B(r) = \mu_0 I$$

We see that the strength of the magnetic field is

$$\mathbf{B} = \frac{\mu_0 I}{2\pi r}\hat{\boldsymbol{\varphi}} \tag{3.5}$$

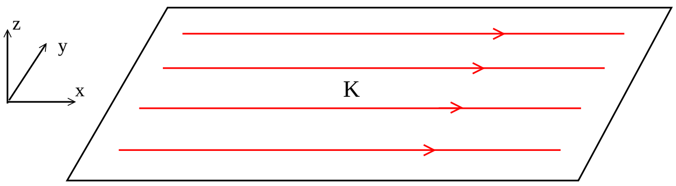
The magnetic field circles the wire using the "right-hand rule": stick the thumb of your right hand in the direction of the current and your fingers curl in the direction of the magnetic field.

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 Section 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

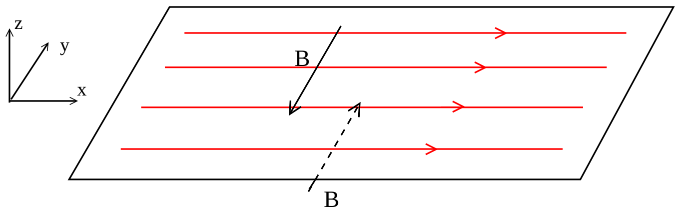
Consider the flat plane lying at $z = 0$ with a surface current density that we'll call $\mathbf{K}$. Note that $\mathbf{K}$ is the current per unit length, as opposed to $\mathbf{J}$ which is the current per unit area. You can think of the surface current as a bunch of wires, all lying parallel to each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

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



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

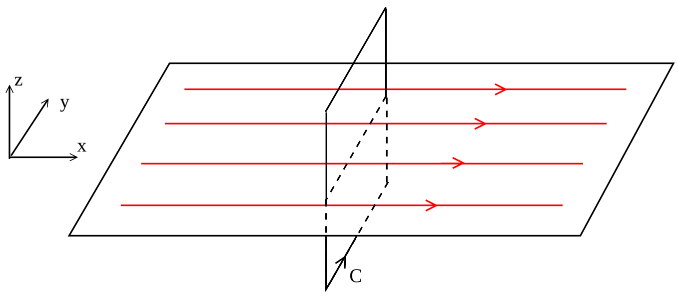
$\mathbf{B}$ is oriented along the $\mathbf{y}$ direction. In fact, from the symmetry of the problem, it must look like



with $\mathbf{B}$ pointing in the $-\hat{\mathbf{y}}$ direction when $z > 0$ and in the $+\hat{\mathbf{y}}$ direction when $z < 0$. We write

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

with $B(z) = -B(-z)$. We invoke Ampère's law using the following open surface:



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

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

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

$$B(z) = \frac{\mu_0 K}{2} \qquad z > 0$$

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

$$B(z \to 0^+) - B(z \to 0^-) = \mu_0 K$$

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 the Figure 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

$$\hat{\mathbf{n}} \times \mathbf{B}|_+ - \hat{\mathbf{n}} \times \mathbf{B}|_- = \mu_0 \mathbf{K} \tag{3.6}$$

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 (2.9) 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.

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.

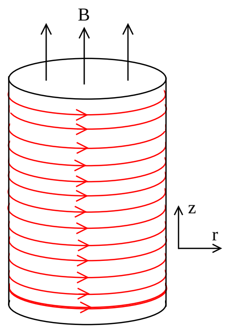


Figure 27:

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 $\mathbf{B}$ 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 $\nabla \cdot \mathbf{B} = 0$.

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

$$\nabla \times \mathbf{B} = 0 \quad \Rightarrow \quad \frac{dB}{dr} = 0 \quad \Rightarrow \quad B(r) = \text{constant}$$

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 Ampère's law and consider the surface S, bounded by the curve C shown in the figure. Only the line that runs inside the solenoid contributes to the line integral. We have

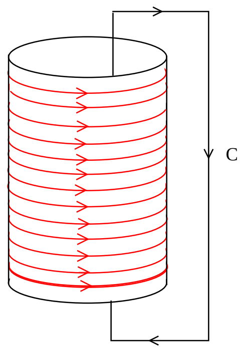


Figure 28:

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = BL = \mu_0 INL$$

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

$$\mathbf{B} = \mu_0 I N \, \hat{\mathbf{z}} \tag{3.7}$$

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 (3.6).

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

$$\nabla \cdot \mathbf{B} = 0 \tag{3.8}$$

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 $\nabla \cdot \mathbf{B} = 0$ if we write the magnetic field as the curl of some vector field,

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{3.9}$$

Here $\mathbf{A}$ is called the *vector potential*. While magnetic fields that can be written in the form (3.9) certainly satisfy $\nabla \cdot \mathbf{B} = 0$, the converse is also true; any divergence-free magnetic field can be written as (3.9) 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 $\mathbf{R}^3$ or some open ball on $\mathbf{R}^3$, we rarely run into subtleties. But if space becomes more interesting then the possible solutions to $\nabla \cdot \mathbf{B} = 0$ 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 (3.9), Ampère's law becomes

$$\nabla \times \mathbf{B} = -\nabla^2 \mathbf{A} + \nabla(\nabla \cdot \mathbf{A}) = \mu_0 \mathbf{J} \tag{3.10}$$

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, $\mathbf{B}$.

3.2.1 Magnetic Monopoles

Above, we dispatched with the Maxwell equation $\nabla \cdot \mathbf{B} = 0$ fairly quickly by writing $\mathbf{B} = \nabla \times \mathbf{A}$. 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

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

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 (3.8) to read $\nabla \cdot \mathbf{B} = \rho_m$ 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 $\mathbf{B} = \nabla \times \mathbf{A}$. I won't explain the details here, but if you're interested then look up the "Aharonov-Bohm effect".

Monopoles After All?

To summarise, magnetic monopoles have never been observed. We have a law of physics (3.8) 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 (3.8) 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

$$ge = 2\pi\hbar n \qquad n \in \mathbf{Z} \tag{3.11}$$

This is known as the *Dirac quantization 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 (3.9) 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 $\nabla\chi$ for some function χ and the magnetic field remains the same,

$$\mathbf{A}' = \mathbf{A} + \nabla\chi \quad \Rightarrow \quad \nabla\times\mathbf{A}' = \nabla\times\mathbf{A}$$

Such a change of $\mathbf{A}$ is called a *gauge transformation*. As we will see in Section 5.3.1, 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.

Claim: We can always find a gauge transformation χ such that $\mathbf{A}'$ satisfies $\nabla\cdot\mathbf{A}' = 0$. Making this choice is usually referred to as *Coulomb gauge*.

Proof: Suppose that we've found some $\mathbf{A}$ which gives us the magnetic field that we want, so $\nabla\times\mathbf{A} = \mathbf{B}$, but when we take the divergence we get some function $\nabla\cdot\mathbf{A} = \psi(\mathbf{x})$. We instead choose $\mathbf{A}' = \mathbf{A} + \nabla\chi$ which now has divergence

$$\nabla\cdot\mathbf{A}' = \nabla\cdot\mathbf{A} + \nabla^2\chi = \psi + \nabla^2\chi$$

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

$$\nabla^2 \chi = -\psi$$

But this is just the Poisson equation again. And we know from our discussion in Section 2 that there is always a solution. (For example, we can write it down in integral form using the Green's function). $\square$

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 $\nabla \times \mathbf{B} = 0$, which you can do by writing

$$\mathbf{B} = -\nabla \Omega$$

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 Section 5.3.1, 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 $\mathbf{B} = -\nabla \Omega$ 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 $\nabla \cdot \mathbf{A} = 0$. Then Ampère's law (3.10) becomes a whole lot easier: we just have to solve

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J} \tag{3.12}$$

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,

$$\nabla^2 A_i = -\mu_0 J_i \quad (i = 1, 2, 3) \tag{3.13}$$

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 $\nabla^2\mathbf{A} = -\nabla\times(\nabla\times\mathbf{A}) + \nabla(\nabla\cdot\mathbf{A})$ 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 (3.13) are of exactly the same form that we had to solve in electrostatics. And, in analogy to (2.21), we know how to write down the most general solution using Green's functions. It is

$$A_i(\mathbf{x}) = \frac{\mu_0}{4\pi}\int_V d^3x'\;\frac{J_i(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}$$

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

$$\mathbf{A}(\mathbf{x}) = \frac{\mu_0}{4\pi}\int_V d^3x'\;\frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} \tag{3.14}$$

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 $\mathbf{x}'$).

Checking Coulomb Gauge

We've derived a solution to (3.12), but this is only a solution to Ampère's equation (3.10) if the resulting $\mathbf{A}$ obeys the Coulomb gauge condition, $\nabla\cdot\mathbf{A} = 0$. Let's now check that it does. We have

$$\nabla\cdot\mathbf{A}(\mathbf{x}) = \frac{\mu_0}{4\pi}\int_V d^3x'\;\nabla\cdot\left(\frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\right)$$

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

$$\begin{aligned}\nabla\cdot\mathbf{A}(\mathbf{x}) &= \frac{\mu_0}{4\pi}\int_V d^3x'\;\mathbf{J}(\mathbf{x}')\cdot\nabla\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right)\\ &= -\frac{\mu_0}{4\pi}\int_V d^3x'\;\mathbf{J}(\mathbf{x}')\cdot\nabla'\left(\frac{1}{|\mathbf{x}-\mathbf{x}'|}\right)\end{aligned}$$

Here we've done something clever. Now our ∇' 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 ∇' sits inside an $\int d^3x'$ integral, it's ripe for integrating by parts. This gives

$$\nabla \cdot \mathbf{A}(\mathbf{x}) = -\frac{\mu_0}{4\pi} \int_V d^3x' \left[\nabla' \cdot \left(\frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} \right) - \nabla' \cdot \mathbf{J}(\mathbf{x}') \left(\frac{1}{|\mathbf{x} - \mathbf{x}'|} \right) \right]$$

The second term vanishes because we're dealing with steady currents obeying $\nabla \cdot \mathbf{J} = 0$. The first term also vanishes if we take the current to be localised in some region of space, $\hat{V} \subset V$ so that $\mathbf{J}(\mathbf{x}) = 0$ on the boundary ∂V. We'll assume that this is the case. We conclude that

$$\nabla \cdot \mathbf{A} = 0$$

and (3.14) is indeed the general solution to the Maxwell equations (3.1) and (3.2) as we'd hoped.

The Magnetic Field

From the solution (3.14), it is simple to compute the magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$. Again, we need to remember that the ∇ acts on the $\mathbf{x}$ in (3.14) rather than the $\mathbf{x}'$. We find

$$\mathbf{B}(\mathbf{x}) = \frac{\mu_0}{4\pi} \int_V d^3x' \, \frac{\mathbf{J}(\mathbf{x}') \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3} \tag{3.15}$$

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 (3.15) 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 C. Then, for a current density $\mathbf{J}$ passing through a small volume δV, we write $\mathbf{J}\delta V = (JA)\delta\mathbf{x}$ where A is the cross-sectional area of the wire and δx lies tangent to C. Assuming that the cross-sectional area is constant throughout the wire, the current $I = JA$ is also constant. The Biot-Savart law becomes

$$\mathbf{B}(\mathbf{x}) = \frac{\mu_0 I}{4\pi} \int_C \frac{d\mathbf{x}' \times (\mathbf{x} - \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3} \tag{3.16}$$

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

An Example: The Straight Wire Revisited

Of course, we already derived the answer for a straight wire in (3.5) 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 $d\mathbf{x}' = \hat{\mathbf{z}}dz$ and, for a point $\mathbf{x}$ away from the wire, the vector $d\mathbf{x}' \times (\mathbf{x} - \mathbf{x}')$ points along the tangent to the circle of radius r,

$$d\mathbf{x}' \times (\mathbf{x} - \mathbf{x}') = r\hat{\boldsymbol{\varphi}}\, dz$$

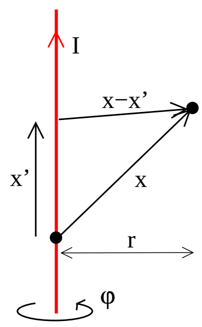


Figure 29:

So we have

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

which is the same result we found earlier (3.5).

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, C and C', in $\mathbf{R}^3$. For each pair of curves, there is an integer $n \in Z$ 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.

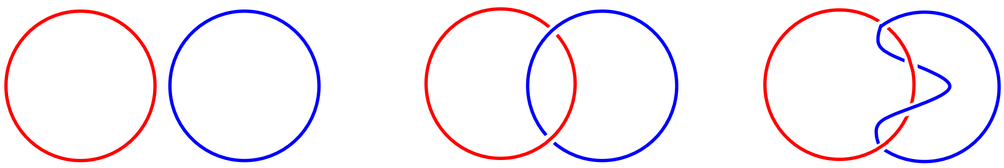


Figure 30: Curves with linking number $n = 0$, $n = 1$ and $n = 2$.

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

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

where we've changed our conventions somewhat from (3.16): now $\mathbf{x}$ labels coordinates on C while $\mathbf{x}'$ labels coordinates on C'.

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

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

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

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

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

$$n = \frac{1}{4\pi}\oint_{C'} d\mathbf{x}'\cdot\oint_C \frac{d\mathbf{x}\times(\mathbf{x}'-\mathbf{x})}{|\mathbf{x}-\mathbf{x}'|^3} \tag{3.17}$$

Note that our final expression is symmetric in C and C', even though these two curves played a rather different physical role in the original definition, with C carrying a current, and 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 (3.17) is rather pretty. It's not at all obvious that the right-hand-side doesn't change under (non-crossing) deformations of C and 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 (3.11). 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 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 the figure However, we can be more accurate. Here we restrict ourselves only to the magnetic field far from the loop.

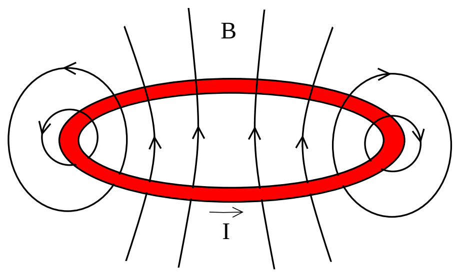


Figure 31:

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 (3.14). 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 $r = |\mathbf{r}|$). The vector potential is then given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_V d^3r' \, \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}$$

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

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi}\oint_C \frac{d\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}$$

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 (2.22),

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \ldots$$

So that

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0 I}{4\pi}\oint_C d\mathbf{r}' \left(\frac{1}{r} + \frac{\mathbf{r}\cdot\mathbf{r}'}{r^3} + \ldots\right) \tag{3.18}$$

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

$$\oint_C d\mathbf{r}' \cdot \mathbf{g}\,(\mathbf{r} \cdot \mathbf{r}')$$

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

$$\oint_C d\mathbf{r}' \cdot \mathbf{g}\,(\mathbf{r} \cdot \mathbf{r}') = \int_S d\mathbf{S} \cdot \nabla \times (\mathbf{g}\,(\mathbf{r} \cdot \mathbf{r}')) = \int_S dS_i\, \epsilon_{ijk} \partial'_j (g_k r_l r'_l)$$

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

$$\oint_C d\mathbf{r}' \cdot \mathbf{g}\,(\mathbf{r} \cdot \mathbf{r}') = \int_S dS_i\, \epsilon_{ijk} g_k r_j = \mathbf{g} \cdot \int_S d\mathbf{S} \times \mathbf{r}$$

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

$$\oint_C d\mathbf{r}'\,(\mathbf{r} \cdot \mathbf{r}') = \boldsymbol{\mathcal{S}} \times \mathbf{r}$$

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

$$\boldsymbol{\mathcal{S}} = \int_S d\mathbf{S}$$

If the boundary 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 (3.18). With the integral over $\mathbf{r}'$, we can treat $\mathbf{r}$ as the constant vector $\mathbf{g}$ that we introduced in the lemma. With the first term vanishing, we're left with

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{m} \times \mathbf{r}}{r^3} \tag{3.19}$$

where we've introduced the *magnetic dipole moment*

$$\mathbf{m} = I\boldsymbol{\mathcal{S}}$$

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

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}}{r^3}\right) \tag{3.20}$$

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 (2.19).

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 (2.22) in the general form of the vector potential (3.14),

$$A_i(\mathbf{r}) = \frac{\mu_0}{4\pi}\int d^3r'\ \frac{J_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} = \frac{\mu_0}{4\pi}\int d^3r'\ \left(\frac{J_i(\mathbf{r}')}{r} + \frac{J_i(\mathbf{r}')\,(\mathbf{r}\cdot\mathbf{r}')}{r^3} + \ldots\right) \tag{3.21}$$

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

$$\partial_j(J_j r_i) = (\partial_j J_j)\, r_i + J_i = J_i \tag{3.22}$$

where the last equality follows from the continuity condition $\nabla\cdot\mathbf{J} = 0$. Using this, we see that the first term in (3.21) is a total derivative (of $\partial/\partial r'_i$ rather than $\partial/\partial r_i$) which vanishes if we take the integral over $\mathbf{R}^3$ and keep the current localised within some interior region.

For the second term in (3.21) we use a similar trick, now with the identity

$$\partial_j(J_j r_i r_k) = (\partial_j J_j) r_i r_k + J_i r_k + J_k r_i = J_i r_k + J_k r_i$$

Because $\mathbf{J}$ in (3.21) 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

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

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

$$\int d^3r'\ \mathbf{J}\,(\mathbf{r}\cdot\mathbf{r}') = \frac{1}{2}\mathbf{r}\times\int d^3r'\ \mathbf{J}\times\mathbf{r}' \tag{3.23}$$

With this in hand, we see that the long distance fall-off of any current distribution again takes the dipole form (3.19)

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

now with the magnetic dipole moment given by the integral,

$$\mathbf{m} = \frac{1}{2}\int d^3r'\ \mathbf{r}'\times\mathbf{J}(\mathbf{r}') \tag{3.24}$$

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

$$\mathbf{F} = q\mathbf{v}\times\mathbf{B}$$

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 (3.5). So if the charges in the second wire are moving with velocity $\mathbf{v}$, they will each experience a force

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

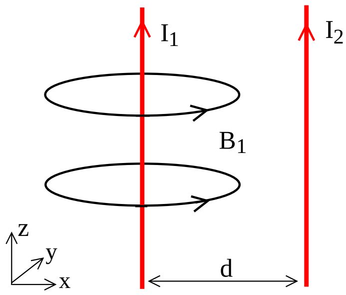


Figure 32:

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

$$\mathbf{J}_2 = nq\mathbf{v}$$

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

$$\mathbf{f} = nA\mathbf{F} = \left(\frac{\mu_0 I_1 I_2}{2\pi d}\right) \hat{\mathbf{z}} \times \hat{\mathbf{y}} = -\left(\frac{\mu_0 I_1 I_2}{2\pi d}\right) \hat{\mathbf{x}} \tag{3.25}$$

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.

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 (3.15) tells us that this can be written as

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

If the current $\mathbf{J}_1$ is localised on a curve C_1, then we can replace this volume integral with the line integral (3.16)

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0 I_1}{4\pi} \oint_{C_1} \frac{d\mathbf{r}_1 \times (\mathbf{r} - \mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|^3}$$

Now we place a second current distribution $\mathbf{J}_2$ in this magnetic field. It experiences a force per unit area given by (1.3), so the total force is

$$\mathbf{F} = \int d^3r \; \mathbf{J}_2(\mathbf{r}) \times \mathbf{B}(\mathbf{r}) \tag{3.26}$$

Again, if the current $\mathbf{J}_2$ is restricted to lie on a curve C_2, then this volume integral can be replaced by the line integral

$$\mathbf{F} = I_2 \oint_{C_2} d\mathbf{r} \times \mathbf{B}(\mathbf{r})$$

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

$$\mathbf{F} = \frac{\mu_0}{4\pi} I_1 I_2 \oint_{C_1} \oint_{C_2} d\mathbf{r}_2 \times \left(d\mathbf{r}_1 \times \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|^3} \right)$$

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

$$\mathbf{F} = \int_V d^3r \; \mathbf{J}(\mathbf{r}) \times \mathbf{B}(\mathbf{r})$$

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

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}(\mathbf{R}) + (\mathbf{r} \cdot \nabla)\mathbf{B}(\mathbf{R}) + \ldots$$

We then get the expression for the force

$$\mathbf{F} = -\mathbf{B}(\mathbf{R}) \times \int_V d^3r\ \mathbf{J}(\mathbf{r}) + \int_V d^3r\ \mathbf{J}(\mathbf{r}) \times [(\mathbf{r}\cdot\nabla)\mathbf{B}(\mathbf{R})] + \ldots$$

The first term vanishes because the currents have to go around in loops; we've already seen a proof of this following equation (3.21). We're going to do some fiddly manipulations with the second term. To help us remember that the derivative ∇ 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

$$\mathbf{F} = \int_V d^3r\ \mathbf{J}(\mathbf{r}) \times [(\mathbf{r}\cdot\nabla')\mathbf{B}(\mathbf{r}')]\Big|_{\mathbf{r}'=\mathbf{R}} \tag{3.27}$$

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

$$\mathbf{J}(\mathbf{r}) \times [(\mathbf{r}\cdot\nabla')\mathbf{B}(\mathbf{r}')] = -\nabla' \times [(\mathbf{r}\cdot\mathbf{B}(\mathbf{r}'))\mathbf{J}(\mathbf{r})]$$

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:

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

Or, subtracting one from the other,

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

But the terms in the brackets are the components of $\nabla\times\mathbf{B}$ and so vanish. So our result is true and we can rewrite the force (3.27) as

$$\mathbf{F} = -\nabla' \times \int_V d^3r\ (\mathbf{r}\cdot\mathbf{B}(\mathbf{r}'))\,\mathbf{J}(\mathbf{r})\Big|_{\mathbf{r}'=\mathbf{R}}$$

Now we need to manipulate this a little more. We make use of the identity (3.23) where we replace the constant vector by $\mathbf{B}$. Thus, up to some relabelling, (3.23) is the same as

$$\int_V d^3r\ (\mathbf{B}\cdot\mathbf{r})\mathbf{J} = \frac{1}{2}\mathbf{B} \times \int_V d^3r\ \mathbf{J}\times\mathbf{r} = -\mathbf{B}\times\mathbf{m}$$

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

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

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 $\nabla \cdot \mathbf{B} = 0$, this simplifies and we're left with a simple expression for the force on a dipole

$$\mathbf{F} = \nabla(\mathbf{B} \cdot \mathbf{m}) \tag{3.28}$$

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,

$$U = -\mathbf{B} \cdot \mathbf{m} \tag{3.29}$$

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 (3.29) 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.

The Force Between Dipoles

As a particular example of the force (3.28), 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 (3.24),

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3(\mathbf{m}_1 \cdot \hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}_1}{r^3}\right) \tag{3.30}$$

Now we'll consider how this affects the second dipole $\mathbf{m} = \mathbf{m}_2$. From (3.28), we have

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

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 (2.30). This is particularly pleasing because we used two rather different methods to calculate these forces. If we act with the derivative, we have

$$\mathbf{F} = \frac{3\mu_0}{4\pi r^4}\left[(\mathbf{m}_1 \cdot \hat{\mathbf{r}})\mathbf{m}_2 + (\mathbf{m}_2 \cdot \hat{\mathbf{r}})\mathbf{m}_1 + (\mathbf{m}_1 \cdot \mathbf{m}_2)\hat{\mathbf{r}} - 5(\mathbf{m}_1 \cdot \hat{\mathbf{r}})(\mathbf{m}_2 \cdot \hat{\mathbf{r}})\hat{\mathbf{r}}\right] \tag{3.31}$$

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 (3.31). 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 (3.31) 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 the picture. How are these related to our discussion above?

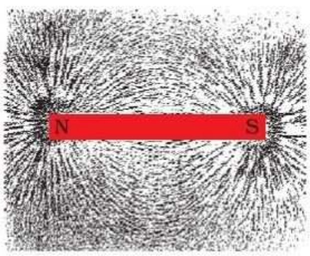


Figure 33:

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 (3.31). 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 (3.31) 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.

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:

$$s = \frac{1}{2}\hbar$$

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

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

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

$$g = 2$$

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

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

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 $\mathbf{m}$ to align themselves in a material. The details involve quantum mechanics and are beyond the scope of this course.

3.5 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 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[2], the Coulomb is defined in terms of the charge $-e$ carried by the electron. This is taken to be exactly

$$e = 1.602176634 \times 10^{-19}\ C$$

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 15 C. The total charge that passes through an AA battery in its lifetime is about 5000 C.

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 $1\ A$ to $10\ A$. There is a current in the Earth's atmosphere, known as the Birkeland current, which creates the aurora and varies between $10^5\ A$ and $10^6\ A$. 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 NC^{-1}. The electrostatic potential ϕ has units of *Volts*, denoted V, where the 1 Volt is the potential difference between two infinite, parallel plates which create an electric field of $1\ NC^{-1}$. A nerve cell sits at

[2]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 1 A 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 1 m, in order to experience an attractive force-per-unit-length of $2 \times 10^{-7}\ Nm^{-1}$. (Recall that a Newton is the unit of force needed to accelerate 1 Kg at 1 ms^{-1}.) From our result (3.25), we see that if we plug in $I_1 = I_2 = 1\ A$ and $d = 1\ m$ then this force is $f = \mu_0/2\pi\ A^2m^{-1}$. This definition is the reason that μ_0 has the strange-looking value $\mu_0 = 4\pi \times 10^{-7}\ m\,Kg\,C^{-2}$. The new definitions of SI units means that we can no longer say with certainty that $\mu_0 = 4\pi \times 10^{-7}\ m\,Kg\,C^{-2}$, 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.

around 10^{-2} V. An AA battery sits at 1.5 V. The largest man-made voltage is 10^7 V 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 V. 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 1 C, passing through a magnetic field of 1 T at 1 ms^{-1} will experience a force of 1 N. From the examples that we've seen above it's clear that 1 C is a lot of charge. Correspondingly, 1 T 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 100 T. Again, Nature beats us quite considerably. The magnetic field around neutron stars can be between 10^6 T and 10^9 T. (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*,

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

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 $e \approx 0.3$.

3.5.1 A History of Magnetostatics

The history of magnetostatics, like electrostatics, starts with the Greeks. The fact that magnetic iron ore, sometimes known as "lodestone", can attract pieces of iron was apparently known to Thales. He thought that he had found the soul in the stone. The word "magnetism" comes from the Greek town Magnesia, which is situated in an area rich in lodestone.

It took over 1500 years to turn Thales' observation into something useful. In the 11^{th} century, the Chinese scientist Shen Kuo realised that magnetic needles could be used to build a compass, greatly improving navigation.

The modern story of magnetism begins, as with electrostatics, with William Gilbert. From the time of Thales, it had been thought that electric and magnetic phenomenon are related. One of Gilbert's important discoveries was, ironically, to show that this is not the case: the electrostatic forces and magnetostatic forces are different.

Yet over the next two centuries, suspicions remained. Several people suggested that electric and magnetic phenomena are intertwined, although no credible arguments were given. The two just smelled alike. The following unisightful quote from Henry Elles, written in 1757 to the Royal Society, pretty much sums up the situation: "There are some things in the power of magnetism very similar to those of electricity. But I do not by any means think them the same". A number of specific relationships between electricity and magnetism were suggested and all subsequently refuted by experiment.

When the breakthrough finally came, it took everyone by surprise. In 1820, the Danish scientist Hans Christian Ørsted noticed that the needle on a magnet was deflected when a current was turned on or off. After that, progress was rapid. Within months, Ørsted was able to show that a steady current produces the circular magnetic field around a wire that we have seen in these lectures. In September that year, Ørsted's experiments were reproduced in front of the French Academy by Francois Arago, a talk which seemed to mobilise the country's entire scientific community. First out of the blocks were Jean-Baptiste Biot and Félix Savart who quickly determined the strength of the magnetic field around a long wire and the mathematical law which bears their name.

Of those inspired by Arago's talk, the most important was André-Marie Ampère. Skilled in both experimental and theoretical physics, Ampère determined the forces that arise between current carrying wires and derived the mathematical law which now bears his name: $\oint \mathbf{B} \cdot d\mathbf{r} = \mu_0 I$. He was also the first to postulate that there exists an atom of electricity, what we would now call the electron. Ampère's work was published in 1827 a book with the catchy title "*Memoir on the Mathematical Theory of Electrodynamic Phenomena, Uniquely Deduced from Experience*". It is now viewed as the beginning of the subject of electrodynamics.

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

> "I was at first almost frightened when I saw such mathematical force made to bear upon the subject, and then wondered to see that the subject stood it so well."
>
> *Faraday to Maxwell, 1857*

One of the Maxwell equations relates time varying magnetic fields to electric fields,

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{4.1}$$

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*.

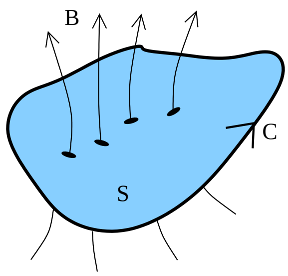


Figure 34:

We'll consider a wire to be a conductor, stretched along a stationary, closed curve, 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 S which is bounded by C,

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S}$$

By Stokes theorem, we can write this as

$$\int_C \mathbf{E} \cdot d\mathbf{r} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} = -\frac{d}{dt}\int_S \mathbf{B} \cdot d\mathbf{S}$$

Recall that the line integral around C should be in the right-handed sense; if the fingers on your right-hand curl around C then your thumb points in the direction of $d\mathbf{S}$. (This means that in the figure $d\mathbf{S}$ points in the same direction as $\mathbf{B}$). To get the last equality above, we need to use the fact that neither C nor S change with time. Both sides of this equation are usually given names. The integral of the electric field around the curve C is called the *electromotive force*, $\mathcal{E}$, or *emf* for short,

$$\mathcal{E} = \int_C \mathbf{E} \cdot d\mathbf{r}$$

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 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 S is called the magnetic *flux* Φ through S,

$$\Phi = \int_S \mathbf{B} \cdot d\mathbf{S}$$

The Maxwell equation (4.1) can be written as

$$\mathcal{E} = -\frac{d\Phi}{dt} \tag{4.2}$$

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 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 S; or you could replace the bar magnet with some other current density, restricted to a second wire C', and move that; or you could keep the second wire C' fixed and vary the current in it, perhaps turning it on and off. All of these will induce a current in C.

However, there is then a secondary effect. When a current flows in C, it will create its own magnetic field. We've seen how this works for steady currents in Section 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 (4.2).

We can illustrate this with a simple example. Consider the case where C is a circle, lying in a plane. We'll place it in a uniform B field and then make 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 C as shown. 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 the figure. You see that the induced current causes $\mathbf{B}$ to increase inside the loop, counteracting the original decrease.

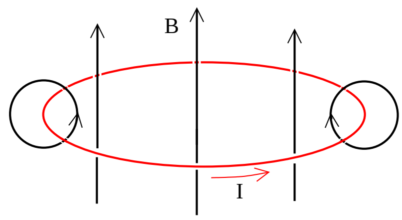


Figure 35: Lenz's law

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.

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 the figure: 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.

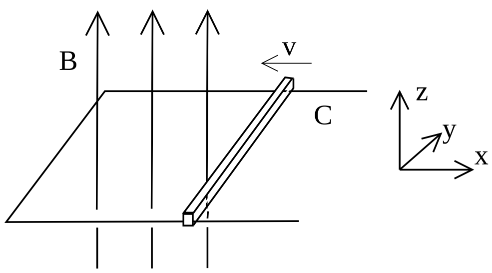


Figure 36: Moving circuit

Slide the bar to the left with speed v. Each charge q in the bar experiences a Lorentz force qvB, pushing it in the $\mathbf{y}$ direction. This results in an emf which, now, is defined as the integrated force per charge. In this case, the resulting emf is

$$\mathcal{E} = vBd$$

where d is the length of the moving bar. But, because the area inside the circuit is getting smaller, the flux through C is also decreasing. In this case, it's simple to compute the change of flux: it is

$$\frac{d\Phi}{dt} = -vBd$$

We see that once again the change of flux is related to the emf through the flux rule

$$\mathcal{E} = -\frac{d\Phi}{dt}$$

Note that this is the same formula (4.2) 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 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.

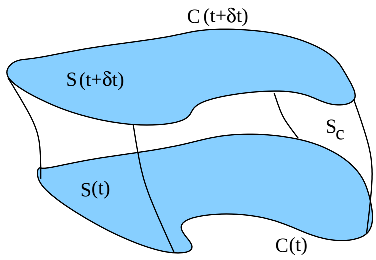


Figure 37: Moving Circuits

The General Case

There is a nice way to include both the effects of time-dependent magnetic fields and the possibility that the circuit C changes with time. We consider the moving loop $C(t)$, as shown in the figure. Now the change in flux through a surface S has two terms: one because B may be changing, and one because C is changing. In a small time δt, we have

$$\begin{aligned} \delta\Phi = \Phi(t+\delta t) - \Phi(t) &= \int_{S(t+\delta t)} \mathbf{B}(t+\delta t)\cdot d\mathbf{S} - \int_{S(t)} \mathbf{B}(t)\cdot d\mathbf{S} \\ &= \int_{S(t)} \frac{\partial \mathbf{B}}{\partial t}\,\delta t\cdot d\mathbf{S} + \left[\int_{S(t+\delta t)} - \int_{S(t)}\right]\mathbf{B}(t)\cdot d\mathbf{S} + \mathcal{O}(\delta t^2) \end{aligned}$$

We can do something with the middle terms. Consider the closed surface created by $S(t)$ and $S(t+\delta t)$, together with the cylindrical region swept out by $C(t)$ which we call S_c. Because $\nabla\cdot\mathbf{B} = 0$, the integral of $\mathbf{B}(t)$ over any closed surface vanishes. But

$\int_{S(t+\delta t)} - \int_{S(t)}$ is the top and bottom part of the closed surface, with the minus sign just ensuring that the integral over the bottom part $S(t)$ is in the outward direction. This means that we must have

$$\left[\int_{S(t+\delta t)} - \int_{S(t)}\right] \mathbf{B}(t) \cdot d\mathbf{S} = -\int_{S_c} \mathbf{B}(t) \cdot d\mathbf{S}$$

For the integral over S_c, we can write the surface element as

$$d\mathbf{S} = (d\mathbf{r} \times \mathbf{v})\delta t$$

where $d\mathbf{r}$ is the line element along $C(t)$ and $\mathbf{v}$ is the velocity of a point on C. We find that the expression for the change in flux can be written as

$$\frac{d\Phi}{dt} = \lim_{\delta t \to 0} \frac{\delta \Phi}{\delta t} = \int_{S(t)} \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} - \int_{C(t)} (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r}$$

where we've taken the liberty of rewriting $(d\mathbf{r} \times \mathbf{v}) \cdot \mathbf{B} = d\mathbf{r} \cdot (\mathbf{v} \times \mathbf{B})$. 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

$$\frac{d\Phi}{dt} = -\int_C (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r}$$

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,

$$\mathcal{E} = \int_C (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r}$$

and we once again get the flux rule $\mathcal{E} = -d\Phi/dt$.

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 Section 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 C. From the results of Section 3 we know that this gives rise to a magnetic field and hence a flux $\Phi = \int_S \mathbf{B} \cdot d\mathbf{S}$ through the surface S bounded by 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 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.

Inductance

If a current I flowing around a curve C gives rise to a flux $\Phi = \int_S \mathbf{B} \cdot d\mathbf{S}$ then the *inductance* L of the circuit is defined to be

$$L = \frac{\Phi}{I}$$

The inductance is a property only of our choice of curve C.

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 (3.7)

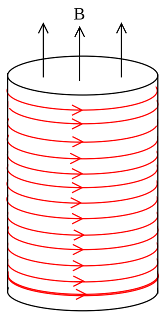


Figure 38:

$$B = \mu_0 I N$$

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

$$\Phi = \mu_0 I N^2 A l = \mu_0 I N^2 V$$

with $V = Al$ the volume inside the solenoid. The inductance of the solenoid is therefore

$$L = \mu_0 N^2 V$$

Magnetostatic Energy

The definition of inductance is useful to derive the energy stored in the magnetic field. Let's take our circuit C with current I. We'll try to increase the current. The induced emf is

$$\mathcal{E} = -\frac{d\Phi}{dt} = -L\frac{dI}{dt}$$

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

$$\delta W = \mathcal{E} I \delta t = -LI\frac{dI}{dt}\delta t \quad \Rightarrow \quad \frac{dW}{dt} = -LI\frac{dI}{dt} = -\frac{L}{2}\frac{dI^2}{dt}$$

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

$$W = \frac{1}{2}LI^2 = \frac{1}{2}I\Phi$$

Following our discussion for electric energy in (2.3), we identify this with the energy U stored in the system. We can write it as

$$U = \frac{1}{2}I\int_S \mathbf{B}\cdot d\mathbf{S} = \frac{1}{2}I\int_S \nabla\times\mathbf{A}\cdot d\mathbf{S} = \frac{1}{2}I\oint_C \mathbf{A}\cdot d\mathbf{r} = \frac{1}{2}\int d^3x\ \mathbf{J}\cdot\mathbf{A}$$

where, in the last step, we've used the fact that the current density $\mathbf{J}$ is localised on the curve C to turn the integral into one over all of space. At this point we turn to the Maxwell equation $\nabla\times\mathbf{B} = \mu_0\mathbf{J}$ to write the energy as

$$U = \frac{1}{2\mu_0}\int d^3x\ (\nabla\times\mathbf{B})\cdot\mathbf{A} = \frac{1}{2\mu_0}\int d^3x\ \ [\nabla\cdot(\mathbf{B}\times\mathbf{A}) + \mathbf{B}\cdot(\nabla\times\mathbf{A})]$$

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

$$U = \frac{1}{2\mu_0}\int d^3x\ \mathbf{B}\cdot\mathbf{B}$$

Combining this with our previous result (2.27) for the electric field, we have the energy stored in the electric and magnetic fields,

$$U = \int d^3x\ \left(\frac{\epsilon_0}{2}\mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\mathbf{B}\cdot\mathbf{B}\right) \tag{4.3}$$

This is a nice result. But there's something a little unsatisfactory behind our derivation of (4.3). 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.

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 the next course.

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

$$\mathcal{E} = IR \tag{4.4}$$

The constant of proportionality R is called the resistance. The emf is $\mathcal{E} = \int \mathbf{E} \cdot d\mathbf{x}$. If we write $\mathbf{E} = -\nabla\phi$, 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 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

$$\mathbf{J} = \sigma \mathbf{E}$$

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 (4.4) 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.

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.

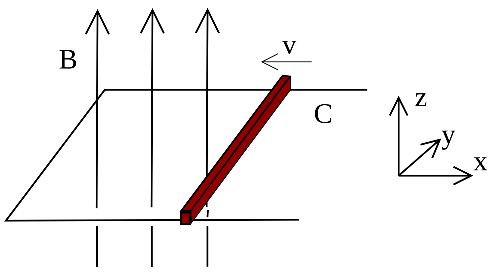


Figure 39:

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

$$\mathbf{F} = IBd\,\hat{\mathbf{x}}$$

The equation of motion for the position of the wire is then

$$m\ddot{x} = IBd$$

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

$$\mathcal{E} = -\frac{d\Phi}{dt} = -Bd\dot{x}$$

Ohm's law tells us that $\mathcal{E} = IR$. Combining these, we get a simple differential equation for the position of the bar

$$m\ddot{x} = -\frac{B^2 d^2}{R}\,\dot{x}$$

which we can solve to see that any initial velocity of the bar, v, decays exponentially:

$$\dot{x}(t) = -ve^{-B^2 d^2 t/mR}$$

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 $\mathbf{B}$ 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

$$\mathcal{E} = \mathcal{E}_0 + \mathcal{E}_{\text{induced}} = \mathcal{E}_0 - Bd\dot{x}$$

The total current is again given by Ohms law $I = \mathcal{E}/R$. The position of the bar is now governed by the equation

$$m\ddot{x} = -\frac{Bd}{R}\,(\mathcal{E}_0 - Bd\dot{x})$$

Again, it's simple to solve this equation.

Joule Heating

In Section 4.1.2, we computed the work done in changing the current in a circuit 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 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 C is $\mathcal{E}_0$ which means that amount of work necessary to keep a current I moving for time δt is

$$\delta W = \mathcal{E}_0 I \delta t = I^2 R\, \delta t$$

We learn that the power (work per unit time) dissipated by a current passing through a circuit of resistance R is $dW/dt = I^2 R$. This is not energy that can be usefully stored like the magnetic and electric energy (4.3); 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*.

4.1.4 Michael Faraday (1791-1867)

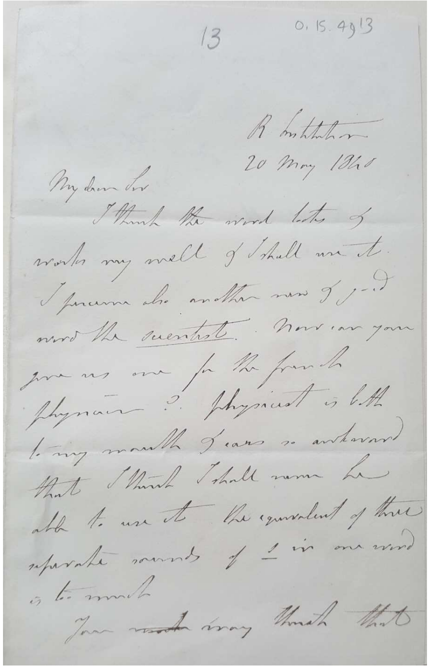
> "The word "physicist" is both to my mouth and ears so awkward that I think I shall never be able to use it. The equivalent of three separate sounds of "s" in one word is too much."
>
> *Faraday in a letter to William Whewell* [3]

Michael Faraday's route into science was far from the standard one. The son of a blacksmith, he had little schooling and, at the age of 14, was apprenticed to a bookbinder. There he remained until the age of 20 when Faraday attended a series of popular lectures at the Royal Institution by the chemist Sir Humphry Davy. Inspired, Faraday wrote up these lectures, lovingly bound them and presented them to Davy as a gift. Davy was impressed and some months later, after suffering an eye injury in an explosion, turned to Faraday to act as his assistant.

Not long after, Davy decided to retire and take a two-year leisurely tour of Europe, meeting many of the continent's top scientists along the way. He asked Faraday to join him and his wife, half as assistant, half as valet. The science part of this was a success; the valet part less so. But Faraday dutifully played his roles, emptying his master's chamber pot each morning, while aiding in a number of important scientific discoveries along the way, including a wonderful caper in Florence where Davy and Faraday used Galileo's old lens to burn a diamond, reducing it, for the first time, to Carbon.

[3] According to the rest of the internet, Faraday complains about three separate sounds of "i". The rest of the internet is wrong and can't read Faraday's writing. The original letter is in the Wren library in Trinity College and is shown on the next page. I'm grateful to Frank James, editor of Faraday's correspondence, for help with this.

Back in England, Faraday started work at the Royal Institution. He would remain there for over 45 years. An early attempt to study electricity and magnetism was abandoned after a priority dispute with his former mentor Davy and it was only after Davy's death in 1829 that Faraday turned his attentions fully to the subject. He made his discovery of induction on 28th October, 1831. The initial experiment involved two, separated coils of wire, both wrapped around the same magnet. Turning on a current in one wire induces a momentary current in the second. Soon after, he found that a current is also induced by passing a loop of wire over a magnet. The discovery of induction underlies the electrical dynamo and motor, which convert mechanical energy into electrical energy and vice-versa.


13
0.15.49[13]

R Institution
20 May 1840

My dear Sir
I think the word [illegible] of
works very well & I shall use it.
I perceive also another new & good
word the scientist. Now can you
give us one for the french
physicien? Physicist is both
to my mouth & ears so awkward
that I think I shall never be
able to use it the equivalent of three
separate sounds of i in one word
is too much
Your ~~[illegible]~~ very thankful ~~that~~

Figure 40:

Faraday was not a great theorist and the mathematical expression that we have called Faraday's law is due to Maxwell. Yet Faraday's intuition led him to make one of the most important contributions of all time to theoretical physics: he was the first to propose the idea of the field.

As Faraday's research into electromagnetism increased, he found himself lacking the vocabulary needed to describe the phenomena he was seeing. Since he didn't exactly receive a classical education, he turned to William Whewell, then Master of Trinity, for some advice. Between them, they cooked up the words 'anode', 'cathode', 'ion', 'dielectric', 'diamagnetic' and 'paramagnetic'. They also suggested the electric charge be renamed 'Franklinic' in honour of Benjamin Franklin. That one didn't stick.

The last years of Faraday's life were spent in the same way as Einstein: seeking a unified theory of gravity and electromagnetism. The following quote describes what is, perhaps, the first genuine attempt at unification:

> Gravity: Surely his force must be capable of an experimental relation to Electricity, Magnetism and the other forces, so as to bind it up with them in reciprocal action and equivalent effect. Consider for a moment how to set about touching this matter by facts and trial . . .
>
> *Faraday, 19th March, 1849.*

As this quote makes clear, Faraday's approach to this problem includes something that Einstein's did not: experiment. Ultimately, neither of them found a connection between electromagnetism and gravity. But it could be argued that Faraday made the more important contribution: while a null theory is useless, a null experiment tells you something about Nature.

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)

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{4.5}$$

and the law that says there are no magnetic monopoles

$$\nabla \cdot \mathbf{B} = 0 \tag{4.6}$$

and Ampère's law

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{4.7}$$

and now also Faraday's law

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{4.8}$$

In fact, there's only one term left to discuss. When fields change with time, there is an extra term that appears in Ampère's law, which reads in full:

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \tag{4.9}$$

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 (4.5), (4.6), (4.7) and (4.8) — which include Ampère'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 (4.5) to (4.8) 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 Ampère'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 Ampère law (4.7) is inconsistent. We simply need to take the divergence to find

$$\mu_0 \nabla \cdot \mathbf{J} = \nabla \cdot (\nabla \times \mathbf{B}) = 0$$

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 Ampère's law.

There is a standard thought experiment involving circuits which is usually invoked to demonstrate the need to amend Ampère's law. This is shown in the figure. 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.

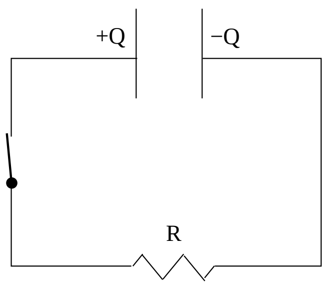


Figure 41:

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 Ampère's law. We can take a curve C that surrounds the wire and surface S with boundary C. If we chose 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

$$\int_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 I \tag{4.10}$$

where I is the current through the wire which, in this case, is changing with time.

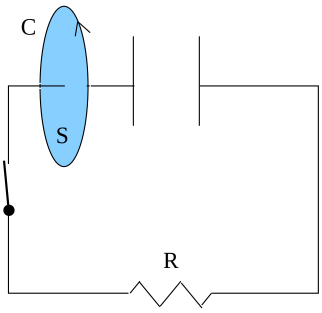


Figure 42: This choice of surface suggests there is a magnetic field

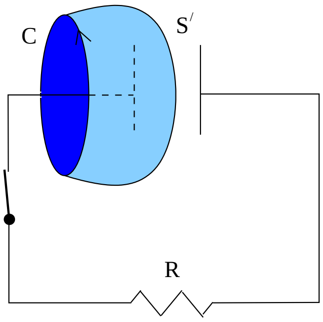


Figure 43: This choice of surface suggests there is none.

Suppose, however, that we instead decided to bound the curve C with the surface S', which now sneaks through the gap between the capacitor plates. Now there is no current passing through S', so if we were to use Ampère's law, we would conclude that there is no magnetic field

$$\int_C \mathbf{B} \cdot d\mathbf{r} = 0 \tag{4.11}$$

This is in contradiction to our first calculation (4.10). So what's going on here? Well, Ampère'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.

Adding the Displacement Current

Let's now see how adding the displacement current (4.9) fixes the situation. We'll first look at the abstract issue that Ampère's law requires $\nabla \cdot \mathbf{J} = 0$. If we add the displacement current, then taking the divergence of (4.9) gives

$$\mu_0 \left(\nabla \cdot \mathbf{J} + \epsilon_0 \nabla \cdot \frac{\partial \mathbf{E}}{\partial t} \right) = \nabla \cdot (\nabla \times \mathbf{B}) = 0$$

But, using Gauss's law, we can write $\epsilon_0 \nabla \cdot \mathbf{E} = \rho$, so the equation above becomes

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0$$

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 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 (2.10): it is

$$E = \frac{Q}{\epsilon_0 A}$$

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

$$\frac{\partial E}{\partial t} = \frac{1}{\epsilon_0 A}\frac{dQ}{dt} = \frac{1}{\epsilon_0 A} I(t)$$

So now if we repeat the calculation of $\mathbf{B}$ using the surface S', we find an extra term from (4.9) which gives

$$\int_C \mathbf{B} \cdot d\mathbf{r} = \int_{S'} \mu_0 \epsilon_0 \frac{\partial E}{\partial t} = \mu_0 I$$

This is the same answer (4.10) that we found using Ampère's law applied to the surface 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:

$$\nabla \cdot \mathbf{E} = 0 \qquad \text{and} \qquad \nabla \times \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$
$$\nabla \cdot \mathbf{B} = 0 \qquad \text{and} \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

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,

$$\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} = \frac{\partial}{\partial t}(\nabla\times\mathbf{B}) = \nabla\times\frac{\partial\mathbf{B}}{\partial t} = -\nabla\times(\nabla\times\mathbf{E}) \tag{4.12}$$

To complete the derivation, we need the identity

$$\nabla\times(\nabla\times\mathbf{E}) = \nabla(\nabla\cdot\mathbf{E}) - \nabla^2\mathbf{E}$$

But, the first of Maxwell equations tells us that $\nabla\cdot\mathbf{E} = 0$ in vacuum, so the first term above vanishes. We find that each component of the electric field satisfies,

$$\frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} - \nabla^2\mathbf{E} = 0 \tag{4.13}$$

This is the wave equation. The speed of the waves, c, is given by

$$c = \sqrt{\frac{1}{\mu_0\epsilon_0}}$$

Identical manipulations hold for the magnetic field. We have

$$\frac{\partial^2\mathbf{B}}{\partial t^2} = -\frac{\partial}{\partial t}(\nabla\times\mathbf{E}) = -\nabla\times\frac{\partial\mathbf{E}}{\partial t} = -\frac{1}{\mu_0\epsilon_0}\nabla\times(\nabla\times\mathbf{B}) = \frac{1}{\mu_0\epsilon_0}\nabla^2\mathbf{B}$$

where, in the last equality, we have made use of the vector identity (4.12), now applied to the magnetic field $\mathbf{B}$, together with the Maxwell equation $\nabla\cdot\mathbf{B} = 0$. We again find that each component of the magnetic field satisfies the wave equation,

$$\frac{1}{c^2}\frac{\partial^2\mathbf{B}}{\partial t^2} - \nabla^2\mathbf{B} = 0 \tag{4.14}$$

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

$$\epsilon_0 = 8.854187817\times 10^{-12}\ m^{-3}\,Kg^{-1}\,s^2\,C^2$$

and the magnetic constant,

$$\mu_0 = 4\pi\times 10^{-7}\ m\,Kg\,C^{-2}$$

Plugging in these numbers gives the speed of electric and magnetic waves to be

$$c = 299792458\ ms^{-1}$$

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. In the words of the man himself

"The velocity of transverse undulations in our hypothetical medium, calculated from the electro-magnetic experiments of MM. Kohlrausch and Weber, agrees so exactly with the velocity of light calculated from the optical experiments of M. Fizeau, that we can scarcely avoid the inference that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena"

James Clerk Maxwell

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 $\nabla \cdot \mathbf{E} = 0$ 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

$$\mathbf{E} = (0, E(x,t), 0)$$

where E satisfies the wave equation (4.13) which is now

$$\frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} - \nabla^2 E = 0$$

The most general solution to the wave equation takes the form

$$E(x,t) = f(x - ct) + g(x + ct)$$

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

$$E = E_0 \sin\left[\omega\left(\frac{x}{c} - t\right)\right]$$

We usually write this as

$$E = E_0 \sin\left(kx - \omega t\right) \tag{4.15}$$

where k is the *wavenumber*. The wave equation (4.13) requires that it is related to the frequency by

$$\omega^2 = c^2 k^2$$

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 $\lambda \sim 3.9 \times 10^{-7}\ m$ and $7 \times 10^{-7}\ m$. At one end of the spectrum, gamma rays have wavelength $\lambda \sim 10^{-12}\ m$ and X-rays around $\lambda \sim 10^{-10}$ to $10^{-8}\ m$. At the other end, radio waves have $\lambda \sim 1\ cm$ to $10\ km$. 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 $\nabla \cdot \mathbf{B} = 0$ 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 (4.14). But their behaviour is dictated by what the electric field is doing through the Maxwell equation $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$. This tells us that

$$\mathbf{B} = (0, 0, B)$$

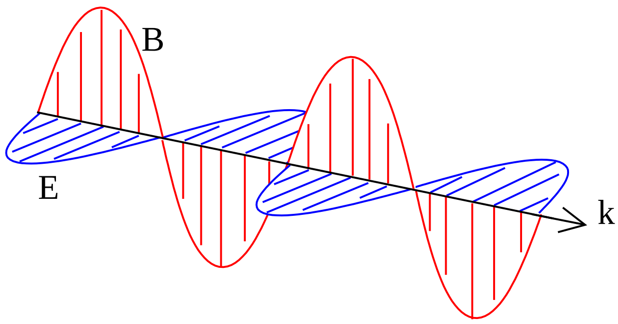
with

$$\frac{\partial B}{\partial t} = -\frac{\partial E}{\partial x} = -kE_0 \cos(kx - \omega t)$$

We find

$$B = \frac{E_0}{c} \sin(kx - \omega t) \tag{4.16}$$

We see that the electric $\mathbf{E}$ and magnetic $\mathbf{B}$ fields oscillate in phase, but in orthogonal directions. And both oscillate in directions which are orthogonal to the direction in which the wave travels.



Because the Maxwell equations are linear, we're allowed to add any number of solutions of the form (4.15) and (4.16) 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 (4.15) and (4.16) in complex notation,

$$\mathbf{E} = E_0\, \hat{\mathbf{y}}\, e^{i(kx-\omega t)} \quad , \quad \mathbf{B} = \frac{E_0}{c}\, \hat{\mathbf{z}}\, e^{i(kx-\omega t)} \tag{4.17}$$

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 $\mathbf{E}$ and $\mathbf{B}$ solve the Maxwell equations. And, more importantly, if we start adding complex $\mathbf{E}$ and $\mathbf{B}$ 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 $\mathbf{E}$ and $\mathbf{B}$ — 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,

$$\mathbf{E} = \mathbf{E}_0\, e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \quad \text{and} \quad \mathbf{B} = \mathbf{B}_0\, e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \tag{4.18}$$

where, for now, both $\mathbf{E}_0$ and $\mathbf{B}_0$ could be complex-valued vectors. (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 (4.18) already satisfy the wave equations (4.13) and (4.14) if ω and $\mathbf{k}$ obey the dispersion relation $\omega^2 = c^2k^2$.

We get further constraints on $\mathbf{E}_0$, $\mathbf{B}_0$ and $\mathbf{k}$ from the original Maxwell equations. These are

$$\begin{aligned} \nabla\cdot\mathbf{E} = 0 \quad &\Rightarrow \quad i\mathbf{k}\cdot\mathbf{E}_0 = 0 \\ \nabla\cdot\mathbf{B} = 0 \quad &\Rightarrow \quad i\mathbf{k}\cdot\mathbf{B}_0 = 0 \\ \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad &\Rightarrow \quad i\mathbf{k}\times\mathbf{E}_0 = i\omega\mathbf{B}_0 \end{aligned}$$

Let's now interpret these equations:

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 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

$$\hat{\mathbf{k}} \times (\mathbf{E}_0/c) = \mathbf{B}_0$$

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.

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 (4.18), but now the polarisation does not point in a fixed direction. To see this, write

$$\mathbf{E}_0 = \boldsymbol{\alpha} - i\boldsymbol{\beta}$$

The real part of the electric field is then

$$\mathbf{E} = \boldsymbol{\alpha}\cos(\mathbf{k}\cdot\mathbf{x} - \omega t) + \boldsymbol{\beta}\sin(\mathbf{k}\cdot\mathbf{x} - \omega t)$$

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 $\mathbf{k}$).

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*.

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,

$$\mathbf{E}(\mathbf{x},t) = \int \frac{d^3k}{(2\pi)^3}\, \mathbf{E}(\mathbf{k})\, e^{i(\mathbf{k}\cdot\mathbf{x} - \omega t)}$$

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 means an incident plane wave, travelling in the x-direction,

$$\mathbf{E}_{\rm inc} = E_0\,\hat{\mathbf{y}}\,e^{i(kx-\omega t)}$$

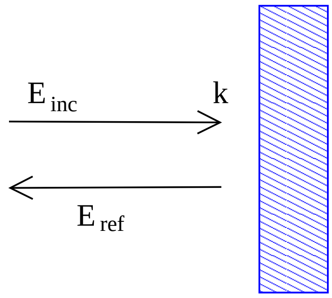


Figure 44:

where, as before, $\omega = ck$. Inside the conductor, we know that we must have $\mathbf{E} = 0$. 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

$$\mathbf{E}_{\rm ref} = -E_0\,\hat{\mathbf{y}}\,e^{i(-kx-\omega t)}$$

So that the combination $\mathbf{E} = \mathbf{E}_{\rm inc} + \mathbf{E}_{\rm ref}$ satisfies $E(x = 0) = 0$ as it must. This is illustrated in the figure. (Note, however, that the figure 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).

We've already seen above that the corresponding magnetic field can be determined by $\nabla\times\mathbf{E} = -\partial\mathbf{B}/\partial t$. It is given by $\mathbf{B} = \mathbf{B}_{\rm inc} + \mathbf{B}_{\rm ref}$, with

$$\mathbf{B}_{\rm inc} = \frac{E_0}{c}\,\hat{\mathbf{z}}\,e^{i(kx-\omega t)} \quad \text{and} \quad \mathbf{B}_{\rm ref} = \frac{E_0}{c}\,\hat{\mathbf{z}}\,e^{i(-kx-\omega t)} \tag{4.19}$$

This obeys $\mathbf{B}\cdot\mathbf{n} = 0$, as it should be continuity. But the tangential component doesn't vanish at the surface. Instead, we have

$$\mathbf{B}\cdot\hat{\mathbf{z}}\big|_{x=0^-} = \frac{2E_0}{c}\,e^{-i\omega t}$$

Since the magnetic field vanishes inside the conductor, we have a discontinuity. But there's no mystery here. We know from our previous discussion (3.6) that this corresponds to a surface current $\mathbf{K}$ induced by the wave

$$\mathbf{K} = \frac{2E_0}{c\mu_0}\,\hat{\mathbf{y}}\,e^{-i\omega t}$$

We see that the surface current oscillates with the frequency of the reflected wave.

Reflection at an Angle

Let's now try something a little more complicated: we'll send in the original ray at an angle, θ, to the normal as shown in the figure. Our incident electric field is

$$\mathbf{E}_{\rm inc} = E_0\,\hat{\mathbf{y}}\,e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$$

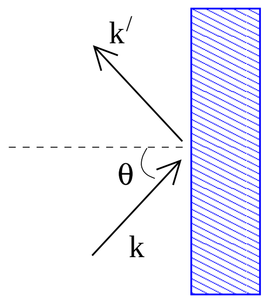


Figure 45:

where

$$\mathbf{k} = k\cos\theta\,\hat{\mathbf{x}} + k\sin\theta\,\hat{\mathbf{z}}$$

Notice that we've made a specific choice for the polarisation of the electric field: it is out of the page in the figure, tangential to the surface. Now we have two continuity conditions to worry about. We want to add a reflected wave,

$$\mathbf{E}_{\rm ref} = -E_0\,\hat{\boldsymbol{\zeta}}\,e^{i(\mathbf{k}'\cdot\mathbf{x}-\omega' t)}$$

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

$$(\mathbf{E}_{\rm inc} + \mathbf{E}_{\rm ref})\cdot\hat{\mathbf{n}} = 0 \quad \text{ and } \quad (\mathbf{E}_{\rm inc} + \mathbf{E}_{\rm ref})\times\hat{\mathbf{n}} = 0$$

where, for this set-up, the normal vector is $\hat{\mathbf{n}} = -\hat{\mathbf{x}}$. This is achieved by taking $\omega' = \omega$ and $\boldsymbol{\zeta} = \hat{\mathbf{y}}$, so that the reflected wave changes neither frequency nor polarisation. The reflected wavevector is

$$\mathbf{k}' = -k\cos\theta\,\hat{\mathbf{x}} + k\sin\theta\,\hat{\mathbf{z}}$$

We can also check what becomes of the magnetic field. It is $\mathbf{B} = \mathbf{B}_{\rm inc} + \mathbf{B}_{\rm ref}$, with

$$\mathbf{B}_{\rm inc} = \frac{E_0}{c}\,(\hat{\mathbf{k}}\times\hat{\mathbf{y}})\,e^{i(\mathbf{k}\cdot\mathbf{x}-\omega' t)} \quad \text{ and } \quad \mathbf{B}_{\rm ref} = -\frac{E_0}{c}\,(\hat{\mathbf{k}}'\times\hat{\mathbf{y}})\,e^{i(\mathbf{k}'\cdot\mathbf{x}-\omega' t)}$$

Note that, in contrast to (4.19), there is now a minus sign in the reflected $\mathbf{B}_{\rm ref}$, but this is simply to absorb a second minus sign coming from the appearance of $\hat{\mathbf{k}}'$ in the polarisation 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.

4.3.4 James Clerk Maxwell (1831-1879)

Still those papers lay before me,
Problems made express to bore me,
When a silent change came o'er me,
In my hard uneasy chair.
Fire and fog, and candle faded,
Spectral forms the room invaded,
Little creatures, that paraded
On the problems lying there.

James Clerk Maxwell, "A Vision of a Wrangler, of a University, of Pedantry, and of Philosophy"

James Clerk Maxwell was a very smart man. Born in Edinburgh, he was a student, first in his hometown, and later in Cambridge, at Peterhouse and then at Trinity. He held faculty positions at the University of Aberdeen (where they fired him) and Kings College London before returning to Cambridge as the first Cavendish professor of physics.

Perhaps the first very smart thing that Maxwell did was to determine the composition of Saturn's rings. He didn't do this using a telescope. He did it using mathematics! He showed that neither a solid nor a fluid ring could be stable. Such rings could only be made of many small particles. For this he was awarded the Adams Prize. (These days you can win this prize for much much less!)

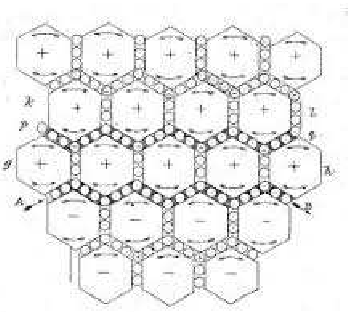


Figure 46: Maxwell's vortices

Maxwell's great work on electromagnetism was accomplished between 1861 and 1862. He started by constructing an elaborate mechanical model of electricity and magnetism in which space is filled by vortices of an incompressible fluid, separated by tiny rotating particles that give rise to electricity. One of his illustrations is shown above. Needless to say, we don't teach this picture of space anymore. From this, he managed to distill everything that was known about electromagnetism into 20 coupled equations in 20 variables. This was the framework in which he discovered the displacement current and its consequences for light.

You might think that the world changed when Maxwell published his work. In fact, no one cared. The equations were too hard for physicists, the physics too hard for mathematicians. Things improved marginally in 1873 when Maxwell reduced his equations to just four, albeit written in quaternion notation. The modern version of Maxwell equations, written in vector calculus notation, is due to Oliver Heaviside in 1881. In all, it took almost 30 years for people to appreciate the significance of Maxwell's acheivement.

Maxwell made a number of other important contributions to science, including the first theory of colour vision and the theory of colour photography. His work on thermodynamics and statistical mechanics deserves at least equal status with his work on electromagnetism. He was the first to understand the distribution of velocities of molecules in a gas, the first to extract an experimental prediction from the theory of atoms and, remarkably, the first (with the help of his wife) to build the experiment and do the measurement, confirming his own theory.

4.4 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 (4.3) for the energy stored in electric and magnetic fields,

$$U = \int_V d^3x \ \left(\frac{\epsilon_0}{2}\,\mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\,\mathbf{B}\cdot\mathbf{B}\right)$$

The expression in brackets is the energy density. Here we have integrated this only over some finite volume 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

$$\begin{aligned}\frac{dU}{dt} &= \int_V d^3x \ \left(\epsilon_0\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t} + \frac{1}{\mu_0}\mathbf{B}\cdot\frac{\partial\mathbf{B}}{\partial t}\right)\\ &= \int_V d^3x \ \left(\frac{1}{\mu_0}\mathbf{E}\cdot(\nabla\times\mathbf{B}) - \mathbf{E}\cdot\mathbf{J} - \frac{1}{\mu_0}\mathbf{B}\cdot(\nabla\times\mathbf{E})\right)\end{aligned}$$

where we've used the two Maxwell equations. Now we use the identity

$$\mathbf{E}\cdot(\nabla\times\mathbf{B})-\mathbf{B}\cdot(\nabla\times\mathbf{E})=-\nabla\cdot(\mathbf{E}\times\mathbf{B})$$

and write

$$\frac{dU}{dt}=-\int_V d^3x\;\mathbf{J}\cdot\mathbf{E}-\frac{1}{\mu_0}\int_S(\mathbf{E}\times\mathbf{B})\cdot d\mathbf{S} \tag{4.20}$$

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 $\int_V d^3x\,\mathbf{J}\cdot\mathbf{E}$ 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 V. Since it appears with a minus sign in (4.20), it is the rate of loss of energy of the particles.

Now we can interpret (4.20). If we write it as

$$\frac{dU}{dt}+\int_V d^3x\;\mathbf{J}\cdot\mathbf{E}=-\frac{1}{\mu_0}\int_S(\mathbf{E}\times\mathbf{B})\cdot d\mathbf{S}$$

then the left-hand side is the combined change in energy of both fields and particles in region V. Since energy is conserved, the right-hand side must describe the energy that escapes through the surface S of region V. We define the *Poynting vector*

$$\mathbf{S}=\frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}$$

This is a vector field. It tells us the magnitude and direction of the flow of energy in any point in space. (It is unfortunate that the canonical name for the Poynting vector is $\mathbf{S}$ because it makes it notationally difficult to integrate over a surface which we usually also like to call $\mathbf{S}$. Needless to say, these two things are not the same and hopefully no confusion will arise).

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 (4.17), but with arbitrary wavevector $\mathbf{k}$,

$$\mathbf{E}=\mathbf{E}_0\sin(\mathbf{k}\cdot\mathbf{x}-\omega t)\quad\text{ and }\quad\mathbf{B}=\frac{1}{c}(\hat{\mathbf{k}}\times\mathbf{E}_0)\sin(\mathbf{k}\cdot\mathbf{x}-\omega t)$$

The Poynting vector is then

$$\mathbf{S} = \frac{E_0^2}{c\mu_0}\,\hat{\mathbf{k}}\,\sin^2(\mathbf{k}\cdot\mathbf{x} - \omega t)$$

Averaging over a period, $T = 2\pi/\omega$, we have

$$\bar{\mathbf{S}} = \frac{E_0^2}{2c\mu_0}\,\hat{\mathbf{k}}$$

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 (4.3). Evaluated on the electromagnetic wave, the energy density is

$$u = \frac{\epsilon_0}{2}\,\mathbf{E}\cdot\mathbf{E} + \frac{1}{2\mu_0}\,\mathbf{B}\cdot\mathbf{B} = \epsilon_0 E_0^2 \sin^2(\mathbf{k}\cdot\mathbf{x} - \omega t)$$

Averaged over a period $T = 2\pi/\omega$, this is

$$\bar{u} = \frac{\epsilon_0 E_0^2}{2}$$

Then, using $c^2 = 1/\epsilon_0\mu_0$,we can write

$$\bar{\mathbf{S}} = c\bar{u}\hat{\mathbf{k}}$$

The interpretation is simply that the energy $\bar{\mathbf{S}}$ is equal to the energy density in the wave $\bar{u}$ times the speed of the wave, c.

4.4.1 The Continuity Equation Revisited

Recall that, way back in Section 1, we introduced the continuity equation for electric charge,

$$\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{J} = 0$$

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 (4.20) is simply the integrated version of a continuity equation for energy. We could equally well write it as

$$\frac{\partial\mathcal{E}}{\partial t} + \nabla\cdot\mathbf{S} = 0$$

We see that the Poynting vector $\mathbf{S}$ is to energy what the current $\mathbf{J}$ is to charge.

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.

5.1 A Review of Special Relativity

We start with a very quick review of the relevant concepts of special relativity. (For more details see the lecture notes on Dynamics and Relativity). The basic postulate of relativity is that the laws of physics are the same in all inertial reference frames. The guts of the theory tell us how things look to observers who are moving relative to each other.

The first observer sits in an inertial frame $\mathcal{S}$ with spacetime coordinates (ct, x, y, z) the second observer sits in an inertial frame $\mathcal{S}'$ with spacetime coordinates (ct', x', y', z').

If we take $\mathcal{S}'$ to be moving with speed v in the x-direction relative to $\mathcal{S}$ then the coordinate systems are related by the Lorentz boost

$$x' = \gamma\left(x - \frac{v}{c}\,ct\right) \quad \text{and} \quad ct' = \gamma\left(ct - \frac{v}{c}x\right) \tag{5.1}$$

while $y' = y$ and $z' = z$. Here c is the speed of light which has the value,

$$c = 299792458\ ms^{-1}$$

Meanwhile γ is the ubiquitous factor

$$\gamma = \sqrt{\frac{1}{1 - v^2/c^2}} \tag{5.2}$$

The Lorentz transformation (5.1) encodes within it all of the fun ideas of time dilation and length contraction that we saw in our first course on relativity.

5.1.1 Four-Vectors

It's extremely useful to package these spacetime coordinates in 4-vectors, with indices running from $\mu = 0$ to $\mu = 3$

$$X^\mu = (ct, x, y, z) \qquad \mu = 0, 1, 2, 3$$

Note that the index is a superscript rather than subscript. This will be important shortly. A general Lorentz transformation is a linear map from X to X' of the form

$$(X')^\mu = \Lambda^\mu{}_\nu X^\nu$$

Here Λ is a 4×4 matrix which obeys the matrix equation

$$\Lambda^T \eta \Lambda = \eta \qquad \Leftrightarrow \qquad \Lambda^\rho{}_\mu \eta_{\rho\sigma} \Lambda^\sigma{}_\nu = \eta_{\mu\nu} \tag{5.3}$$

with $\eta_{\mu\nu}$ the *Minkowski metric*

$$\eta_{\mu\nu} = \text{diag}(+1, -1, -1, -1)$$

The solutions to (5.3) fall into two classes. The first class is simply rotations. Given a 3×3 rotation matrix R obeying $R^T R = 1$, we can construct a Lorentz transformation Λ obeying (5.3) by embedding R in the spatial part,

$$\Lambda^\mu{}_\nu = \left(\begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & & & \\ 0 & & \text{R} & \\ 0 & & & \end{array}\right) \tag{5.4}$$

These transformations describe how to relate the coordinates of two observers who are rotated with respect to each other.

The other class of solutions to (5.3) are the Lorentz boosts. These are the transformations appropriate for observers moving relative to each other. The Lorentz transformation (5.1) is equivalent to

$$\Lambda^\mu{}_\nu = \left(\begin{array}{cc|cc} \gamma & -\gamma v/c & 0 & 0 \\ -\gamma v/c & \gamma & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) \tag{5.5}$$

There are similar solutions associated to boosts along the y and z axes.

The beauty of 4-vectors is that it's extremely easy to write down invariant quantities. These are things which all observers, no matter which their reference frame, can agree on. To construct these we take the inner product of two 4-vectors. The trick is that this inner product uses the Minkowski metric and so comes with some minus signs. For example, the square of the distance from the origin to some point in spacetime labelled by X is

$$X \cdot X = X^\mu \eta_{\mu\nu} X^\nu = c^2t^2 - x^2 - y^2 - z^2$$

which is the invariant interval. Similarly, if we're given two four-vectors X and Y then the inner product $X \cdot Y = X^\mu \eta_{\mu\nu} Y^\nu$ is also a Lorentz invariant.

5.1.2 Proper Time

The key to building relativistic theories of Nature is to find the variables that have nice properties under Lorentz transformations. The 4-vectors X, labelling spacetime points, are a good start. But we need more. Here we review how the other kinematical variables of velocity, momentum and acceleration fit into 4-vectors.

Suppose that, in some frame, the particle traces out a worldline. The clever trick is to find a way to parameterise this path in a way that all observers agree upon. The natural choice is the *proper time* τ, the duration of time experienced by the particle itself. If you're sitting in some frame, watching some particle move with an old-fashioned Newtonian 3-velocity $\mathbf{u}(t)$, then it's simple to show that the relationship between your time t and the proper time of the particle τ is given by

$$\frac{dt}{d\tau} = \gamma(\mathbf{u})$$

The proper time allows us to define the 4-velocity and the 4-momentum. Suppose that the particle traces out a path $X(\tau)$ in some frame. Then the 4-velocity is

$$U = \frac{dX}{d\tau} = \gamma \begin{pmatrix} c \\ \mathbf{u} \end{pmatrix}$$

Similarly, the 4-momentum is $P = mU$ where m is the rest mass of the particle. We write

$$P = \begin{pmatrix} E/c \\ \mathbf{p} \end{pmatrix} \tag{5.6}$$

where $E = m\gamma c^2$ is the energy of the particle and $\mathbf{p} = \gamma m\mathbf{u}$ is the 3-momentum in special relativity.

The importance of U and P is that they too are 4-vectors. Because all observers agree on τ, the transformation law of U and P are inherited from X. This means that under a Lorentz transformation, they too change as $U \to \Lambda U$ and $P \to \Lambda P$. And it means that inner products of U and P are guaranteed to be Lorentz invariant.

5.1.3 Indices Up, Indices Down

Before we move on, we do need to introduce one extra notational novelty. The minus signs in the Minkowski metric η means that it's useful to introduce a slight twist to the usual summation convention of repeated indices. For all the 4-vectors that we introduced above, we always place the spacetime index $\mu = 0, 1, 2, 3$ as a superscript (i.e. up) rather than a subscript.

$$X^\mu = \begin{pmatrix} ct \\ \mathbf{x} \end{pmatrix}$$

This is because the same object with an index down, X_μ, will mean something subtly different. We define

$$X_\mu = \begin{pmatrix} ct \\ -\mathbf{x} \end{pmatrix}$$

With this convention, the Minkowski inner product can be written using the usual convention of summing over repeated indices as

$$X^\mu X_\mu = c^2t^2 - \mathbf{x} \cdot \mathbf{x}$$

In contrast, $X^\mu X^\mu = c^2t^2 + \mathbf{x}^2$ is a dumb thing to write in the context of special relativity since it looks very different to observers in different inertial frames. In fact, we will shortly declare it illegal to write things like $X^\mu X^\mu$.

There is a natural way to think of X_μ in terms of X^μ using the Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$. The following equation is trivially true:

$$X_\mu = \eta_{\mu\nu} X^\nu$$

This means that we can think of the Minkowski metric as allowing us to lower indices. To raise indices back up, we need the inverse of $\eta_{\mu\nu}$ which, fortunately, is the same matrix: $\eta^{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$ which means we have $\eta^{\mu\rho}\eta_{\rho\nu} = \delta^\mu{}_\nu$ and we can write

$$X^\nu = \eta^{\nu\mu} X_\mu$$

From now on, we're going to retain this distinction between all upper and lower indices. All the four-vectors that we've met so far have upper indices. But all can be lowered in the same way. For example, we have

$$U_\mu = \gamma \begin{pmatrix} c \\ -\mathbf{u} \end{pmatrix} \tag{5.7}$$

This trick of distinguishing between indices up and indices down provides a simple formalism to ensure that all objects have nice transformation properties under the Lorentz group. We insist that, just as in the usual summation convention, repeated indices only ever appear in pairs. But now we further insist that pairs always appear with one index up and the other down. The result will be an object which is invariant under Lorentz transformations.

5.1.4 Vectors, Covectors and Tensors

In future courses, you will learn that there is somewhat deeper mathematics lying behind distinguishing X^μ and X_μ: formally, these objects live in different spaces (sometimes called dual spaces). We'll continue to refer to X^μ as vectors, but to distinguish them, we'll call X_μ *covectors*. (In slightly fancier language, the components of the vector X^μ are sometimes said to be *contravariant* while the components of the covector X_μ are said to be *covariant*).

For now, the primary difference between a vector and covector is how they transform under rotations and boosts. We know that, under a Lorentz transformation, any 4-vector changes as

$$X^\mu \to X'^\mu = \Lambda^\mu{}_\nu X^\nu$$

From this, we see that a covector should transform as

$$\begin{aligned} X_\mu \to X'_\mu &= \eta_{\mu\rho} X'^{\rho} \\ &= \eta_{\mu\rho} \Lambda^\rho{}_\sigma X^\sigma \\ &= \eta_{\mu\rho} \Lambda^\rho{}_\sigma \eta^{\sigma\nu} X_\nu \end{aligned}$$

Using our rule for raising and lowering indices, now applied to the Lorentz transformation Λ, we can also write this as

$$X_\mu \to \Lambda_\mu{}^\nu X_\nu$$

where our notation is now getting dangerously subtle: you have to stare to see whether the upper or lower index on the Lorentz transformation comes first.

There is a sense in which $\Lambda_\mu{}^\nu$ can be thought of a the components of the inverse matrix Λ^{-1}. To see this, we go back to the definition of the Lorentz transformation (5.3), and start to use our new rules for raising and lowering indices

$$\begin{aligned} \Lambda^\rho{}_\mu \eta_{\rho\sigma} \Lambda^\sigma{}_\nu = \eta_{\mu\nu} \quad &\Rightarrow \quad \Lambda^\rho{}_\mu \Lambda_{\rho\nu} = \eta_{\mu\nu} \\ &\Rightarrow \quad \Lambda^\rho{}_\mu \Lambda_\rho{}^\sigma = \delta^\sigma_\mu \\ &\Rightarrow \quad \Lambda_\rho{}^\sigma \Lambda^\rho{}_\mu = \delta^\sigma_\mu \end{aligned}$$

In the last line above, we've simply reversed the order of the two terms on the left. (When written in index notation, these are just the entries of the matrix so there's no problem with commuting them). Now we compare this to the formula for the inverse of a matrix,

$$(\Lambda^{-1})^\sigma{}_\rho \Lambda^\rho{}_\mu = \delta^\sigma_\mu \quad \Rightarrow \quad (\Lambda^{-1})^\sigma{}_\rho = \Lambda_\rho{}^\sigma \tag{5.8}$$

Note that you need to be careful where you place the indices in equations like this. The result (5.8) is analogous to the statement that the inverse of a rotation matrix is the transpose matrix. For general Lorentz transformations, we learn that the inverse is sort of the transpose where "sort of" means that there are minus signs from raising and lowering. The placement of indices in (5.8) tells us where those minus signs go.

The upshot of (5.8) is that if we want to abandon index notation all together then vectors transform as $X \to \Lambda X$ while covectors – which, for the purpose of this sentence, we'll call $\tilde{X}$ – transform as $\tilde{X} \to \Lambda^{-1}\tilde{X}$. However, in what follows, we have no intention of abandoning index notation. Instead, we will embrace it. It will be our friend and our guide in showing that the Maxwell equations are consistent with special relativity.

A particularly useful example of a covector is the *four-derivative*. This is the relativistic generalisation of ∇, defined by

$$\partial_\mu = \frac{\partial}{\partial X^\mu} = \left(\frac{1}{c}\frac{\partial}{\partial t}\,,\nabla\right)$$

Notice that the superscript on the spacetime 4-vector X^μ has migrated to a subscript on the derivative ∂_μ. For this to make notational sense, we should check that ∂_μ does indeed transform as covector. This is a simple application of the chain rule. Under a Lorentz transformation, $X^\mu \to X'^\mu = \Lambda^\mu{}_\nu X^\nu$, so we have

$$\partial_\mu = \frac{\partial}{\partial X^\mu} \to \frac{\partial}{\partial X'^\mu} = \frac{\partial X^\nu}{\partial X'^\mu}\frac{\partial}{\partial X^\nu} = (\Lambda^{-1})^\nu{}_\mu \partial_\nu = \Lambda_\mu{}^\nu \partial_\nu$$

which is indeed the transformation of a co-vector.

Tensors

Vectors and covectors are the simplest examples of objects which have nice transformation properties under the Lorentz group. But there are many more examples. The most general object can have a bunch of upper indices and a bunch of lower indices, $T^{\mu_1\ldots\mu_n}{}_{\nu_1\ldots\nu_m}$. These objects are also called *tensors* of type (n, m). In order to qualify as a tensor, they must transform under a Lorentz transformation as

$$T'^{\mu_1\ldots\mu_n}{}_{\nu_1\ldots\nu_m} = \Lambda^{\mu_1}{}_{\rho_1}\ldots\Lambda^{\mu_n}{}_{\rho_n}\Lambda_{\nu_1}{}^{\sigma_1}\ldots\Lambda_{\nu_m}{}^{\sigma_m} T^{\rho_1\ldots\rho_n}{}_{\sigma_1\ldots\sigma_m} \tag{5.9}$$

You can always use the Minkowski metric to raise and lower indices on tensors, changing the type of tensor but keeping the total number of indices $n + m$ fixed.

Tensors of this kind are the building blocks of all our theories. This is because if you build equations only out of tensors which transform in this manner then, as long as the $\mu, \nu, \ldots$ indices match up on both sides of the equation, you're guaranteed to have an equation that looks the same in all inertial frames. Such equations are said to be *covariant*. You'll see more of this kind of thing in courses on *General Relativity* and *Differential Geometry*.

In some sense, this index notation is too good. Remember all those wonderful things that you first learned about in special relativity: time dilation and length contraction and twins and spaceships so on. You'll never have to worry about those again. From now on, you can guarantee that you're working with a theory consistent with relativity by ensuring two simple things

- That you only deal with tensors.
- That the indices match up on both sides of the equation.

It's sad, but true. It's all part of growing up and not having fun anymore.

5.2 Conserved Currents

We started these lectures by discussing the charge density $\rho(\mathbf{x}, t)$, the current density $\mathbf{J}(\mathbf{x}, t)$ and their relation through the continuity equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0$$

which tells us that charge is locally conserved.

The continuity equation is already fully consistent with relativity. To see this, we first need to appreciate that the charge and current densities sit nicely together in a 4-vector,

$$J^\mu = \begin{pmatrix} \rho c \\ \mathbf{J} \end{pmatrix}$$

Of course, placing objects in a four-vector has consequence: it tells us how these objects look to different observers. Let's quickly convince ourselves that it makes sense that charge density and current do indeed transform in this way. We can start by considering a situation where there are only static charges with density ρ_0 and no current. So $J^\mu = (\rho_0, 0)$. Now, in a frame that is boosted by velocity $\mathbf{v}$, the current will appear as $J'^\mu = \Lambda^\mu{}_\nu J^\nu$ with the Lorentz transformation given by (5.5). The new charge density and current are then

$$\rho' = \gamma \rho_0 \quad , \quad \mathbf{J}' = -\gamma \rho \mathbf{v}$$

The first of these equations tells us that different observers see different charge densities. This is because of Lorentz contraction: charge *density* means charge per unit volume. And the volume gets squeezed because lengths parallel to the motion undergo Lorentz contraction. That's the reason for the factor of γ in the observed charge density. Meanwhile, the second of these equations is just the relativistic extension of the formula $\mathbf{J} = \rho \mathbf{v}$ that we first saw in the introduction. (The extra minus sign is because $\mathbf{v}$ here denotes the velocity of the boosted observer; the charge is therefore moving with relative velocity $-\mathbf{v}$).

In our new, relativistic, notation, the continuity equation takes the particularly simple form

$$\partial_\mu J^\mu = 0 \tag{5.10}$$

This equation is Lorentz invariant. This follows simply because the indices are contracted in the right way: one up, and one down.

5.2.1 Magnetism and Relativity

We've learned something unsurprising: boosted charge gives rise to a current. But, combined with our previous knowledge, this tells us something new and important: boosted electric fields must give rise to magnetic fields. The rest of this chapter will be devoted to understanding the details of how this happens. But first, we're going to look at a simple example where we can re-derive the magnetic force purely from the Coulomb force and a dose of Lorentz contraction.

To start, consider a bunch of positive charges $+q$ moving along a line with speed $+v$ and a bunch of negative charges $-q$ moving in the opposite direction with speed $-v$ as shown in the figure. If there is equal density, n, of positive and negative charges then the charge density vanishes while the current is

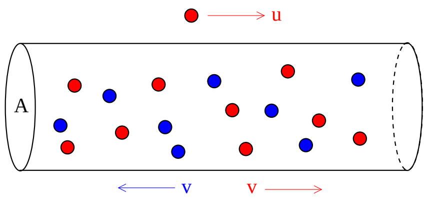


Figure 47:

$$I = 2nAqv$$

where A is the cross-sectional area of the wire. Now consider a test particle, also carrying charge q, which is moving parallel to the wire with some speed u. It doesn't feel any electric force because the wire is neutral, but we know it experiences a magnetic force. Here we will show how to find an expression for this force without ever invoking the phenomenon of magnetism.

The trick is to move to the rest frame of the test particle. This means we have to boost by speed u. The usual addition formula tells us that the velocities of the positive and negative charges now differ, given by

$$v_\pm = \frac{v \mp u}{1 \mp uv/c^2}$$

But with the boost comes a Lorentz contraction which means that the charge density changes. Moreover, because the velocities of positive and negative charges are now different, this will mean that, viewed from the rest frame of our particle, the wire is no longer neutral. Let's see how this works. First, we'll introduce n_0, the density of charges when the particles in the wire are at rest. Then the charge density in the original frame is

$$\rho = qn = \gamma(v)qn_0$$

In this frame the wire is neutral because the positive and negative charges travel at the same speed, albeit in opposite directions. However, in our new frame, the charge densities are

$$\rho_\pm = qn_\pm = q\gamma(v_\pm)n_0 = \left(1 \mp \frac{uv}{c^2}\right)\gamma(u)\gamma(v)\,qn_0$$

where you've got to do a little bit of algebra to get to the last result. Since $v_- > v_+$, we have $n_- > n_+$ and the wire carries negative charge. The overall net charge density in the new frame is

$$\rho' = qn' = q(n_+ - n_-) = -\frac{2uv}{c^2}\,\gamma(u)\,qn$$

But we know that a line of electric charge creates an electric field; we calculated it in (2.6); it is

$$E(r) = -\frac{2uv}{c^2}\,\frac{\gamma(u)\,qnA}{2\pi\epsilon_0 r}\hat{\mathbf{r}}$$

where r is the radial direction away from the wire. This means that, in its rest frame, the particle experiences a force

$$F' = -u\gamma(u)\,\frac{nAq^2v}{\pi\epsilon_0 c^2 r}$$

where the minus sign tells us that the force is towards the wire for $u > 0$. But if there's a force in one frame, there must also be a force in another. Transforming back to where we came from, we conclude that even when the wire is neutral there has to be a force

$$F = \frac{F'}{\gamma(u)} = -u\,\frac{nq^2Av}{\pi\epsilon_0 c^2 r} = -uq\,\frac{\mu_0 I}{2\pi r} \tag{5.11}$$

But this precisely agrees with the Lorentz force law, with the magnetic field given by the expression (3.5) that we computed for a straight wire. Notice that if $u > 0$ then the test particle – which has charge q – is moving in the same direction as the particles in the wire which have charge q and the force is attractive. If $u < 0$ then it moves in the opposite direction and the force is repulsive.

This analysis provides an explicit demonstration of how an electric force in one frame of reference is interpreted as a magnetic force in another. There's also something rather surprising about the result. We're used to thinking of length contraction as an exotic result which is only important when we approach the speed of light. Yet the electrons in a wire crawl along. They take around an hour to travel a meter! Nonetheless, we can easily detect the magnetic force between two wires which, as we've seen above, can be directly attributed to the length contraction in the electron density.

The discussion above needs a minor alteration for actual wires. In the rest frame of the wire the positive charges – which are ions, atoms stripped of some of their electrons – are stationary while the electrons move. Following the explanation above, you might think that there is an imbalance of charge density already in this frame. But that's not correct. The current is due to some battery feeding electrons into the wire and taking them out the other end. And this is done in such a way that the wire is neutral in the rest frame, with the electron density exactly compensating the ion density. In contrast, if we moved to a frame in which the ions and electrons had equal and opposite speeds, the wire would appear charged. Although the starting point is slightly different, the end result remains.

5.3 Gauge Potentials and the Electromagnetic Tensor

Under Lorentz transformations, electric and magnetic fields will transform into each other. In this section, we want to understand more precisely how this happens. At first sight, it looks as if it's going to be tricky. So far the objects which have nice transformation properties under Lorentz transformations are 4-vectors. But here we've got two 3-vectors, $\mathbf{E}$ and $\mathbf{B}$. How do we make those transform into each other?

5.3.1 Gauge Invariance and Relativity

To get an idea for how this happens, we first turn to some objects that we met previously: the scalar and vector potentials ϕ and $\mathbf{A}$. Recall that we introduced these to solve some of the equations of electrostatics and magnetostatics,

$$\begin{aligned} \nabla \times \mathbf{E} = 0 \quad &\Rightarrow \quad \mathbf{E} = -\nabla\phi \\ \nabla \cdot \mathbf{B} = 0 \quad &\Rightarrow \quad \mathbf{B} = \nabla \times \mathbf{A} \end{aligned}$$

However, in general these expressions can't be correct. We know that when $\mathbf{B}$ and $\mathbf{E}$ change with time, the two source-free Maxwell equations are

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \quad \text{and} \quad \nabla \cdot \mathbf{B} = 0$$

Nonetheless, it's still possible to use the scalar and vector potentials to solve both of these equations. The solutions are

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t} \quad \text{and} \quad \mathbf{B} = \nabla \times \mathbf{A}$$

where now $\phi = \phi(\mathbf{x}, t)$ and $\mathbf{A} = \mathbf{A}(\mathbf{x}, t)$.

Just as we saw before, there is no unique choice of ϕ and $\mathbf{A}$. We can always shift $\mathbf{A} \to \mathbf{A} + \nabla\chi$ and $\mathbf{B}$ remains unchanged. However, now this requires a compensating shift of ϕ.

$$\phi \to \phi - \frac{\partial \chi}{\partial t} \quad \text{and} \quad \mathbf{A} \to \mathbf{A} + \nabla\chi \tag{5.12}$$

with $\chi = \chi(\mathbf{x}, t)$. These are *gauge transformations.* They reproduce our earlier gauge transformation for $\mathbf{A}$, while also encompassing constant shifts in ϕ.

How does this help with our attempt to reformulate electromagnetism in a way compatible with special relativity? Well, now we have a scalar, and a 3-vector: these are ripe to place in a 4-vector. We define

$$A^\mu = \begin{pmatrix} \phi/c \\ \mathbf{A} \end{pmatrix}$$

Or, equivalently, $A_\mu = (\phi/c, -\mathbf{A})$. In this language, the gauge transformations (5.12) take a particularly nice form,

$$A_\mu \to A_\mu - \partial_\mu \chi \tag{5.13}$$

where χ is any function of space and time

5.3.2 The Electromagnetic Tensor

We now have all the ingredients necessary to determine how the electric and magnetic fields transform. From the 4-derivative $\partial_\mu = (\partial/\partial(ct), \nabla)$ and the 4-vector $A_\mu = (\phi/c, -\mathbf{A})$, we can form the anti-symmetric tensor

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$$

This is constructed to be invariant under gauge transformations (5.13). We have

$$F_{\mu\nu} \to F_{\mu\nu} + \partial_\mu \partial_\nu \chi - \partial_\nu \partial_\mu \chi = F_{\mu\nu}$$

This already suggests that the components involve the $\mathbf{E}$ and $\mathbf{B}$ fields. To check that this is indeed the case, we can do a few small computations,

$$F_{01} = \frac{1}{c}\frac{\partial(-A_x)}{\partial t} - \frac{\partial(\phi/c)}{\partial x} = \frac{E_x}{c}$$

and

$$F_{12} = \frac{\partial(-A_y)}{\partial x} - \frac{\partial(-A_x)}{\partial y} = -B_z$$

Similar computations for all other entries give us a matrix of electric and magnetic fields,

$$F_{\mu\nu} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix} \tag{5.14}$$

This, then, is the answer to our original question. You can make a Lorentz covariant object consisting of two 3-vectors by arranging them in an anti-symmetric tensor. $F_{\mu\nu}$ is called the *electromagnetic tensor*. Equivalently, we can raise both indices using the Minkowski metric to get

$$F^{\mu\nu} = \eta^{\mu\rho}\eta^{\nu\sigma}F_{\rho\sigma} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix}$$

Both $F_{\mu\nu}$ and $F^{\mu\nu}$ are tensors. They are tensors because they're constructed out of objects, A_μ, ∂_μ and $\eta_{\mu\nu}$, which themselves transform nicely under the Lorentz group. This means that the field strength must transform as

$$F'^{\mu\nu} = \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma F^{\rho\sigma} \tag{5.15}$$

Alternatively, if you want to get rid of the indices, this reads $F' = \Lambda F \Lambda^T$. The observer in a new frame sees electric and magnetic fields $\mathbf{E}'$ and $\mathbf{B}'$ that differ from the original observer. The two are related by (5.15). Let's look at what this means in a couple of illustrative examples.

Rotations

To compute the transformation (5.15), it's probably simplest to just do the sums that are implicit in the repeated ρ and σ labels. Alternatively, if you want to revert to matrix multiplication then this is the same as $F' = \Lambda F \Lambda^T$. Either way, we get the same result. For a rotation, the 3×3 matrix R is embedded in the lower-right hand block of Λ as shown in (5.4). A quick calculation shows that the transformation of the electric and magnetic fields in (5.15) is as expected,

$$\mathbf{E}' = R\,\mathbf{E} \quad \text{and} \quad \mathbf{B}' = R\,\mathbf{B}$$

Boosts

Things are more interesting for boosts. Let's consider a boost v in the x-direction, with Λ given by (5.5). Again, you need to do a few short calculations. For example, we have

$$\begin{aligned}
-\frac{E'_x}{c} = F'^{\,01} &= \Lambda^0{}_\rho \Lambda^1{}_\sigma F^{\rho\sigma} \\
&= \Lambda^0{}_0 \Lambda^1{}_1 F^{01} + \Lambda^0{}_1 \Lambda^1{}_0 F^{10} \\
&= \frac{\gamma^2 v^2}{c^2}\frac{E_x}{c} - \gamma^2 \frac{E_x}{c} = -\frac{E_x}{c}
\end{aligned}$$

and

$$\begin{aligned}
-\frac{E'_y}{c} = F'^{\,02} &= \Lambda^0{}_\rho \Lambda^2{}_\sigma F^{\rho\sigma} \\
&= \Lambda^0{}_0 \Lambda^2{}_2 F^{02} + \Lambda^0{}_1 \Lambda^2{}_2 F^{12} \\
&= -\gamma \frac{E_y}{c} + \frac{\gamma v}{c} B_z = -\frac{\gamma}{c}(E_y - vB_z)
\end{aligned}$$

and

$$\begin{aligned}
-B'_z = F'^{\,12} &= \Lambda^1{}_\rho \Lambda^2{}_\sigma F^{\rho\sigma} \\
&= \Lambda^1{}_0 \Lambda^2{}_2 F^{02} + \Lambda^1{}_1 \Lambda^2{}_2 F^{12} \\
&= \frac{\gamma v}{c^2} E_y - \gamma B_z = -\gamma(B_z - vE_y/c^2)
\end{aligned}$$

The final result for the transformation of the electric field after a boost in the x-direction is

$$\begin{aligned}
E'_x &= E_x \\
E'_y &= \gamma(E_y - vB_z) \\
E'_z &= \gamma(E_z + vB_y)
\end{aligned} \tag{5.16}$$

and, for the magnetic field,

$$\begin{aligned}
B'_x &= B_x \\
B'_y &= \gamma\left(B_y + \frac{v}{c^2}E_z\right) \\
B'_z &= \gamma\left(B_z - \frac{v}{c^2}E_y\right)
\end{aligned} \tag{5.17}$$

As we anticipated above, what appears to be a magnetic field to one observer looks like an electric field to another, and vice versa.

Note that in the limit $v \ll c$, we have $\mathbf{E}' = \mathbf{E} + \mathbf{v} \times \mathbf{B}$ and $\mathbf{B}' = \mathbf{B}$. This can be thought of as the Galilean boost of electric and magnetic fields. We recognise $\mathbf{E}+\mathbf{v}\times\mathbf{B}$ as the combination that appears in the Lorentz force law. We'll return to this force in Section 5.4.1 where we'll see how it's compatible with special relativity.

5.3.3 An Example: A Boosted Line Charge

In Section 2.1.3, we computed the electric field due to a line with uniform charge density η per unit length. If we take the line to lie along the x-axis, we have (2.6)

$$\mathbf{E} = \frac{\eta}{2\pi\epsilon_0(y^2+z^2)}\begin{pmatrix}0\\ y\\ z\end{pmatrix} \tag{5.18}$$

Meanwhile, the magnetic field vanishes for static electric charges: $\mathbf{B} = 0$. Let's see what this looks like from the perspective of an observer moving with speed v in the x-direction, parallel to the wire. In the moving frame the electric and magnetic fields are given by (5.16) and (5.17). These read

$$\begin{aligned}\mathbf{E}' &= \frac{\eta\gamma}{2\pi\epsilon_0(y^2+z^2)}\begin{pmatrix}0\\ y\\ z\end{pmatrix} = \frac{\eta\gamma}{2\pi\epsilon_0(y'^2+z'^2)}\begin{pmatrix}0\\ y'\\ z'\end{pmatrix}\\ \mathbf{B}' &= \frac{\eta\gamma v}{2\pi\epsilon_0 c^2(y^2+z^2)}\begin{pmatrix}0\\ z\\ -y\end{pmatrix} = \frac{\eta\gamma v}{2\pi\epsilon_0 c^2(y'^2+z'^2)}\begin{pmatrix}0\\ z'\\ -y'\end{pmatrix}\end{aligned} \tag{5.19}$$

In the second equality, we've rewritten the expression in terms of the coordinates of $\mathcal{S}'$ which, because the boost is in the x-direction, are trivial: $y = y'$ and $z = z'$.

From the perspective of an observer in frame $\mathcal{S}'$, the charge density in the wire is $\eta' = \gamma\eta$, where the factor of γ comes from Lorentz contraction. This can be seen in the expression above for the electric field. Since the charge density is now moving, the observer in frame $\mathcal{S}'$ sees a current $I' = -\gamma\eta v$. Then we can rewrite (5.19) as

$$\mathbf{B}' = \frac{\mu_0 I'}{2\pi\sqrt{y'^2+z'^2}}\hat{\boldsymbol{\varphi}}' \tag{5.20}$$

But this is something that we've seen before. It's the magnetic field due to a current in a wire (3.5). We computed this in Section 3.1.1 using Ampére's law. But here we've re-derived the same result without ever mentioning Ampére's law! Instead, our starting point (5.18) needed Gauss' law and we then used only the Lorentz transformation of electric and magnetic fields. We can only conclude that, under a Lorentz transformation, Gauss' law must be related to Ampére's law. Indeed, we'll shortly see explicitly that this is the case. For now, it's worth repeating the lesson that we learned in Section 5.2.1: the magnetic field can be viewed as a relativistic effect.

5.3.4 Another Example: A Boosted Point Charge

Consider a point charge Q, stationary in an inertial frame $\mathcal{S}$. We know that it's electric field is given by

$$\mathbf{E} = \frac{Q}{4\pi\epsilon_0 r^2}\,\hat{\mathbf{r}} = \frac{Q}{4\pi\epsilon_0[x^2+y^2+z^2]^{3/2}}\begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

while its magnetic field vanishes. Now let's look at this same particle from the frame $\mathcal{S}'$, moving with velocity $\mathbf{v} = (v,0,0)$ with respect to $\mathcal{S}$. The Lorentz boost which relates the two is given by (5.5) and so the new electric field are given by (5.16),

$$\mathbf{E}' = \frac{Q}{4\pi\epsilon_0[x^2+y^2+z^2]^{3/2}}\begin{pmatrix} x \\ \gamma y \\ \gamma z \end{pmatrix}$$

But this is still expressed in terms of the original coordinates. We should now rewrite this in terms of the coordinates of $\mathcal{S}'$, which are $x' = \gamma(x - vt)$ and $y' = y$ and $z' = z$. Inverting these, we have

$$\mathbf{E}' = \frac{Q\gamma}{4\pi\epsilon_0[\gamma^2(x'+vt')^2 + y'^2 + z'^2]^{3/2}}\begin{pmatrix} x'+vt' \\ y' \\ z' \end{pmatrix} \tag{5.21}$$

In the frame $\mathcal{S}'$, the particle sits at $\mathbf{x}' = (-vt', 0, 0)$, so we see that the electric field emanates from the position of the charge, as it should. For now, let's look at the electric field when $t' = 0$ so that the particle sits at the origin in the new frame. The electric field points outwards radially, along the direction

$$\mathbf{r}' = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}$$

However, the electric field is not isotropic. This arises from the denominator of (5.21) which is not proportional to r'^3 because there's an extra factor of γ^2 in front of the x' component. Instead, at $t' = 0$, the denominator involves the combination

$$\begin{aligned}
\gamma^2 x'^2 + y'^2 + z'^2 &= (\gamma^2 - 1)x'^2 + \mathbf{r}'^2 \\
&= \frac{v^2\gamma^2}{c^2}x'^2 + \mathbf{r}'^2 \\
&= \left(\frac{v^2\gamma^2}{c^2}\cos^2\theta + 1\right)\mathbf{r}'^2 \\
&= \gamma^2\left(1 - \frac{v^2}{c^2}\sin^2\theta\right)\mathbf{r}'^2
\end{aligned}$$

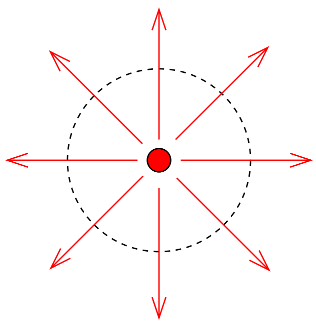


Figure 48: The isotropic field lines of a static charge

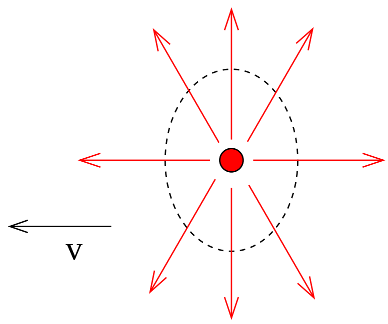


Figure 49: The squeezed field lines of a moving charge

where the θ is the angle between $\mathbf{r}'$ and the x'-axis and, in the last line, we've just used some simple trig and the definition of $\gamma^2 = 1/(1 - v^2/c^2)$. This means that we can write the electric field in frame S' as

$$\mathbf{E}' = \frac{1}{\gamma^2(1 - v^2 \sin^2\theta/c^2)^{3/2}} \frac{Q}{4\pi\epsilon_0 r'^2} \hat{\mathbf{r}}'$$

The pre-factor is responsible for the fact that the electric field is not isotropic. We see that it reduces the electric field along the x'-axis (i.e when $\theta = 0$) and increases the field along the perpendicular y' and z' axes (i.e. when $\theta = \pi/2$). This can be thought of as a consequence of Lorentz contraction, squeezing the electric field lines in the direction of travel.

The moving particle also gives rise to a magnetic field. This is easily computed using the Lorentz transformations (5.17). It is

$$\mathbf{B} = \frac{\mu_0 Q\gamma v}{4\pi[\gamma^2(x' + vt')^2 + y'^2 + z'^2]^{3/2}} \begin{pmatrix} 0 \\ z' \\ -y' \end{pmatrix}$$

5.3.5 Lorentz Scalars

We can now ask a familiar question: is there any combination of the electric and magnetic fields that all observers agree upon? Now we have the power of index notation at our disposal, this is easy to answer. We just need to write down an object that doesn't have any floating μ or ν indices. Unfortunately, we don't get to use the obvious choice of $\eta_{\mu\nu}F^{\mu\nu}$ because this vanishes on account of the anti-symmetry of $F^{\mu\nu}$. The simplest thing we can write down is

$$\frac{1}{2}F_{\mu\nu}F^{\mu\nu} = -\frac{\mathbf{E}^2}{c^2} + \mathbf{B}^2$$

Note the relative minus sign between $\mathbf{E}$ and $\mathbf{B}$, mirroring a similar minus sign in the spacetime interval.

However, this isn't the only Lorentz scalar that we can construct from $\mathbf{E}$ and $\mathbf{B}$. There is another, somewhat more subtle, object. To build this, we need to appreciate that Minkowski spacetime comes equipped with another natural tensor object, beyond the familiar metric $\eta_{\mu\nu}$. This is the fully anti-symmetric object known as the *alternating tensor*,

$$\epsilon^{\mu\nu\rho\sigma} = \begin{cases} +1 & \text{if } \mu\nu\rho\sigma \text{ is an even permutation of } 0123 \\ -1 & \text{if } \mu\nu\rho\sigma \text{ is an odd permutation of } 0123 \end{cases}$$

while $\epsilon^{\mu\nu\rho\sigma} = 0$ if there are any repeated indices.

To see why this is a natural object in Minkowski space, let's look at how it changes under Lorentz transformations. The usual tensor transformation is

$$\epsilon'^{\,\mu\nu\rho\sigma} = \Lambda^\mu{}_\kappa \Lambda^\nu{}_\lambda \Lambda^\rho{}_\alpha \Lambda^\sigma{}_\beta \epsilon^{\kappa\lambda\alpha\beta}$$

It's simple to check that $\epsilon'^{\,\mu\nu\rho\sigma}$ is also full anti-symmetric; it inherits this property from $\epsilon^{\kappa\lambda\alpha\beta}$ on the right-hand side. But this means that $\epsilon'^{\,\mu\nu\rho\sigma}$ must be proportional to $\epsilon^{\mu\nu\rho\sigma}$. We only need to determine the constant of proportionality. To do this, we can look at

$$\epsilon'^{\,0123} = \Lambda^0{}_\kappa \Lambda^1{}_\lambda \Lambda^2{}_\alpha \Lambda^3{}_\beta \epsilon^{\kappa\lambda\alpha\beta} = \det(\Lambda)$$

Now any Lorentz transformations have $\det(\Lambda) = \pm 1$. Those with $\det(\Lambda) = 1$ make up the "proper Lorentz group" $SO(1,3)$. (This was covered in the *Dynamics and Relativity* notes). These proper Lorentz transformations do not include reflections or time reversal. We learn that the alternating tensor $\epsilon^{\mu\nu\rho\sigma}$ is invariant under proper Lorentz transformations. What it's really telling us is that Minkowski space comes with an oriented orthonormal basis. By lowering indices with the Minkowski metric, we can also construct the tensor $\epsilon_{\mu\nu\rho\sigma}$ which has $\epsilon_{0123} = -1$.

The alternating tensor allows us to construct a second tensor field, sometimes called the *dual electromagnetic tensor* (although "dual" is perhaps the most overused word in physics),

$$\tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma} = \begin{pmatrix} 0 & -B_x & -B_y & -B_z \\ B_x & 0 & E_z/c & -E_y/c \\ B_y & -E_z/c & 0 & E_x/c \\ B_z & E_y/c & -E_x/c & 0 \end{pmatrix} \tag{5.22}$$

$\tilde{F}^{\mu\nu}$ is sometimes also written as ${}^\star F^{\mu\nu}$. We see that this is looks just like $F^{\mu\nu}$ but with the electric and magnetic fields swapped around. Actually, looking closely you'll see that there's a minus sign difference as well: $\tilde{F}^{\mu\nu}$ arises from $F^{\mu\nu}$ by the substitution $\mathbf{E} \to c\mathbf{B}$ and $\mathbf{B} \to -\mathbf{E}/c$.

The statement that $\tilde{F}^{\mu\nu}$ is a tensor means that it too has nice properties under Lorentz transformations,

$$\tilde{F}'^{\mu\nu} = \Lambda^\mu{}_\rho \Lambda^\nu{}_\sigma \tilde{F}^{\rho\sigma}$$

and we can use this to build new Lorentz invariant quantities. Taking the obvious square of $\tilde{F}$ doesn't give us anything new, since

$$\tilde{F}^{\mu\nu}\tilde{F}_{\mu\nu} = -F^{\mu\nu}F_{\mu\nu}$$

But by contracting $\tilde{F}$ with the original F we do find a new Lorentz invariant

$$\frac{1}{4}\tilde{F}^{\mu\nu}F_{\mu\nu} = \frac{1}{c}\,\mathbf{E}\cdot\mathbf{B}$$

This tells us that the inner-product of $\mathbf{E}$ and $\mathbf{B}$ is the same viewed in all frames.

5.4 Maxwell Equations

We now have the machinery to write the Maxwell equations in a way which is manifestly compatible with special relativity. They take a particularly simple form:

$$\partial_\mu F^{\mu\nu} = \mu_0 J^\nu \quad \text{and} \quad \partial_\mu \tilde{F}^{\mu\nu} = 0 \tag{5.23}$$

Pretty aren't they!

The Maxwell equations are not *invariant* under Lorentz transformations. This is because there is the dangling ν index on both sides. However, because the equations are built out of objects which transform nicely – $F^{\mu\nu}$, $\tilde{F}^{\mu\nu}$, J^μ and ∂_μ – the equations themselves also transform nicely. For example, we will see shortly that Gauss' law transforms into Ampére's law under a Lorentz boost, something we anticipated in Section 5.3.3. We say that the equations are *covariant* under Lorentz transformations.

This means that an observer in a different frame will mix everything up: space and time, charges and currents, and electric and magnetic fields. Although observers disagree on what these things are, they all agree on how they fit together. This is what it means for an equation to be covariant: the ingredients change, but the relationship between them stays the same. All observers agree that, in their frame, the electric and magnetic fields are governed by the same Maxwell equations.

Given the objects $F^{\mu\nu}$, $\tilde{F}^{\mu\nu}$, J^μ and ∂_μ, the Maxwell equations are not the only thing you could write down compatible with Lorentz invariance. But they are by far the simplest . Any other equation would be non-linear in F or $\tilde{F}$ or contain more derivative terms or some such thing. Of course, simplicity is no guarantee that equations are correct. For this we need experiment. But surprisingly often in physics we find that the simplest equations are also the right ones.

Unpacking the Maxwell Equations

Let's now check that the Maxwell equations (5.23) in relativistic form do indeed coincide with the vector calculus equations that we've been studying in this course. We just need to expand the different parts of the equation. The components of the first Maxwell equation give

$$\begin{aligned} \partial_i F^{i0} = \mu_0 J^0 \quad &\Rightarrow \quad \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \\ \partial_\mu F^{\mu i} = \mu_0 J^i \quad &\Rightarrow \quad -\frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} = \mu_0 \mathbf{J} \end{aligned}$$

In the first equation, which arises from $\nu = 0$, we sum only over spatial indices $i = 1, 2, 3$ because $F^{00} = 0$. Meanwhile the components of the second Maxwell equation give

$$\begin{aligned} \partial_i \tilde{F}^{i0} = 0 \quad &\Rightarrow \quad \nabla \cdot \mathbf{B} = 0 \\ \partial_\mu \tilde{F}^{\mu i} = 0 \quad &\Rightarrow \quad \frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0 \end{aligned}$$

These, of course, are the familiar equations that we've all grown to love over this course.

Here a few further, simple comments about the advantages of writing the Maxwell equations in relativistic form. First, the Maxwell equations imply that current is conserved. This follows because $F^{\mu\nu}$ is anti-symmetric, so $\partial_\mu \partial_\nu F^{\mu\nu} = 0$ automatically, simply because $\partial_\mu \partial_\nu$ is symmetric. The first of the Maxwell equations (5.23) then requires that the continuity equation holds

$$\partial_\mu J^\mu = 0$$

This is the same calculation that we did in vector notation in Section 4.2.1. Note that it's marginally easier in the relativistic framework.

The second Maxwell equation can be written in a number of different ways. It is equivalent to:

$$\partial_\mu \tilde{F}^{\mu\nu} = 0 \quad \Leftrightarrow \quad \epsilon^{\mu\nu\rho\sigma} \partial_\nu F_{\rho\sigma} = 0 \quad \Leftrightarrow \quad \partial_\rho F_{\mu\nu} + \partial_\nu F_{\rho\mu} + \partial_\mu F_{\nu\rho} = 0$$

where the last of these equalities follows because the equation is constructed so that it is fully anti-symmetric with respect to exchanging any of the indices ρ, μ and ν. (Just expand out for a few examples to see this).

The gauge potential A_μ was originally introduced to solve the two Maxwell equations which are contained in $\partial_\mu \tilde{F}^{\mu\nu} = 0$. Again, this is marginally easier to see in relativistic notation. If we write $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ then

$$\partial_\mu \tilde{F}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}\partial_\mu F_{\rho\sigma} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma}\partial_\mu(\partial_\rho A_\sigma - \partial_\sigma A_\rho) = 0$$

where the final equality holds because of the symmetry of the two derivatives, combined with the anti-symmetry of the ϵ-tensor. This means that we could equally well write the Maxwell equations as

$$\partial_\mu F^{\mu\nu} = \mu_0 J^\nu \quad \text{where} \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$$

In more advanced formulations of electromagnetism (for example, in the Lagrangian formulation), this is the form in which the Maxwell equations arise.

5.4.1 The Lorentz Force Law

There's one last aspect of electromagnetism that we need to show is compatible with relativity: the Lorentz force law. In the Newtonian world, the equation of motion for a particle moving with velocity $\mathbf{u}$ and momentum $\mathbf{p} = m\mathbf{u}$ is

$$\frac{d\mathbf{p}}{dt} = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \tag{5.24}$$

We want to write this equation in 4-vector notation in a way that makes it clear how all the objects change under Lorentz transformations.

By now it should be intuitively clear how this is going to work. A moving particle experiences the magnetic force. But if we boost to its rest frame, there is no magnetic force. Instead, the magnetic field transforms into an electric field and we find the same force, now interpreted as an electric force.

The relativistic version of (5.24) involves the 4-momentum P^μ, defined in (5.6), the proper time τ, reviewed in Section 5.1.2, and our new friend the electromagnetic tensor $F^{\mu\nu}$. The electromagnetic force acting on a point particle of charge q can then be written as

$$\frac{dP^\mu}{d\tau} = q\, F^{\mu\nu} U_\nu \tag{5.25}$$

where the 4-velocity is

$$U^\mu = \frac{dX^\mu}{d\tau} = \gamma \begin{pmatrix} c \\ \mathbf{u} \end{pmatrix} \tag{5.26}$$

and the 4-momentum is $P = mU$. Again, we see that the relativistic form of the equation (5.25) is somewhat prettier than the original equation (5.24).

Unpacking the Lorentz Force Law

Let's check to see that the relativistic equation (5.25) is giving us the right physics. It is, of course, four equations: one for each $\mu = 0, 1, 2, 3$. It's simple to multiply out the right-hand side, remembering that U_μ comes with an extra minus sign in the spatial components relative to (5.26). We find that the $\mu = 1, 2, 3$ components of (5.25) arrange themselves into a familiar vector equation,

$$\frac{d\mathbf{p}}{d\tau} = q\gamma(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \quad \Rightarrow \quad \frac{d\mathbf{p}}{dt} = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \tag{5.27}$$

where we've used the relationship $dt/d\tau = \gamma$. We find that we recover the Lorentz force law. Actually, there's a slight difference from the usual Newtonian force law (5.24), although the difference is buried in our notation. In the Newtonian setting, the momentum is $\mathbf{p} = m\mathbf{u}$. However, in the relativistic setting above, the momentum is $\mathbf{p} = m\gamma\mathbf{u}$. Needless to say, the relativistic version is correct, although the difference only shows up at high speeds.

The relativistic formulation of the Lorentz force (5.25) also contains an extra equation coming from $\mu = 0$. This reads

$$\frac{dP^0}{d\tau} = \frac{q}{c}\gamma\,\mathbf{E} \cdot \mathbf{u} \tag{5.28}$$

Recall that the temporal component of the four-momentum is the energy $P^0 = E/c$. Here the energy is $E = m\gamma c^2$ which includes both the rest-mass of the particle and its kinetic energy. The extra equation in (5.25) is simply telling us that the kinetic energy increases when work is done by an electric field

$$\frac{d(\text{Energy})}{dt} = q\mathbf{E} \cdot \mathbf{u}$$

where I've written energy as a word rather than as E to avoid confusing it with the electric field $\mathbf{E}$.

5.4.2 Motion in Constant Fields

We already know how electric and magnetic fields act on particles in a Newtonian world. Electric fields accelerate particles in straight lines; magnetic fields make particles go in circles. Here we're going to redo this analysis in the relativistic framework. The Lorentz force law remains the same. The only difference is that momentum is now $\mathbf{p} = m\gamma\mathbf{u}$. We'll see how this changes things.

Constant Electric Field

Consider a vanishing magnetic field and constant electric field $\mathbf{E} = (E, 0, 0)$. (Note that E here denotes electric field, not energy!). The equation of motion (5.27) for a charged particle with velocity $\mathbf{u} = (u, 0, 0)$ is

$$m\frac{d(\gamma u)}{dt} = qE \quad \Rightarrow \quad m\gamma u = qEt$$

where we've implicitly assumed that the particle starts from rest at $t = 0$. Rearranging, we get

$$u = \frac{dx}{dt} = \frac{qEt}{\sqrt{m^2 + q^2E^2t^2/c^2}}$$

Reassuringly, the speed never exceeds the speed of light. Instead, $u \to c$ as $t \to \infty$ as one would expect. It's simple to integrate this once more. If the particle starts from the origin, we have

$$x = \frac{mc^2}{qE}\left(\sqrt{1 + \frac{q^2E^2t^2}{m^2c^2}} - 1\right)$$

For early times, when the speeds are not too high, this reduces to

$$mx \approx \frac{1}{2}qEt^2 + \ldots$$

which is the usual non-relativistic result for particles undergoing constant acceleration in a straight line.

Constant Magnetic Field

Now let's turn the electric field off and look at the case of constant magnetic field $\mathbf{B} = (0, 0, B)$. In the non-relativistic world, we know that particles turn circles with frequency $\omega = qB/m$. Let's see how relativity changes things.

We start by looking at the zeroth component of the force equation (5.28) which, in the absence of an electric field, reads

$$\frac{dP^0}{d\tau} = 0$$

This tells us that magnetic fields do no work. We knew this from our course on Newtonian physics, but it remains true in the relativistic context. So we know that energy, $E = m\gamma c^2$, is constant. But this tells us that the speed (i.e. the magnitude of the velocity) remains constant. In other words, the velocity, and hence the position, once again turn circles. The equation of motion is now

$$m\frac{d(\gamma\mathbf{u})}{dt} = q\mathbf{u}\times\mathbf{B}$$

Since γ is constant, the equation takes the same form as in the non-relativistic case and the solutions are circles (or helices if the particle also moves in the z-direction). The only difference is that the frequency with which the particle moves in a circle now depends on how fast the particle is moving,

$$\omega = \frac{qB}{m\gamma}$$

If you wanted, you could interpret this as due to the relativistic increase in the mass of a moving particle. Naturally, for small speeds $\gamma \approx 1$ and we reproduce the more familiar cyclotron frequency $\gamma \approx qB/m$.

So far we have looked at situations in which $\mathbf{E} = 0$ and in which $\mathbf{B} = 0$. But we've seen that $\mathbf{E}\cdot\mathbf{B} = 0$ and $\mathbf{E}^2 - \mathbf{B}^2$ are both Lorentz invariant quantities. This means that the solutions we've described above can be boosted to apply to any situation where $\mathbf{E}\cdot\mathbf{B} = 0$ and $\mathbf{E}^2 - \mathbf{B}^2$ is either > 0 or < 0. In the general situation, both electric and magnetic fields are turned on so $\mathbf{E}\cdot\mathbf{B} \neq 0$ and we have three possibilities to consider depending on whether $\mathbf{E}^2 - \mathbf{B}^2$ is > 0 or < 0 or $= 0$.

6. Electromagnetic Radiation

We've seen that Maxwell's equations allow for wave solutions. This is light. Or, more generally, electromagnetic radiation. But how do you generate these waves from a collection of electric charges? In other words, how do you make light?

We know that a stationary electric charge produce a stationary electric field. If we boost this charge so it moves at a constant speed, it produces a stationary magnetic field. In this section, we will see that propagating electromagnetic waves are created by *accelerating* charges.

6.1 Retarded Potentials

We start by simply solving the Maxwell equations for a given current distribution $J^\mu = (\rho c, \mathbf{J})$. We did this in Section 2 and Section 3 for situations where both charges and currents are independent of time. Here we're going to solve the Maxwell equations in full generality where the charges and currents are time dependent.

We know that we can solve half of Maxwell's equations by introducing the gauge potential $A_\mu = (\phi/c, -\mathbf{A})$ and writing $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Then the remaining equations become

$$\partial_\nu F^{\nu\mu} = \mu_0 J^\mu \quad \Rightarrow \quad \Box A^\mu - \partial^\mu(\partial_\nu A^\nu) = \mu_0 J^\mu \tag{6.1}$$

where $\Box$ is the wave operator: $\Box = \partial_\mu \partial^\mu = (1/c^2)\partial^2/\partial t^2 - \nabla^2$.

This equation is invariant under gauge transformations

$$A^\mu \to A^\mu + \partial^\mu \chi \tag{6.2}$$

Any two gauge potentials related by the transformation (6.2) are considered physically equivalent. We will use this symmetry to help us solve (6.1). To do this we make a gauge choice:

Claim: We can use the gauge symmetry (6.2) to choose A^μ to satisfy

$$\partial_\mu A^\mu = 0 \tag{6.3}$$

This is known as *Lorentz Gauge*. It was actually discovered by a guy named Lorenz who had the misfortune to discover a gauge choice that is Lorentz invariant: all observers will agree on the gauge condition (6.3).

Proof: Suppose you are handed a gauge potential A_μ which doesn't obey (6.3) but, instead, satisfies $\partial_\mu A^\mu = f$ for some function f. Then do a gauge transformation of the form (6.2). Your new gauge potential will obey $\partial_\mu A^\mu + \Box\chi = f$. This means that if you can find a gauge transformation χ which satisfies $\Box\chi = f$ then your new gauge potential will be in Lorentz gauge. Such a χ can always be found. This follows from general facts about differential equations. (Note that this proof is essentially the same as we used in Section 3.2.2 when proving that we could always choose Coulomb gauge $\nabla \cdot \mathbf{A} = 0$). $\Box$

If we are in Lorentz gauge then the Maxwell equations (6.1) become particularly simple; they reduce to the sourced wave equation

$$\Box A^\mu = \left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) A^\mu = \mu_0 J^\mu \tag{6.4}$$

Our goal is to solve this equation, subject to the condition (6.3). We'll assume that J has compact spatial support, meaning that the charges and currents are restricted to some finite region of space. As an aside, notice that this is the same kind of equation as $\Box\chi = -f$ which we needed to solve to go Lorentz gauge in the first place. This means that the methods we develop below will allow us to solve figure out both how to go to Lorentz gauge, and also how to solve for A_μ once we're there.

In the following, we'll solve (6.4) in two (marginally) different ways. The first way is quicker; the second way gives us a deeper understanding of what's going on.

6.1.1 Green's Function for the Helmholtz Equation

For our first method, we will Fourier transform A_μ and J_μ in time, but not in space. We write

$$A_\mu(\mathbf{x}, t) = \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \tilde{A}_\mu(\mathbf{x}, \omega)\, e^{-i\omega t} \quad \text{ and } \quad J_\mu(\mathbf{x}, t) = \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \tilde{J}_\mu(\mathbf{x}, \omega)\, e^{-i\omega t}$$

Now the Fourier components $\tilde{A}_\mu(\mathbf{x}, \omega)$ obey the equation

$$\left(\nabla^2 + \frac{\omega^2}{c^2}\right) \tilde{A}_\mu = -\mu_0 \tilde{J}_\mu \tag{6.5}$$

This is the *Helmholtz equation* with source given by the current $\tilde{J}$.

When $\omega = 0$, the Helmholtz equation reduces to the Poisson equation that we needed in our discussion of electrostatics. We solved the Poisson equation using the method of Green's functions when discussing electrostatics in Section 2.2.3. Here we'll do the same for the Helmholtz equation. The Green's function for the Helmholtz equation obeys

$$\left(\nabla^2 + \frac{\omega^2}{c^2}\right) G_\omega(\mathbf{x};\mathbf{x}') = \delta^3(\mathbf{x} - \mathbf{x}')$$

Translational and rotational invariance ensure that the solutions to this equation are of the form $G_\omega(\mathbf{x};\mathbf{x}') = G_\omega(r)$ with $r = |\mathbf{x} - \mathbf{x}'|$. We can then write this as the ordinary differential equation,

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{dG_\omega}{dr}\right) + \frac{\omega^2}{c^2} G_\omega = \delta^3(r) \tag{6.6}$$

We want solutions that vanish as $r \to \infty$. However, even with this restriction, there are still two such solutions. Away from the origin, they take the form

$$G_\omega \sim \frac{e^{\pm i\omega r/c}}{r}$$

We will see shortly that there is a nice physical interpretation of these two Green's functions. First, let's figure out the coefficient that sits in front of the Green's function. This is determined by the delta-function. We integrate both sides of (6.6) over a ball of radius R. We get

$$4\pi \int_0^R dr\, r^2 \left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dG_\omega}{dr}\right) + \frac{\omega^2}{c^2} G_\omega\right] = 1$$

Now, taking the limit $R \to 0$, only the first term on the left-hand side survives. Moreover, only the first term of $dG_\omega/dr \sim (-1/r^2 \pm i\omega/cr)e^{\pm i\omega r/c}$ survives. We find that the two Green's functions for the Helmholtz equation are

$$G_\omega(r) = -\frac{1}{4\pi}\frac{e^{\pm i\omega r/c}}{r}$$

Note that this agrees with the Green's function for the Poisson equation when $\omega = 0$.

Retarded Potentials

So which $\pm$ sign should we take? The answer depends on what we want to do with the Green's function. For our purposes, we'll nearly always need $G_\omega \sim e^{+i\omega r/c}/r$. Let's see why. The Green's function G_ω allows us to write the Fourier components $\tilde{A}_\mu$ in (6.5) as

$$\tilde{A}_\mu(\mathbf{x},\omega) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{e^{+i\omega|\mathbf{x}-\mathbf{x}'|/c}}{|\mathbf{x}-\mathbf{x}'|}\,\tilde{J}_\mu(\mathbf{x}',\omega)$$

which, in turn, means that the time-dependent gauge potential becomes

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int \frac{d\omega}{2\pi}\int d^3x'\ \frac{e^{-i\omega(t-|\mathbf{x}-\mathbf{x}'|/c)}}{|\mathbf{x}-\mathbf{x}'|}\,\tilde{J}_\mu(\mathbf{x}')$$

But now the integral over ω is just the inverse Fourier transform. With one difference: what was the time variable t has become the *retarted time*, $t_{\rm ret}$, with

$$ct_{\rm ret} = ct - |\mathbf{x}-\mathbf{x}'|$$

We have our final result,

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} \tag{6.7}$$

This is called the *retarded potential.* To determine the contribution at point $\mathbf{x}$ and time t, we integrate the current over all of space, weighted with the Green's function factor $1/|\mathbf{x}-\mathbf{x}'|$ which captures the fact that points further away contribute more weakly.

After all this work, we've arrived at something rather nice. The general form of the answer is very similar to the result for electrostatic potential and magnetostatic vector potential that we derived in Sections 2 and 3. Recall that when the charge density and current were independent of time, we found

$$\phi(\mathbf{x}) = \frac{1}{4\pi\epsilon_0}\int d^3x'\ \frac{\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} \quad \text{and} \quad \mathbf{A}(\mathbf{x}) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{\mathbf{J}(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}$$

But when the charge density and current do depend on time, we see from (6.7) that something new happens: the gauge field at point $\mathbf{x}$ and time t depends on the current configuration at point $\mathbf{x}'$ and the *earlier* time $t_{\rm ret} = t - |\mathbf{x}-\mathbf{x}'|/c$. This, of course, is due to causality. The difference $t - t_{\rm ret}$ is just the time it took the signal to propagate from $\mathbf{x}'$ to $\mathbf{x}$, travelling at the speed of light. Of course, we know that Maxwell's equations are consistent with relativity so something like this had to happen; we couldn't have signals travelling instantaneously. Nonetheless, it's pleasing to see how this drops out of our Green's functionology.

Finally, we can see what would happen were we to choose the other Green's function, $G_\omega \sim e^{-i\omega r/c}/r$. Following through the steps above, we see that the retarded time $t_{\rm ret}$ is replaced by the advanced time $t_{\rm adv} = t + |\mathbf{x} - \mathbf{x}'|/c$. Such a solution would mean that the gauge field depends on what the current is doing in the future, rather than in the past. These solutions are typically thrown out as being unphysical. We'll have (a little) more to say about them at the end of the next section.

6.1.2 Green's Function for the Wave Equation

The expression for the retarded potential (6.7) is important. In this section, we provide a slightly different derivation. This will give us more insight into the origin of the retarded and advanced solutions. Moreover, the techniques below will also be useful in later courses[4].

We started our previous derivation by Fourier transforming only the time coordinate, to change the wave equation into the Helmholtz equation. Here we'll treat time and space on more equal footing and solve the wave equation directly. We again make use of Green's functions. The Green's function for the wave equation obeys

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G(\mathbf{x},t;\mathbf{x}',t') = \delta^3(\mathbf{x}-\mathbf{x}')\delta(t-t') \tag{6.8}$$

Translational invariance in space and time means that the Green's function takes the form $G(\mathbf{x},t;\mathbf{x}',t) = G(\mathbf{x}-\mathbf{x}',t-t')$. To determine this function $G(\mathbf{r},t)$, with $\mathbf{r} = \mathbf{x}-\mathbf{x}'$, we Fourier transform both space and time coordinates,

$$G(\mathbf{x},t) = \int \frac{dw\, d^3k}{(2\pi)^4}\, \tilde{G}(\mathbf{k},\omega)\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{6.9}$$

Choosing $\mathbf{x}' = 0$ and $t' = 0$, the wave equation (6.8) then becomes

$$\begin{aligned}
\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G(\mathbf{r},t) &= \int \frac{dw\, d^3k}{(2\pi)^4}\, \tilde{G}(\omega,\mathbf{k}) \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\
&= \int \frac{dw\, d^3k}{(2\pi)^4}\, \tilde{G}(\mathbf{k},\omega) \left(-k^2 + \frac{\omega^2}{c^2}\right) e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\
&= \delta^3(\mathbf{r})\delta(t) = \int \frac{dw\, d^3k}{(2\pi)^4}\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}
\end{aligned}$$

Equating the terms inside the integral, we see that the Fourier transform of the Green's function takes the simple form

$$\tilde{G}(\mathbf{k},\omega) = -\frac{1}{k^2 - \omega^2/c^2}$$

[4] A very similar discussion can be found in the lecture notes on *Quantum Field Theory*.

But notice that this diverges when $\omega^2 = c^2k^2$. This pole results in an ambiguity in the Green's function in real space which, from (6.9), is given by

$$G(\mathbf{r},t) = -\int \frac{dw\, d^3k}{(2\pi)^4}\frac{1}{k^2-\omega^2/c^2}\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$$

We need some way of dealing with that pole in the integral. To see what's going on, it's useful to change to polar coordinates for the momentum integrals over $\mathbf{k}$. This will allow us to deal with that $e^{i\mathbf{k}\cdot\mathbf{r}}$ factor. The best way to do this is to think of fixing $\mathbf{r}$ and then choosing the k_z axis to point along x. We then write $\mathbf{k}\cdot\mathbf{r} = kr\cos\theta$, and the Green's function becomes

$$G(\mathbf{r},t) = -\frac{1}{(2\pi)^4}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta\, \sin\theta \int_0^{\infty} dk\, k^2 \int_{-\infty}^{+\infty} d\omega\, \frac{1}{k^2-\omega^2/c^2}\, e^{i(kr\cos\theta-\omega t)}$$

Now the $d\phi$ integral is trivial, while the $d\theta$ integral is

$$\int_0^{\pi} d\theta\, \sin\theta\, e^{ikr\cos\theta} = -\frac{1}{ikr}\int_0^{\pi} d\theta\left[\frac{d}{d\theta}e^{ikr\cos\theta}\right] = -\frac{1}{ikr}\left[e^{-ikr}-e^{+ikr}\right] = 2\frac{\sin kr}{kr}$$

After performing these angular integrals, the real space Green's function becomes

$$G(\mathbf{r},t) = \frac{1}{4\pi^3}\int_0^{\infty} dk\ c^2k^2\, \frac{\sin kr}{kr}\int_{-\infty}^{+\infty} d\omega \frac{e^{-i\omega t}}{(\omega-ck)(\omega+ck)}$$

Now we have to face up to those poles. We'll work by fixing k and doing the ω integral first. (Afterwards, we'll then have to do the k integral). It's clear that we run into two poles at $\omega = \pm ck$ when we do the ω integral and we need a prescription for dealing with these. To do this, we need to pick a contour C in the complex ω plane which runs along the real axis but skips around the poles. There are different choices for C. Each of them provides a Green's function which obeys (6.8) but, as we will now see, these Green's functions are different. What's more, this difference has a nice physical interpretation.

Retarded Green's Function

To proceed, let's just pick a particular C and see what happens. We choose a contour which skips above the poles at $\omega = \pm ck$ as shown in the diagram. This results in what's called the *retarded Greens function*; we denote it as $G_{\rm ret}(\mathbf{r},t)$. As we now show, it depends crucially on whether $t<0$ or $t>0$.

Let's first look at the case with $t < 0$. Here, $e^{-i\omega t} \to 0$ when $\omega \to +i\infty$. This means that, for $t < 0$, we can close the contour C in the upper-half plane as shown in the figure and the extra semi-circle doesn't give rise to any further contribution. But there are no poles in the upper-half plane. This means that, by the Cauchy residue theorem, $G_{\rm ret}(\mathbf{r}, t) = 0$ when $t < 0$.

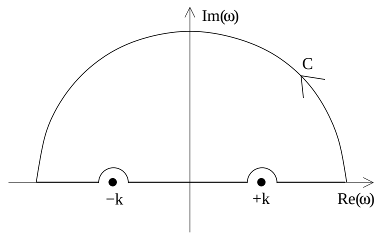


Figure 50:

In contrast, when $t > 0$ we have $e^{-i\omega t} \to 0$ when $\omega \to -i\infty$, which means that we get to close the contour in the lower-half plane. Now we do pick up contributions to the integral from the two poles at $\omega = \pm ck$. This time the Cauchy residue theorem gives

$$\begin{aligned}\int_C d\omega\, \frac{e^{-i\omega t}}{(\omega - ck)(\omega + ck)} &= -2\pi i \left[\frac{e^{-ickt}}{2ck} - \frac{e^{+ickt}}{2ck}\right] \\ &= -\frac{2\pi}{ck} \sin\, ckt \qquad\qquad (t > 0)\end{aligned}$$

So, for $t > 0$, the Green's function becomes

$$\begin{aligned}G_{\rm ret}(\mathbf{r}, t) &= -\frac{1}{2\pi^2}\frac{1}{r}\int_0^\infty dk\ c\, \sin kr\, \sin\, ckt \\ &= \frac{1}{4\pi^2}\frac{1}{r}\int_{-\infty}^\infty dk\ \frac{c}{4}\,(e^{ikr} - e^{-ikr})(e^{ickt} - e^{-ickt}) \\ &= \frac{1}{4\pi^2}\frac{1}{r}\int_{-\infty}^\infty dk\ \frac{c}{4}\,(e^{ik(r+ct)} + e^{-ik(r+ct)} - e^{ik(r-ct)} - e^{-ik(r-ct)})\end{aligned}$$

Each of these final integrals is a delta-function of the form $\delta(r \pm ct)$. But, obviously, $r > 0$ while this form of the Green's function is only valid for $t > 0$. So the $\delta(r + ct)$ terms don't contribute and we're left with

$$G_{\rm ret}(\mathbf{x}, t) = -\frac{1}{4\pi}\frac{c}{r}\,\delta(r - ct) \qquad\quad t > 0$$

We can absorb the factor of c into the delta-function. (Recall that $\delta(x/a) = |a|\delta(x)$ for any constant a). So we finally get the answer for the *retarded Green's function*

$$G_{\rm ret}(\mathbf{r}, t) = \begin{cases} 0 & t < 0 \\ -\dfrac{1}{4\pi r}\,\delta(t_{\rm ret}) & t > 0\end{cases}$$

where $t_{\rm ret}$ is the retarded time that we met earlier,

$$t_{\rm ret} = t - \frac{r}{c}$$

The delta-function ensures that the Green's function is only non-vanishing on the light-cone emanating from the origin.

Finally, with the retarded Green's function in hand, we can construct what we really want: solutions to the wave equation (6.4). These solutions are given by

$$\begin{aligned} A_\mu(\mathbf{x},t) &= -\mu_0 \int d^3x'dt'\ G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')\, J_\mu(\mathbf{x}',t') \\ &= \frac{\mu_0}{4\pi}\int d^3x'dt'\ \frac{\delta(t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} J_\mu(\mathbf{x}',t') \\ &= \frac{\mu_0}{4\pi}\int d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} \end{aligned} \tag{6.10}$$

Happily, we find the same expression for the retarded potential that we derived previously in (6.7).

Advanced Green's Function

Let us briefly look at other Green's functions. We can pick the contour C in the complex ω-plane to skip below the two poles on the real axis. This results in what's called the *advanced Green's function*. Now, when $t > 0$, we complete the contour in the lower-half plane, as shown in the figure, where the lack of poles means that the advanced Green's function vanishes. Meanwhile, for $t < 0$, we complete the contour in the upper-half plane and get

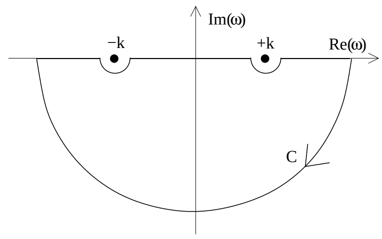


Figure 51:

$$G_{\rm adv}(\mathbf{r},t) = \begin{cases} -\dfrac{1}{4\pi r}\,\delta(t_{\rm adv}) & t<0 \\ 0 & t>0 \end{cases}$$

where

$$t_{\rm adv} = t + \frac{r}{c}$$

The resulting solution gives a solution known as the advanced potential,

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm adv})}{|\mathbf{x}-\mathbf{x}'|}$$

It's hard to think of this solution as anything other than unphysical. Taken at face value, the effect of the current and charges now propagates backwards in time to determine the gauge potential A_μ. The sensible thing is clearly to throw these solutions away.

However, it's worth pointing out that the choice of the retarded propagator $G_{\rm ret}$ rather than the advanced propagator $G_{\rm adv}$ is an extra ingredient that we should add to the theory of electromagnetism. The Maxwell equations themselves are time symmetric; the choice of which solutions are physical is not.

There is some interesting history attached to this. A number of physicists have felt uncomfortable at imposing this time asymmetry only at the level of solutions, and attempted to rescue the advanced propagator in some way. The most well-known of these is the Feynman-Wheeler absorber theory, which uses a time symmetric propagator, with the time asymmetry arising from boundary conditions. However, I think it's fair to say that these ideas have not resulted in any deeper understanding of how time emerges in physics.

Finally, there is yet another propagator that we can use. This comes from picking a contour C that skips under the first pole and over the second. It is known as the *Feynman propagator* and plays an important role in quantum field theory.

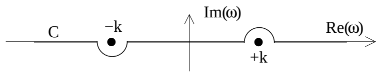


Figure 52:

6.1.3 Checking Lorentz Gauge

There is a loose end hanging over from our previous discussion. We have derived the general solution to the wave equation (6.4) for A_μ.This is given by the retarded potential

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} \tag{6.11}$$

But the wave equation is only equivalent to the Maxwell equations if it obeys the Lorentz gauge fixing condition, $\partial_\mu A^\mu = 0$. We still need to check that this holds. In fact, this follows from the conservation of the current: $\partial_\mu J^\mu = 0$. To show this, it's actually simplest to return to a slightly earlier form of this expression (6.10)

$$A_\mu(\mathbf{x},t) = -\mu_0\int d^3x'dt'\ G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')\,J_\mu(\mathbf{x}',t')$$

The advantage of this is that both time and space remain on an equal footing. We have

$$\partial_\mu A^\mu(\mathbf{x},t) = -\mu_0\int d^3x'dt'\ \partial_\mu G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')\,J^\mu(\mathbf{x}',t')$$

But now we use the fact that $G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')$ depends on $\mathbf{x}-\mathbf{x}'$ and $t-t'$ to change the derivative ∂_μ acting on x into a derivative ∂'_μ acting on x'. We pick up a minus sign for our troubles. We then integrate by parts to find,

$$\begin{aligned}\partial_\mu A^\mu(\mathbf{x},t) &= +\mu_0 \int d^3x'dt'\ \partial'_\mu G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')\,J^\mu(\mathbf{x}',t') \\ &= -\mu_0 \int d^3x'dt'\ G_{\rm ret}(\mathbf{x},t;\mathbf{x}',t')\,\partial'_\mu J^\mu(\mathbf{x}',t') \\ &= 0\end{aligned}$$

as required. If you prefer, you can also run through the same basic steps with the form of the solution (6.11). You have to be a little careful because $t_{\rm ret}$ now also depends on $\mathbf{x}$ and $\mathbf{x}'$ so you get extra terms at various stages when you differentiate. But it all drops out in the wash and you again find that Lorentz gauge is satisfied courtesy of current conservation.

6.2 Dipole Radiation

Let's now use our retarded potential do understand something new. This is the set-up: there's some localised region V in which there is a time-dependent distribution of charges and currents. But we're a long way from this region. We want to know what the resulting electromagnetic field looks like.

Our basic formula is the retarded potential,

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int_V d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} \tag{6.12}$$

The current $J^\mu(\mathbf{x}',t)$ is non-zero only for $\mathbf{x}' \in V$. We denote the size of the region V as d and we're interested in what's happening at a point $\mathbf{x}$ which is a distance $r = |\mathbf{x}|$ away. (A word of warning: in this section we're using $r = |\mathbf{x}|$ which differs from our notation in Section 6.1 where we used $r = |\mathbf{x}-\mathbf{x}'|$). If $|\mathbf{x}-\mathbf{x}'| \gg d$ for all $\mathbf{x}' \in V$ then we can approximate $|\mathbf{x}-\mathbf{x}'| \approx |\mathbf{x}| = r$. In fact, we will keep the leading order correction to this which we get by Taylor expansion. (This is the same Taylor expansion that we needed when deriving the multipole expansion for electrostatics in Section 2.2.3). We have

$$\begin{aligned} |\mathbf{x}-\mathbf{x}'| &= r - \frac{\mathbf{x}\cdot\mathbf{x}'}{r} + \ldots \\ \Rightarrow \qquad \frac{1}{|\mathbf{x}-\mathbf{x}'|} &= \frac{1}{r} + \frac{\mathbf{x}\cdot\mathbf{x}'}{r^3} + \ldots \end{aligned} \tag{6.13}$$

There is a new ingredient compared to the electrostatic case: we have a factor of $|\mathbf{x}-\mathbf{x}'|$ that sits inside $t_{\rm ret} = t - |\mathbf{x}-\mathbf{x}'|/c$ as well, so that

$$J_\mu(\mathbf{x}',t_{\rm ret}) = J_\mu(\mathbf{x}',t - r/c + \mathbf{x}\cdot\mathbf{x}'/rc + \ldots)$$

Now we'd like to further expand out this argument. But, to do that, we need to know something about what the current is doing. We will assume that the motion of the charges and current are *non-relativistic* so that the current doesn't change very much over the time $\tau \sim d/c$ that it takes light to cross the region V. For example, if the current varies with characteristic frequency ω (so that $J \sim e^{-i\omega t}$), then this requirement becomes $d/c \ll 1/\omega$. Then we can further Taylor expand the current to write

$$J_\mu(\mathbf{x}', t_{\rm ret}) = J_\mu(\mathbf{x}', t - r/c) + \dot{J}_\mu(\mathbf{x}', t - r/c)\,\frac{\mathbf{x}\cdot\mathbf{x}'}{rc} + \ldots \tag{6.14}$$

We start by looking at the leading order terms in both these Taylor expansions.

6.2.1 Electric Dipole Radiation

At leading order in d/r, the retarded potential becomes simply

$$A_\mu(\mathbf{x}, t) \approx \frac{\mu_0}{4\pi r}\int_V d^3x'\; J_\mu(\mathbf{x}', t - r/c)$$

This is known as the *electric dipole approximation*. (We'll see why very shortly). We want to use this to compute the electric and magnetic fields far from the localised source. It turns out to be simplest to first compute the magnetic field using the 3-vector form of the above equation,

$$\mathbf{A}(\mathbf{x}, t) \approx \frac{\mu_0}{4\pi r}\int_V d^3x'\; \mathbf{J}(\mathbf{x}', t - r/c)$$

We can manipulate the integral of the current using the conservation formula $\dot{\rho} + \nabla\cdot\mathbf{J} = 0$. (The argument is basically a repeat of the kind of arguments we used in the magnetostatics section 3.3.2). We do this by first noting the identity

$$\partial_j(J_j x_i) = (\partial_j J_j)\, x_i + J_i = -\dot{\rho}\, x_i + J_i$$

We integrate this over all of space and discard the total derivative to find

$$\int d^3x'\; \mathbf{J}(\mathbf{x}') = \frac{d}{dt}\int d^3x'\; \rho(\mathbf{x}')\,\mathbf{x}' = \dot{\mathbf{p}}$$

where we recognise $\mathbf{p}$ as the electric dipole moment of the configuration. We learn that the vector potential is determined by the change of the electric dipole moment,

$$\mathbf{A}(\mathbf{x}, t) \approx \frac{\mu_0}{4\pi r}\,\dot{\mathbf{p}}(t - r/c)$$

This, of course, is where the *electric dipole* approximation gets its name.

We now use this to compute the magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$. There are two contributions: one when ∇ acts on the $1/r$ term, and another when ∇ acts on the r in the argument of $\dot{\mathbf{p}}$. These give, respectively,

$$\mathbf{B} \approx -\frac{\mu_0}{4\pi r^2}\,\hat{\mathbf{x}} \times \dot{\mathbf{p}}(t - r/c) - \frac{\mu_0}{4\pi rc}\,\hat{\mathbf{x}} \times \ddot{\mathbf{p}}(t - r/c)$$

where we've used the fact that $\nabla r = \hat{\mathbf{x}}$. Which of these two terms is bigger? As we get further from the source, we would expect that the second, $1/r$, term dominates over the first, $1/r^2$ term. We can make this more precise. Suppose that the source is oscillating at some frequency ω, so that $\ddot{\mathbf{p}} \sim \omega\dot{\mathbf{p}}$. We expect that it will make waves at the characteristic wavelength $\lambda = c/\omega$. Then, as long we're at distances $r \gg \lambda$, the second term dominates and we have

$$\mathbf{B}(t, \mathbf{x}) \approx -\frac{\mu_0}{4\pi rc}\,\hat{\mathbf{x}} \times \ddot{\mathbf{p}}(t - r/c) \tag{6.15}$$

The region $r \gg \lambda$ is called the *far-field zone* or, sometimes, the *radiation zone*. We've now made two successive approximations, valid if we have a hierarchy of scales in our problem: $r \gg \lambda \gg d$.

To get the corresponding electric field, it's actually simpler to use the Maxwell equation $\dot{\mathbf{E}} = c^2\nabla \times \mathbf{B}$. Again, if we care only about large distances, $r \gg \lambda$, the curl of $\mathbf{B}$ is dominated by ∇ acting on the argument of $\ddot{\mathbf{p}}$. We get

$$\begin{aligned} \nabla \times \mathbf{B} &\approx \frac{\mu_0}{4\pi rc^2}\,\hat{\mathbf{x}} \times (\hat{\mathbf{x}} \times \dddot{\mathbf{p}}(t - r/c)) \\ \Rightarrow \qquad \mathbf{E} &\approx \frac{\mu_0}{4\pi r}\,\hat{\mathbf{x}} \times (\hat{\mathbf{x}} \times \ddot{\mathbf{p}}(t - r/c)) \end{aligned} \tag{6.16}$$

Notice that the electric and magnetic field are related in the same way that we saw for plane waves, namely

$$\mathbf{E} = -c\,\hat{\mathbf{x}} \times \mathbf{B}$$

although, now, this only holds when we're suitably far from the source, $r \gg \lambda$. What's happening here is that oscillating dipole is emitting spherical waves. At radius $r \gg \lambda$ these can be thought of as essentially planar.

Notice, also, that the electric field is dropping off slowly as $1/r$. This, of course, is even slower than the usual Coulomb force fall-off.

6.2.2 Power Radiated: Larmor Formula

We can look at the power radiated away by the source. This is computed by the Poynting vector which we first met in Section 4.4. It is given by

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{c}{\mu_0}|\mathbf{B}|^2\hat{\mathbf{x}} = \frac{\mu_0}{16\pi^2 r^2 c}|\hat{\mathbf{x}}\times\ddot{\mathbf{p}}|^2\,\hat{\mathbf{x}}$$

The fact that $\mathbf{S}$ lies in the direction $\hat{\mathbf{x}}$ means that the power is emitted radially. The fact that it drops off as $1/r^2$ follows from the conservation of energy. It means that the total energy flux, computed by integrating $\mathbf{S}$ over a large surface, is constant, independent of r.

Although the radiation is radial, it is not uniform. Suppose that the dipole oscillates in the $\hat{\mathbf{z}}$ direction. Then we have

$$\mathbf{S} = \frac{\mu_0}{16\pi^2 r^2 c}|\ddot{\mathbf{p}}|^2\sin^2\theta\,\hat{\mathbf{z}} \tag{6.17}$$

where θ is the angle between $\hat{\mathbf{x}}$ and the z-axis. The emitted power is largest in the plane perpendicular to the dipole. A sketch of this is shown in the figure.

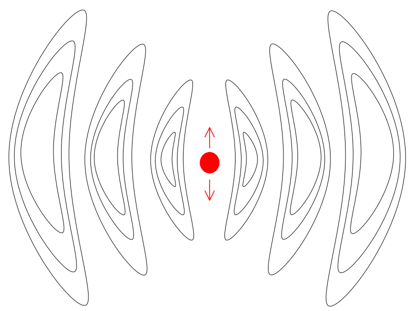


Figure 53:

A device which converts currents into electromagnetic waves (typically in the radio spectrum) is called an *antenna*. We see that it's not possible to create a dipole antenna which emits radiation uniformly. There's actually some nice topology underlying this observation. Look at a sphere which surrounds the antenna at large distance. The radiation is emitted radially, which means that the magnetic field $\mathbf{B}$ lies tangent to the sphere. But there's an intuitive result in topology called the *hairy ball theorem* which says that you can't smoothly comb the hair on a sphere. Or, more precisely, there does not exist a nowhere vanishing vector field on a sphere. Instead, any vector field like $\mathbf{B}$ must vanish at two or more points. In this present context, that ensures that $\mathbf{S}$ too vanishes at two points.

The total radiated power, $\mathcal{P}$, is computed by integrating over a sphere,

$$\mathcal{P} = \int_{S^2} d^2\mathbf{r}\cdot\mathbf{S} = \frac{\mu_0}{16\pi^2 c}|\ddot{\mathbf{p}}|^2\int_0^{2\pi} d\phi\int_0^{\pi} d\theta\,\sin^3\theta$$

where one of the factors of $\sin\theta$ comes from the Jacobian. The integral is easily performed, to get

$$\mathcal{P} = \frac{\mu_0}{6\pi c}|\ddot{\mathbf{p}}|^2 \tag{6.18}$$

Finally, the dipole term $\ddot{\mathbf{p}}$ is still time dependent. It's common practice to compute the time averaged power. The most common example is when the dipole oscillates with frequency ω, so that $|\ddot{\mathbf{p}}|^2 \sim \cos^2(\omega t)$. (Recall that we're only allowed to work with complex expressions when we have linear equations). Then, integrating over a period, $T = 2\pi/\omega$, just gives an extra factor of 1/2.

Let's look at a simple example. Take a particle of charge Q, oscillating in the $\hat{\mathbf{z}}$ direction with frequency ω and amplitude d. Then we have $\mathbf{p} = p\hat{\mathbf{z}}e^{i\omega t}$ with the dipole moment $p = Qd$. Similarly, $\ddot{\mathbf{p}} = -\omega^2 p\hat{\mathbf{z}}e^{i\omega t}$. The end result for the time averaged power $\bar{\mathcal{P}}$ is

$$\bar{\mathcal{P}} = \frac{\mu_0 p^2 \omega^4}{12\pi c} \tag{6.19}$$

This is the *Larmor formula* for the time-averaged power radiated by an oscillating charge. The formula is often described in terms of the acceleration, $a = d\omega^2$. Then it reads

$$\bar{\mathcal{P}} = \frac{Q^2 a^2}{12\pi\epsilon_0 c^3} \tag{6.20}$$

where we've also swapped the μ_0 in the numerator for $\epsilon_0 c^2$ in the denominator.

6.2.3 An Application: Instability of Classical Matter

The popular picture of an atom consists of a bunch of electrons orbiting a nucleus, like planets around a star. But this isn't what an atom looks like. Let's see why.

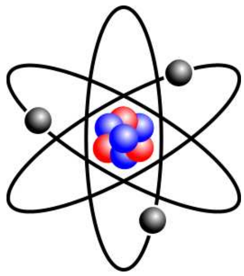


Figure 54: This is not what an atom looks like.

We'll consider a Hydrogen atom, with an electron orbiting around a proton, fixed at the origin. (The two really orbit each other around their common centre of mass, but the mass of the electron is $m_e \approx 9 \times 10^{-31}\ Kg$, while the mass of the proton is about 1800 bigger, so this is a good approximation). The equation of motion for the electron is

$$m_e \ddot{\mathbf{r}} = -\frac{e^2}{4\pi\epsilon_0}\frac{\hat{\mathbf{r}}}{r^2}$$

The dipole moment of the atom is $\mathbf{p} = e\mathbf{r}$ so the equation of motion tells us $\ddot{\mathbf{p}}$. Plugging this into (6.18), we can get an expression for the amount of energy emitted by the electron,

$$\mathcal{P} = \frac{\mu_0}{6\pi c}\left(\frac{e^3}{4\pi\epsilon_0 m_e r^2}\right)^2$$

As the electron emits radiation, it loses energy and must, therefore, spiral towards the nucleus. We know from classical mechanics that the energy of the orbit depends on its eccentricity. For simplicity, let's assume that the orbit is circular with energy

$$E = -\frac{e^2}{4\pi\epsilon_0}\frac{1}{2r}$$

Then we can equate the change in energy with the emitted power to get

$$\dot{E} = \frac{e^2}{8\pi\epsilon_0 r^2}\,\dot{r} = -\mathcal{P} = -\frac{\mu_0}{6\pi c}\left(\frac{e^3}{4\pi\epsilon_0 m_e r^2}\right)^2$$

which gives us an equation that tells us how the radius of the orbit changes,

$$\dot{r} = -\frac{\mu_0 e^4}{12\pi^2 c\epsilon_0 m_e^2 r^2}$$

Suppose that we start at some time, $t = 0$, with a classical orbit with radius r_0. Then we can calculate how long it takes for the electron to spiral down to the origin at $r = 0$. It is

$$T = \int_0^T dt = \int_{r_0}^0 \frac{1}{\dot{r}}\,dr = \frac{4\pi^2 c\epsilon_0 m_e^2 r_0^3}{\mu_0 e^4}$$

Now let's plug in some small numbers. We can take the size of the atom to be $r_0 \approx 5 \times 10^{-11} m$. (This is roughly the Bohr radius that can be derived theoretically using quantum mechanics). Then we find that the lifetime of the hydrogen atom is

$$T \approx 10^{-11}\ s$$

That's a little on the small size. The Universe is 14 billion years old and hydrogen atoms seem in no danger of decaying.

Of course, what we're learning here is something dramatic: the whole framework of classical physics breaks down when we look at the atomic scale and has to be replaced with quantum mechanics. And, although we talk about electron orbits in quantum mechanics, they are very different objects than the classical orbits drawn in the picture. In particular, an electron in the ground state of the hydrogen atom emits no radiation. (Electrons in higher states do emit radiation with some probability, ultimately decaying down to the ground state).

6.2.4 Magnetic Dipole and Electric Quadrupole Radiation

The electric dipole approximation to radiation is sufficient for most applications. Obvious exceptions are when the dipole **p** vanishes or, for some reason, doesn't change in time. For completeness, we describe here the leading order corrections to the electric dipole approximations.

The Taylor expansion of the retarded potential was given in (6.13) and (6.14). Putting them together, we get

$$\begin{aligned}A_\mu(\mathbf{x},t) &= \frac{\mu_0}{4\pi}\int_A d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|}\\ &= \frac{\mu_0}{4\pi r}\int_A d^3x'\ \left(J_\mu(\mathbf{x}',t-r/c)+\dot{J}_\mu(\mathbf{x}',t-r/c)\,\frac{\mathbf{x}\cdot\mathbf{x}'}{rc}\right)\left(1+\frac{\mathbf{x}\cdot\mathbf{x}'}{r^2}\right)+\ldots\end{aligned}$$

The first term is the electric dipole approximation that we discussed in above. We will refer to this as $A_\mu^{\rm ED}$. Corrections to this arise as two Taylor series. Ultimately we will only be interested in the far-field region. At far enough distance, the terms in the first bracket will always dominate the terms in the second bracket, which are suppressed by $1/r$. We therefore have

$$A_\mu(\mathbf{x},t)\approx A_\mu^{\rm ED}(\mathbf{x},t)+\frac{\mu_0}{4\pi r^2 c}\int_A d^3x'\ (\mathbf{x}\cdot\mathbf{x}')\dot{J}_\mu(\mathbf{x}',t-r/c)$$

As in the electric dipole case, it's simplest if we focus on the vector potential

$$\mathbf{A}(\mathbf{x},t)\approx \mathbf{A}^{\rm ED}(\mathbf{x},t)+\frac{\mu_0}{4\pi r^2 c}\int d^3x'\ (\mathbf{x}\cdot\mathbf{x}')\,\dot{\mathbf{J}}(\mathbf{x}',t-r/c) \tag{6.21}$$

The integral involves the kind of expression that we met first when we discussed magnetic dipoles in Section 3.3.2. We use the slightly odd expression,

$$\partial_j(J_jx_ix_k)=(\partial_jJ_j)x_ix_k+J_ix_k+J_kx_i=-\dot{\rho}x_ix_k+J_ix_k+J_kx_i$$

Because $\mathbf{J}$ in (6.21) is a function of $\mathbf{x}'$, we apply this identity to the $J_ix'_j$ terms in the expression. We drop the boundary term at infinity, remembering that we're actually dealing with $\dot{J}$ rather than J, write the integral above as

$$\int d^3x'\ x_jx'_j\dot{J}_i=\frac{x_j}{2}\int d^3x'\ (x'_j\dot{J}_i-x'_i\dot{J}_j+\ddot{\rho}x'_ix'_j)$$

Then, using the appropriate vector product identity, we have

$$\int d^3x'\ (\mathbf{x}\cdot\mathbf{x}')\dot{\mathbf{J}}=\frac{1}{2}\mathbf{x}\times\int d^3x'\ \dot{\mathbf{J}}\times\mathbf{x}'+\frac{1}{2}\int d^3x'\ (\mathbf{x}\cdot\mathbf{x}')\mathbf{x}'\,\ddot{\rho}$$

Using this, we may write (6.21) as

$$\mathbf{A}(\mathbf{x},t)\approx\mathbf{A}^{\rm ED}(\mathbf{x},t)+\mathbf{A}^{\rm MD}(\mathbf{x},t)+\mathbf{A}^{EQ}(\mathbf{x},t)$$

where $\mathbf{A}^{\rm MD}$ is the *magnetic dipole* contribution and is given by

$$\mathbf{A}^{\rm MD}(\mathbf{x},t)=-\frac{\mu_0}{8\pi r^2c}\mathbf{x}\times\int d^3x'\ \mathbf{x}'\times\dot{\mathbf{J}}(\mathbf{x}',t-r/c) \tag{6.22}$$

and $\mathbf{A}^{\rm EQ}$ is the *electric quadrupole* contribution and is given by

$$\mathbf{A}^{\rm EQ}(\mathbf{x},t) = \frac{\mu_0}{8\pi r^2 c}\int d^3x'\;(\mathbf{x}\cdot\mathbf{x}')\mathbf{x}'\,\ddot{\rho}(\mathbf{x}',t-r/c) \tag{6.23}$$

The names we have given to each of these contributions will become clearer as we look at their properties in more detail.

Magnetic Dipole Radiation

Recall that, for a general current distribution, the magnetic dipole $\mathbf{m}$ is defined by

$$\mathbf{m} = \frac{1}{2}\int d^3x'\;\mathbf{x}'\times\mathbf{J}(\mathbf{x}')$$

The magnetic dipole contribution to radiation (6.22) can then be written as

$$\mathbf{A}^{\rm MD}(\mathbf{x},t) = -\frac{\mu_0}{4\pi rc}\,\hat{\mathbf{x}}\times\dot{\mathbf{m}}(t-r/c)$$

This means that varying loops of current will also emit radiation. Once again, the leading order contribution to the magnetic field, $\mathbf{B} = \nabla\times\mathbf{A}$, arises when the curl hits the argument of $\mathbf{m}$. We have

$$\mathbf{B}^{\rm MD}(\mathbf{x},t) \approx \frac{\mu_0}{4\pi rc^2}\,\hat{\mathbf{x}}\times(\hat{\mathbf{x}}\times\ddot{\mathbf{m}}(t-r/c))$$

Using the Maxwell equation $\dot{\mathbf{E}}^{\rm MD} = c^2\nabla\times\mathbf{B}^{\rm MD}$ to compute the electric field, we have

$$\mathbf{E}^{\rm MD}(\mathbf{x},t) \approx \frac{\mu_0}{4\pi rc}\,\hat{\mathbf{x}}\times\ddot{\mathbf{m}}(t-r/c)$$

The end result is very similar to the expression for $\mathbf{B}$ and $\mathbf{E}$ that we saw in (6.15) and (6.16) for the electric dipole radiation. This means that the radiated power has the same angular form, with the Poynting vector now given by

$$\mathbf{S}^{\rm MD} = \frac{\mu_0}{16\pi^2 r^2 c^3}|\ddot{\mathbf{m}}|^2\sin^2\theta\,\hat{\mathbf{z}} \tag{6.24}$$

Integrating over all space gives us the power emitted,

$$\mathcal{P}^{\rm MD} = \frac{\mu_0}{6\pi c^3}\,|\ddot{\mathbf{m}}|^2 \tag{6.25}$$

This takes the same form as the electric dipole result (6.18), but with the electric dipole replaced by the magnetic dipole. Notice, however, that for non-relativistic particles, the magnetic dipole radiation is substantially smaller than the electric dipole contribution. For a particle of charge Q, oscillating a distance d with frequency ω, we have $p\sim Qd$ and $m\sim Qd^2\omega$. This means that the ratio of radiated powers is

$$\frac{\mathcal{P}^{\rm MD}}{\mathcal{P}^{\rm ED}} \sim \frac{d^2\omega^2}{c^2} \sim \frac{v^2}{c^2}$$

where v is the speed of the particle.

Electric Quadrupole Radiation

The electric quadrupole tensor $\mathbb{Q}_{ij}$ arises as the $1/r^4$ term in the expansion of the electric field for a general, static charge distribution. It is defined by

$$\mathbb{Q}_{ij} = \int d^3x'\ \rho(\mathbf{x}')\left(3x_i'x_j' - \delta_{ij}x'^2\right)$$

This is not quite of the right form to account for the contribution to the potential (6.23). Instead, we have

$$A_i^{\rm EQ}(\mathbf{x},t) = -\frac{\mu_0}{24\pi r^2 c}\left(x_j\ddot{\mathbb{Q}}_{ij}(t-r/c) + x_i\int d^3x'\ x'^2\ddot{\rho}(x', t-r/c)\right)$$

The second term looks like a mess, but it doesn't do anything. This is because it's radial and so vanishes when we take the curl to compute the magnetic field. Neither does it contribute to the electric field which, in our case, we will again determine from the Maxwell equation. This means we are entitled to write

$$\mathbf{A}^{\rm EQ}(\mathbf{x},t) = -\frac{\mu_0}{24\pi r^2 c}\,\mathbf{x}\cdot\ddot{\mathbb{Q}}(t-r/c)$$

where $(\mathbf{x}\cdot\mathbb{Q})_i = x_j\mathbb{Q}_{ij}$. Correspondingly, the magnetic and electric fields at large distance are

$$\begin{aligned}\mathbf{B}^{\rm EQ}(\mathbf{x},t) &\approx \frac{\mu_0}{24\pi rc^2}\,\hat{\mathbf{x}}\times(\hat{\mathbf{x}}\cdot\ddot{\mathbb{Q}})\\ \mathbf{E}^{\rm EQ}(\mathbf{x},t) &\approx \frac{\mu_0}{24\pi rc}\left((\hat{\mathbf{x}}\cdot\ddot{\mathbb{Q}}\cdot\hat{\mathbf{x}})\hat{\mathbf{x}} - (\hat{\mathbf{x}}\cdot\ddot{\mathbb{Q}})\right)\end{aligned}$$

We may again compute the Poynting vector and radiated power. The details depend on the exact structure of $\mathbb{Q}$, but the angular dependence of the radiation is now different from that seen in the dipole cases.

Finally, you may wonder about the cross terms between the ED, MD and EQ components of the field strengths when computing the quadratic Poynting vector. It turns out that, courtesy of their different spatial structures, these cross-term vanish when computing the total integrated power.

6.2.5 An Application: Pulsars

Pulsars are lighthouses in the sky, spinning neutron stars continuously beaming out radiation which sweeps past our line of sight once every rotation. They have been observed with periods between 10^{-3} seconds and 8 seconds.

Neutron stars typically carry a very large magnetic field. This arises from the parent star which, as it collapses, reduces in size by a factor of about 10^5. This squeezes the magnetic flux lines, which gets multiplied by a factor of 10^{10}. The resulting magnetic field is typically around 10^8 Tesla, but can be as high as 10^{11} Tesla. For comparison, the highest magnetic field that we have succeeded in creating in a laboratory is a paltry 100 Tesla or so.

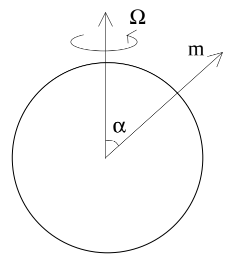


Figure 55:

The simplest model of a pulsar has the resulting magnetic dipole moment $\mathbf{m}$ of the neutron star misaligned with the angular velocity. This resulting magnetic dipole radiation creates the desired lighthouse effect. Consider the set-up shown in the picture. We take the pulsar to rotate about the z-axis with frequency Ω. The magnetic moment sits at an angle α relative to the z-axis, so rotates as

$$\mathbf{m} = m_0 \left(\sin(\alpha)\cos(\Omega t)\hat{\mathbf{x}} + \sin(\alpha)\cos(\Omega t)\hat{\mathbf{y}} + \cos\alpha\hat{\mathbf{z}}\right)$$

The power emitted (6.25) is then

$$\mathcal{P} = \frac{\mu_0}{6\pi c}\, m_0^2\Omega^2\sin^2\alpha$$

At the surface of the neutron star, it's reasonable to assume that the magnetic field is given by the dipole moment. In Section 3.3, we computed the magnetic field due to a dipole moment: it is

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

where R is the radius of the star. This means that $B_{\rm max} = \mu_0 m_0/2\pi R^3$ and the power emitted is

$$\mathcal{P} = \frac{2\pi R^6 B_{\rm max}^2}{3c\mu_0}\,\Omega^2\sin^2\alpha \tag{6.26}$$

Because the pulsar is emitting radiation, it must lose energy. And this means it slows down. The rotational energy of a the pulsar is given by

$$E = \frac{1}{2}I\Omega^2$$

where $I = \frac{2}{5}MR^2$ is the moment of inertia of a sphere of mass M and radius R. Equating the power emitted with the loss of rotational kinetic energy gives

$$\mathcal{P} = -\dot{E} = -I\Omega\dot{\Omega} \tag{6.27}$$

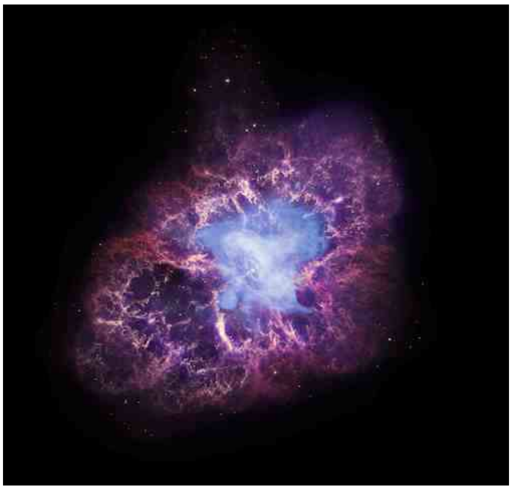


Figure 56: A composite image of the Crab Nebula, taken by the Hubble, Chandra and Spitzer space telescopes.

Let's put some big numbers into these equations. In 1054, Chinese astronomers saw a new star appear in the sky. 6500 light years away, a star had gone supernova. It left behind a pulsar which, today, emits large quantities of radiation, illuminating the part of the sky we call the Crab nebula. This is shown in the picture.

The Crab pulsar has mass $M \approx 1.4 M_{\rm Sun} \approx 3 \times 10^{30}\ Kg$ and radius $R \approx 15\ km$. It spins about 30 times a second, so $\Omega \approx 60\pi\ s^{-1}$. It's also seen to be slowing down with $\dot{\Omega} = -2 \times 10^{-9}\ s^{-2}$. From this information alone, we can calculate that it loses energy at a rate of $\dot{E} = I\Omega\dot{\Omega} \approx -10^{32}\ Js^{-1}$. That's a whopping amount of energy to be losing every second. In fact, it's enough energy to light up the entire Crab nebula. Which, of course, it has to be! Moreover, we can use (6.26) and (6.27) to estimate the magnetic field on the surface of the pulsar. Plugging in the numbers give $B_{\rm max} \sin\alpha \approx 10^8$ Tesla.

6.3 Scattering

In this short section, we describe the application of our radiation formulae to the phenomenon of *scattering*. Here's the set-up: an electromagnetic wave comes in and hits a particle. In response, the particle oscillates and, in doing so, radiates. This new radiation moves out in different directions from the incoming wave. This is the way that light is scattered.

6.3.1 Thomson Scattering

We start by considering free, charged particles where the process is known as Thomson scattering. The particles respond to an electric field by accelerating, as dictated by Newton's law

$$m\ddot{\mathbf{x}} = q\mathbf{E}$$

The incoming radiation takes the form $\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$. To solve for the motion of the particle, we're going to assume that it doesn't move very far from its central position, which we can take to be the origin $\mathbf{r} = 0$. Here, "not very far" means small compared to the wavelength of the electric field. In this case, we can replace the electric field by $\mathbf{E} \approx \mathbf{E}_0 e^{-i\omega t}$, and the particle undergoes simple harmonic motion

$$\mathbf{x}(t) = -\frac{q\mathbf{E}_0}{m\omega^2}\sin(\omega t)$$

We should now check that the motion of the particle is indeed small compared to the wavelength of light. The maximum distance that the particle gets is $x_{\rm max} = qE_0/m\omega^2$, so our analysis will only be valid if we satisfy

$$\frac{qE_0}{m\omega^2} \ll \frac{c}{\omega} \quad \Rightarrow \quad \frac{qE_0}{m\omega c} \ll 1 \tag{6.28}$$

This requirement has a happy corollary, since it also ensures that the maximum speed of the particle $v_{\rm max} = qE_0/m\omega \ll c$, so the particle motion is non-relativistic. This means that we can use the dipole approximation to radiation that we developed in the previous section. We computed the time-averaged radiated power in (6.20): it is given by

$$\bar{\mathcal{P}}_{\rm radiated} = \frac{\mu_0 q^4 E_0^2}{12\pi m^2 c}$$

It's often useful to compare the strength of the emitted radiation to that of the incoming radiation. The relevant quantity to describe the incoming radiation is the time-averaged magnitude of the Poynting vector. Recall from Section 4.4 that the Poynting vector for a wave with wavevector $\mathbf{k}$ is

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{cE_0^2}{\mu_0}\hat{\mathbf{k}}\,\sin^2(\mathbf{k}\cdot\mathbf{x}-\omega t)$$

Taking the time average over a single period, $T = 2\pi/\omega$, gives us the average energy flux of the incoming radiation,

$$\bar{S}_{\rm incident} = \frac{cE_0^2}{2\mu_0}$$

with the factor of two coming from the averaging. The ratio of the outgoing to incoming powers is called the *cross-section* for scattering. It is given by

$$\sigma = \frac{\bar{\mathcal{P}}_{\text{radiated}}}{\bar{S}_{\text{incident}}} = \frac{\mu_0^2 q^4}{6\pi m^2 c^2}$$

The cross-section has the dimensions of area. To highlight this, it's useful to write it as

$$\sigma = \frac{8\pi}{3} r_q^2 \tag{6.29}$$

where the length scale r_q is known as the *classical radius* of the particle and is given by

$$\frac{q^2}{4\pi\epsilon_0 r_q} = mc^2$$

This equation tells us how to think of r_q. Up to some numerical factors, it equates the Coulomb energy of a particle in a ball of size r_q with its relativistic rest mass. Ultimately, this is not the right way to think of the size of point particles. (The right way involves quantum mechanics). But it is a useful concept in the classical world. For the electron, $r_e \approx 2.8 \times 10^{-15}\ m$.

The Thompson cross-section (6.29) is slightly smaller than the (classical) geometric cross-section of the particle (which would be the area of the disc, $4\pi r_q^2$). For us, the most important point is that the cross-section does not depend on the frequency of the incident light. It means that all wavelengths of light are scattered equally by free, charged particles, at least within the regime of validity (6.28). For electrons, the Thomson cross-section is $\sigma \approx 6 \times 10^{-30}\ m^2$.

6.3.2 Rayleigh Scattering

Rayleigh scattering describes the scattering of light off a neutral atom or molecule. Unlike in the case of Thomson scattering, the centre of mass of the atom does not accelerate. Instead, as we saw in Section 7.1.1, the atom undergoes polarisation

$$\mathbf{p} = \alpha \mathbf{E}$$

We presented a simple atomic model to compute the proportionality constant in Section 7.5.1, where we showed that it takes the form (7.29),

$$\alpha = \frac{q^2/m}{-\omega^2 + \omega_0^2 - i\gamma\omega}$$

Figure 57: Now you know why.

Here ω_0 is the natural oscillation frequency of the atom while ω is the frequency of incoming light. For many cases of interest (such as visible light scattering off molecules in the atmosphere), we have $\omega_0 \gg \omega$, and we can approximate α as a constant,

$$\alpha \approx \frac{q^2}{\omega_0^2 m}$$

We can now compute the time-average power radiated in this case. It's best to use the version of Larmor's fomula involving the electric dipole (6.19), since we can just substitute in the results above. We have

$$\bar{\mathcal{P}}_{\rm radiated} = \frac{\mu_0 \alpha^2 E_0^2 \omega^4}{12\pi c}$$

In this case, the cross-section for Rayleigh scattering is given by

$$\sigma = \frac{\bar{\mathcal{P}}_{\rm radiated}}{\bar{\mathcal{S}}_{\rm incident}} = \frac{\mu_0^2 q^4}{6\pi m^2 c^2}\left(\frac{\omega}{\omega_0}\right)^4 = \frac{8\pi r_q^2}{3}\left(\frac{\omega}{\omega_0}\right)^4$$

We see that the cross-section now has more structure. It increases for high frequencies, $\sigma \sim \omega^4$ or, equivalently, for short wavelengths $\sigma \sim 1/\lambda^4$. This is important. The most famous example is the colour of the sky. Nitrogen and oxygen in the atmosphere scatter short-wavelength blue light more than the long-wavelength red light. This means that the blue light from the Sun gets scattered many times and so appears to come from all regions of the sky. In contrast, the longer wavelength red and yellow light gets scattered less, which is why the Sun appears to be yellow. (In the absence of an atmosphere, the light from the Sun would be more or less white). This effect is particularly apparent at sunset, when the light from the Sun passes through a much larger slice of atmosphere and, correspondingly, much more of the blue light is scattered, leaving behind only red.

6.4 Radiation From a Single Particle

In the previous section, we have developed the multipole expansion for radiation emitted from a source. We needed to invoke a couple of approximations. First, we assumed that we were far from the source. Second, we assumed that the motion of charges and currents within the source was non-relativistic.

In this section, we're going to develop a formalism which does not rely on these approximations. We will determine the field generated by a particle with charge q, moving on an arbitrary trajectory $\mathbf{r}(r)$, with velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}(t)$. It won't matter how far we are from the particle; it won't matter how fast the particle is moving. The particle has charge density

$$\rho(\mathbf{x}, t) = q\delta^3(\mathbf{x} - \mathbf{r}(t)) \tag{6.30}$$

and current

$$\mathbf{J}(\mathbf{x}, t) = q\,\mathbf{v}(t)\delta^3(\mathbf{x} - \mathbf{r}(t)) \tag{6.31}$$

Our goal is find the general solution to the Maxwell equations by substituting these expressions into the solution (6.7) for the retarded potential,

$$A_\mu(\mathbf{x}, t) = \frac{\mu_0}{4\pi} \int d^3x' \, \frac{J_\mu(\mathbf{x}', t_{\rm ret})}{|\mathbf{x} - \mathbf{x}'|} \tag{6.32}$$

The result is known as Liénard-Wierchert potentials.

6.4.1 Liénard-Wierchert Potentials

If we simply plug (6.30) into the expression for the retarded electric potential (6.32), we get

$$\phi(\mathbf{x}, t) = \frac{q}{4\pi\epsilon_0} \int d^3x' \, \frac{1}{|\mathbf{x} - \mathbf{x}'|} \, \delta^3(\mathbf{x}' - \mathbf{r}(t_{\rm ret}))$$

Here we're denoting the position of the particle as $\mathbf{r}(t)$, while we're interested in the value of the electric potential at some different point $\mathbf{x}$ which tdoes not lie on the trajectory $\mathbf{r}(t)$. We can use the delta-function to do the spatial integral, but it's a little cumbersome because the $\mathbf{x}'$ appears in the argument of the delta-function both in the obvious place, and also in $t_{\rm ret} = t - |\mathbf{x}-\mathbf{x}'|/c$. It turns out to be useful to shift this awkwardness into a slightly different delta-function over time. We write,

$$\begin{aligned}\phi(\mathbf{x},t) &= \frac{q}{4\pi\epsilon_0}\int dt'\int d^3x'\ \frac{1}{|\mathbf{x}-\mathbf{x}'|}\,\delta^3(\mathbf{x}'-\mathbf{r}(t'))\delta(t'-t_{\rm ret})\\ &= \frac{q}{4\pi\epsilon_0}\int dt'\ \frac{1}{|\mathbf{x}-\mathbf{r}(t')|}\,\delta(t-t'-|\mathbf{x}-\mathbf{r}(t')|/c)\end{aligned} \tag{6.33}$$

We still have the same issue in doing the $\int dt'$ integral, with t' appearing in two places in the argument. But it's more straightforward to see how to deal with it. We introduce the separation vector

$$\mathbf{R}(t) = \mathbf{x} - \mathbf{r}(t)$$

Then, if we define $f(t') = t' + R(t')/c$, we can write

$$\begin{aligned}\phi(\mathbf{x},t) &= \frac{q}{4\pi\epsilon_0}\int dt'\ \frac{1}{R(t')}\,\delta(t-f(t'))\\ &= \frac{q}{4\pi\epsilon_0}\int df\ \frac{dt'}{df}\frac{1}{R(t')}\,\delta(t-f(t'))\\ &= \frac{q}{4\pi\epsilon_0}\left[\frac{dt'}{df}\frac{1}{R(t')}\right]_{f(t')=t}\end{aligned}$$

A quick calculation gives

$$\frac{df}{dt'} = 1 - \frac{\hat{\mathbf{R}}(t')\cdot\mathbf{v}(t')}{c}$$

with $\mathbf{v}(t) = \dot{\mathbf{r}}(t) = -\dot{\mathbf{R}}(t)$. This leaves us with our final expression for the scalar potential

$$\phi(\mathbf{x},t) = \frac{q}{4\pi\epsilon_0}\left[\frac{c}{c-\hat{\mathbf{R}}(t')\cdot\mathbf{v}(t')}\,\frac{1}{R(t')}\right]_{\rm ret} \tag{6.34}$$

Exactly the same set of manipulations will give us a similar expression for the vector potential,

$$\mathbf{A}(\mathbf{x},t) = \frac{q\mu_0}{4\pi}\left[\frac{c}{c-\hat{\mathbf{R}}(t')\cdot\mathbf{v}(t')}\,\frac{\mathbf{v}(t')}{R(t')}\right]_{\rm ret} \tag{6.35}$$

Equations (6.34) and (6.35) are the *Liénard-Wierchert potentials*. In both expressions "ret" stands for "retarded" and means that they should be evaluated at time t' determined by the requirement that

$$t' + R(t')/c = t \tag{6.36}$$

This equation has an intuitive explanation. If you trace back light-sheets from the point $\mathbf{x}$, they intersect the trajectory of the particle at time t', as shown in the figure. The Liénard-Wierchert potentials are telling us that the field at point $\mathbf{x}$ is determined by what the particle was doing at this time t'.

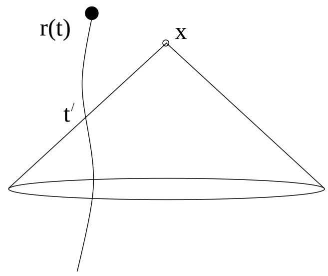


Figure 58:

6.4.2 A Simple Example: A Particle Moving with Constant Velocity

The Liénard-Wierchert potentials (6.34) and (6.35) have the same basic structure that we for the Coulomb law in electrostatics and the Biot-Savart law in magnetostatics. The difference lies in the need to evaluate the potentials at time t'. But there is also the extra factor $1/(1-\hat{\mathbf{R}}\cdot\mathbf{v}/c)$. To get a feel for this, let's look at a simple example. We'll take a particle which moves at constant speed in the $\hat{\mathbf{z}}$ direction, so that

$$\mathbf{r}(t) = vt\hat{\mathbf{z}} \quad \Rightarrow \quad \mathbf{v}(t) = v\hat{\mathbf{z}}$$

To simplify life even further, we'll compute the potentials at a point in the $z = 0$ plane, so that $\mathbf{x} = (x, y, 0)$. We'll ask how the fields change as the particle passes through. The equation (6.36) to determine the retarded time becomes

$$t' + \sqrt{x^2 + y^2 + v^2 t'^2}/c = t$$

Squaring this equation (after first making the right-hand side $t - t'$) gives us a quadratic in t',

$$t'^2 - 2\gamma^2 t t' + \gamma^2(t^2 - r^2/c^2) = 0$$

where we see the factor $\gamma = (1 - v^2/c^2)^{-1/2}$, familiar from special relativity naturally emerging. The quadratic has two roots. We're interested in the one with the minus sign, corresponding to the retarded time. This is

$$t' = \gamma^2 t - \frac{\gamma}{c}\sqrt{v^2t^2 + r^2/\gamma^2} \tag{6.37}$$

We now need to deal with the various factors in the numerator of the Liénard-Wierchert potential (6.34). Pleasingly, they combine together nicely. We have $R(t') = c(t - t')$. Meanwhile, $\mathbf{R}(t')\cdot\mathbf{v}(t') = (\mathbf{x} - \mathbf{r}(t'))\cdot\mathbf{v} = -\mathbf{r}(t')\cdot\mathbf{v} = -v^2 t'$ since we've taken $\mathbf{x}$ to lie perpendicular to $\mathbf{v}$. Put together, this gives us

$$\begin{aligned}\phi(\mathbf{x},t) &= \frac{q}{4\pi\epsilon_0}\frac{1}{[1+v^2t'/c(t-t')]}\frac{1}{c(t-t')}\\ &= \frac{q}{4\pi\epsilon_0}\frac{1}{c(t-t')+v^2t'}\\ &= \frac{1}{4\pi\epsilon_0}\frac{1}{c(t-t'/\gamma^2)}\end{aligned}$$

But, using our solution (6.37), this becomes

$$\phi(\mathbf{x},t) = \frac{q}{4\pi\epsilon_0}\frac{1}{[v^2t^2+(x^2+y^2)/\gamma^2]}$$

Similarly, the vector potential is

$$\mathbf{A}(\mathbf{x},t) = \frac{q\mu_0}{4\pi}\frac{\mathbf{v}}{[v^2t^2+(x^2+y^2)/\gamma^2]}$$

How should we interpret these results? The distance from the particle to the point $\mathbf{x}$ is $r^2 = x^2+y^2+v^2t^2$. The potentials look very close to those due to a particle a distance r away, but with one difference: there is a contraction in the x and y directions. Of course, we know very well what this means: it is the usual Lorentz contraction in special relativity.

In fact, we previously derived the expression for the electric and magnetic field of a moving particle in Section 5.3.4, simply by acting with a Lorentz boost on the static fields. The calculation here was somewhat more involved, but it didn't assume any relativity. Instead, the Lorentz contraction follows only by solving the Maxwell equations. Historically, this kind of calculation is how Lorentz first encountered his contractions.

6.4.3 Computing the Electric and Magnetic Fields

We now compute the electric and magnetic fields due to a particle undergoing arbitrary motion. In principle this is straightforward: we just need to take our equations (6.34) and (6.35)

$$\begin{aligned}\phi(\mathbf{x},t) &= \frac{q}{4\pi\epsilon_0}\left[\frac{c}{c-\hat{\mathbf{R}}(t')\cdot\mathbf{v}(t')}\frac{1}{R(t')}\right]_{\text{ret}}\\ \mathbf{A}(\mathbf{x},t) &= \frac{q\mu_0}{4\pi}\left[\frac{c}{c-\hat{\mathbf{R}}(t')\cdot\mathbf{v}(t')}\frac{\mathbf{v}(t')}{R(t')}\right]_{\text{ret}}\end{aligned}$$

where $\mathbf{R}(t') = \mathbf{x}-\mathbf{r}(t')$. We then plug these into the standard expressions for the electric field $\mathbf{E} = -\nabla\phi - \partial\mathbf{A}/\partial t$ and the magnetic field $\mathbf{B} = \nabla\times\mathbf{A}$. However, in practice, this is a little fiddly. It's because the terms in these equations are evaluated at the retarded time t' determined by the equation $t' + R(t')/c = t$. This means that when we differentiate (either by $\partial/\partial t$ or by ∇), the retarded time also changes and so gives a contribution. It turns out that it's actually simpler to return to our earlier expression (6.33),

$$\phi(\mathbf{x}, t) = \frac{q}{4\pi\epsilon_0} \int dt' \; \frac{1}{R(t')} \, \delta(t - t' - R(t')/c)$$

and a similar expression for the vector potential,

$$\mathbf{A}(\mathbf{x}, t) = \frac{q\mu_0}{4\pi} \int dt' \; \frac{\mathbf{v}(t')}{R(t')} \delta(t - t' - R(t')/c) \tag{6.38}$$

This will turn out to be marginally easier to deal with.

The Electric Field

We start with the electric field $\mathbf{E} = -\nabla\phi - \partial\mathbf{A}/\partial t$. We call the argument of the delta-function

$$s = t - t' - R(t')$$

We then have

$$\begin{aligned} \nabla\phi &= \frac{q}{4\pi\epsilon_0} \int dt' \; \left[-\frac{\nabla R}{R^2} \delta(s) - \frac{1}{R} \delta'(s) \frac{\nabla R}{c} \right] \\ &= \frac{q}{4\pi\epsilon_0} \int ds \; \left| \frac{\partial t'}{\partial s} \right| \left[-\frac{\nabla R}{R^2} \delta(s) - \frac{\nabla R}{Rc} \delta'(s) \right] \end{aligned} \tag{6.39}$$

The Jacobian factor from changing the integral variable is the given by

$$\frac{\partial s}{\partial t'} = -1 + \hat{\mathbf{R}}(t') \cdot \mathbf{v}(t')/c$$

This quantity will appear a lot in what follows, so we give it a new name. We define

$$\kappa = 1 - \hat{\mathbf{R}}(t') \cdot \mathbf{v}(t')/c$$

so that $\partial t'/\partial s = -1/\kappa$. Integrating the second term in (6.39) by parts, we can then write

$$\begin{aligned} \nabla\phi &= \frac{q}{4\pi\epsilon_0} \int ds \; \left[-\frac{\nabla R}{\kappa R^2} + \frac{d}{ds} \left(\frac{\nabla R}{\kappa R c} \right) \right] \delta(s) \\ &= \frac{q}{4\pi\epsilon_0} \int ds \; \left[-\frac{\nabla R}{\kappa R^2} - \frac{1}{\kappa} \frac{d}{dt'} \left(\frac{\nabla R}{\kappa R c} \right) \right] \delta(s) \end{aligned}$$

Meanwhile, the vector potential term gives

$$\frac{\partial \mathbf{A}}{\partial t} = \frac{q\mu_0}{4\pi}\int dt'\ \frac{\mathbf{v}}{R}\delta'(s)\frac{\partial s}{\partial t}$$

But $\partial s/\partial t = 1$. Moving forward, we have

$$\begin{aligned}\frac{\partial \mathbf{A}}{\partial t} &= \frac{q\mu_0}{4\pi}\int ds\ \left|\frac{\partial t'}{\partial s}\right|\ \frac{\mathbf{v}}{R}\,\delta'(s)\\ &= -\frac{q\mu_0}{4\pi}\int ds\ \left[\frac{d}{ds}\left(\frac{\mathbf{v}}{\kappa R}\right)\right]\delta(s)\\ &= \frac{q\mu_0}{4\pi}\int ds\ \frac{1}{\kappa}\left[\frac{d}{dt'}\left(\frac{\mathbf{v}}{\kappa R}\right)\right]\delta(s)\end{aligned}$$

Putting this together, we get

$$\begin{aligned}\mathbf{E} &= \frac{q}{4\pi\epsilon_0}\int ds\ \left[\frac{\nabla R}{\kappa R^2} + \frac{1}{\kappa c}\frac{d}{dt'}\left(\frac{\nabla R - \mathbf{v}/c}{\kappa R}\right)\right]\delta(s)\\ &= \frac{q}{4\pi\epsilon_0}\left[\frac{\hat{\mathbf{R}}}{\kappa R^2} + \frac{1}{\kappa c}\frac{d}{dt'}\left(\frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa R}\right)\right]_{\rm ret}\end{aligned} \tag{6.40}$$

We're still left with some calculations to do. Specifically, we need to take the derivative d/dt'. This involves a couple of small steps. First,

$$\frac{d\hat{\mathbf{R}}}{dt'} = \frac{d}{dt'}\left(\frac{\mathbf{R}}{R}\right) = -\frac{\mathbf{v}}{R} + \frac{\mathbf{R}}{R^2}(\hat{\mathbf{R}}\cdot\mathbf{v}) = -\frac{1}{R}\left(\mathbf{v} - (\mathbf{v}\cdot\hat{\mathbf{R}})\hat{\mathbf{R}}\right)$$

Also,

$$\frac{d}{dt'}(\kappa R) = \frac{d}{dt'}\left(R - \mathbf{R}\cdot\mathbf{v}/c\right) = -\mathbf{v}\cdot\hat{\mathbf{R}} + v^2/c - \mathbf{R}\cdot\mathbf{a}/c$$

Putting these together, we get

$$\frac{d}{dt'}\left(\frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa R}\right) = -\frac{1}{\kappa R^2}\left(\mathbf{v} - (\mathbf{v}\cdot\hat{\mathbf{R}})\hat{\mathbf{R}}\right) - \frac{\mathbf{a}}{\kappa R c} + \frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa^2 R^2}\left(\mathbf{v}\cdot\hat{\mathbf{R}} - v^2/c + \mathbf{R}\cdot\mathbf{a}/c\right)$$

We write the $\mathbf{v}\cdot\hat{\mathbf{R}}$ terms as $\mathbf{v}\cdot\hat{\mathbf{R}} = c(1-\kappa)$. Then, expanding this out, we find that a bunch of terms cancel, until we're left with

$$\begin{aligned}\frac{d}{dt'}\left(\frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa R}\right) &= -\frac{c\hat{\mathbf{R}}}{R^2} + \frac{c(\hat{\mathbf{R}} - \mathbf{v}/c)}{\kappa^2 R^2}(1 - v^2/c^2) + \frac{1}{\kappa^2 R c}\left[(\hat{\mathbf{R}} - \mathbf{v}/c)\,\hat{\mathbf{R}}\cdot\mathbf{a} - \kappa a\right]\\ &= -\frac{c\hat{\mathbf{R}}}{R^2} + \frac{c(\hat{\mathbf{R}} - \mathbf{v}/c)}{\gamma^2\kappa^2 R^2} + \frac{\hat{\mathbf{R}}\times[(\hat{\mathbf{R}} - \mathbf{v}/c)\times\mathbf{a}]}{\kappa^2 R c}\end{aligned} \tag{6.41}$$

where we've introduced the usual γ factor from special relativity: $\gamma^2 = 1/(1 - v^2/c^2)$. Now we can plug this into (6.40) to find our ultimate expression for the electric field,

$$\mathbf{E}(\mathbf{x},t) = \frac{q}{4\pi\epsilon_0}\left[\frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\gamma^2\kappa^3 R^2} + \frac{\hat{\mathbf{R}} \times [(\hat{\mathbf{R}} - \mathbf{v}/c) \times \mathbf{a}]}{\kappa^3 R c^2}\right]_{\rm ret} \tag{6.42}$$

Since it's been a long journey, let's recall what everything in this expression means. The particle traces out a trajectory $\mathbf{r}(t)$, while we sit at some position $\mathbf{x}$ which is where the electric field is evaluated. The vector $\mathbf{R}(t)$ is the difference: $\mathbf{R} = \mathbf{x} - \mathbf{r}$. The *ret* subscript means that we evaluate everything in the square brackets at time t', determined by the condition $t' + R(t')/c = t$. Finally,

$$\kappa = 1 - \frac{\hat{\mathbf{R}}\cdot\mathbf{v}}{c} \quad \text{ and } \quad \gamma^2 = \frac{1}{1 - v^2/c^2}$$

The electric field (6.42) has two terms.

- The first term drops off as $1/R^2$. This is what becomes of the usual Coulomb field. It can be thought of as the part of the electric field that remains bound to the particle. The fact that it is proportional to $\hat{\mathbf{R}}$, with a slight off-set from the velocity, means that it is roughly isotropic.

- The second term drops off as $1/R$ and is proportional to the acceleration. This describes the radiation emitted by the particle. Its dependence on the acceleration means that it's highly directional.

The Magnetic Field

To compute the magnetic field, we start with the expression (6.38),

$$\mathbf{A}(\mathbf{x},t) = \frac{q\mu_0}{4\pi}\int dt' \, \frac{\mathbf{v}(t')}{R(t')}\delta(s)$$

with $s = t - t' - R(t')/c$. Then, using similar manipulations to those above, we have

$$\begin{aligned}\mathbf{B} = \nabla\times\mathbf{A} &= \frac{q\mu_0}{4\pi}\int dt' \, \left[-\frac{\nabla R}{R^2}\times\mathbf{v}\,\delta(s) + \frac{\nabla s\times\mathbf{v}}{R}\,\delta'(s)\right] \\ &= \frac{q\mu_0}{4\pi}\int ds \, \left[-\frac{\nabla R}{\kappa R^2}\times\mathbf{v} - \frac{1}{\kappa}\frac{d}{dt'}\left(\frac{\nabla R\times\mathbf{v}}{\kappa R c}\right)\right]\delta(s) \end{aligned} \tag{6.43}$$

We've already done the hard work necessary to compute this time derivative. We can write,

$$\frac{d}{dt'}\left(\frac{\nabla R \times \mathbf{v}}{\kappa R}\right) = \frac{d}{dt'}\left(\frac{(\hat{\mathbf{R}} - \mathbf{v}/c) \times \mathbf{v}}{\kappa R}\right)$$
$$= \frac{d}{dt'}\left(\frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa R}\right) \times \mathbf{v} + \frac{\hat{\mathbf{R}} - \mathbf{v}/c}{\kappa R} \times \mathbf{a}$$

Now we can use (6.41). A little algebra shows that terms of the form $\mathbf{v} \times \mathbf{a}$ cancel, and we're left with

$$\frac{d}{dt'}\left(\frac{\hat{\mathbf{R}} \times \mathbf{v}}{\kappa R}\right) = -\frac{c\hat{\mathbf{R}} \times \mathbf{v}}{R^2} + \frac{c\hat{\mathbf{R}} \times \mathbf{v}}{\gamma^2\kappa^2 R^2} + \frac{(\mathbf{R} \cdot \mathbf{a})\,\hat{\mathbf{R}} \times \mathbf{v}}{c\kappa^2 R^2} + \frac{\hat{\mathbf{R}} \times \mathbf{a}}{\kappa R}$$

Substituting this into (6.43), a little re-arranging of the terms gives us our final expression for the magnetic field,

$$\mathbf{B} = -\frac{q\mu_0}{4\pi}\left[\frac{\hat{\mathbf{R}} \times \mathbf{v}}{\gamma^2\kappa^3 R^2} + \frac{(\hat{\mathbf{R}} \cdot \mathbf{a})(\hat{\mathbf{R}} \times \mathbf{v}/c) + \kappa\hat{\mathbf{R}} \times \mathbf{a}}{c\kappa^3 R}\right]_{\text{ret}} \tag{6.44}$$

We see that this has a similar form to the electric field (6.42). The first term falls off as $1/R^2$ and is bound to the particle. It vanishes when $\mathbf{v} = 0$ which tells us that a charged particle only gives rise to a magnetic field when it moves. The second term falls off as $1/R$. This is generated by the acceleration and describes the radiation emitted by the particle. You can check that $\mathbf{E}$ in (6.42) and $\mathbf{B}$ in (6.44) are related through

$$\mathbf{B} = \frac{1}{c}[\hat{\mathbf{R}}]_{\text{ret}} \times \mathbf{E} \tag{6.45}$$

as you might expect.

6.4.4 A Covariant Formalism for Radiation

Before we make use of the Liénard-Wierchert potentials, we're going to do something a little odd: we're going to derive them again. This time, however, we'll make use of the Lorentz invariant notation of electromagnetism. This won't teach us anything new about physics and the results of this section aren't needed for what follows. But it will give us some practice on manipulating these covariant quantities. Moreover, the final result will be pleasingly concise.

A Covariant Retarded Potential

We start with our expression for the retarded potential (6.32) in terms of the current,

$$A_\mu(\mathbf{x},t) = \frac{\mu_0}{4\pi}\int d^3x'\ \frac{J_\mu(\mathbf{x}',t_{\rm ret})}{|\mathbf{x}-\mathbf{x}'|} \tag{6.46}$$

with $t_{\rm ret} = t - |\mathbf{x}-\mathbf{x}'|/c$. This has been the key formula that we've used throughout this section. Because it was derived from the Maxwell equations, this formula should be Lorentz covariant, meaning that someone in a different inertial frame will write down the same equation. Although this *should* be true, it's not at all obvious from the way that (6.46) is written that it actually is true. The equation involves only integration over space, and the denominator depends only on the spatial distance between two points. Neither of these are concepts that different observers agree upon.

So our first task is to rewrite (6.46) in a way which is manifestly Lorentz covariant. To do this, we work with four-vectors $X^\mu = (ct,\mathbf{x})$ and take a quantity which everyone agrees upon: the spacetime distance between two points

$$(X-X')^2 = \eta_{\mu\nu}(X^\mu - X'^\mu)(X^\nu - X'^\nu) = c^2(t-t')^2 - |\mathbf{x}-\mathbf{x}'|^2$$

Consider the delta-function $\delta((X-X')^2)$, which is non-vanishing only when X and X' are null-separated. This is a Lorentz-invariant object. Let's see what it looks like when written in terms of the time coordinate t. We will need the general result for delta-functions

$$\delta(f(x)) = \sum_{x_i}\frac{\delta(x-x_i)}{|f'(x_i)|} \tag{6.47}$$

where the sum is over all roots $f(x_i)=0$. Using this, we can write

$$\begin{aligned}
\delta\left((X-X')^2\right) &= \delta\left([c(t'-t)+|\mathbf{x}-\mathbf{x}'|][c(t'-t)-|\mathbf{x}-\mathbf{x}'|]\right) \\
&= \frac{\delta(ct'-ct+|\mathbf{x}-\mathbf{x}'|)}{2c|t-t'|} + \frac{\delta(ct'-ct-|\mathbf{x}-\mathbf{x}'|)}{2c|t-t'|} \\
&= \frac{\delta(ct'-ct+|\mathbf{x}-\mathbf{x}'|)}{2|\mathbf{x}-\mathbf{x}'|} + \frac{\delta(ct'-ct-|\mathbf{x}-\mathbf{x}'|)}{2|\mathbf{x}-\mathbf{x}'|}
\end{aligned}$$

The argument of the first delta-function is $ct'-ct_{\rm ret}$ and so this term contributes only if $t'<t$. The argument of the second delta-function is $ct'-ct_{\rm adv}$ and so this term can contribute only contribute if $t'>t$. But the temporal ordering of two spacetime points is also something all observers agree upon, as long as those points are either timelike or null separated. And here the delta-function requires the points to be null separated.

This means that if we picked just one of these terms, that choice would be Lorentz invariant. Mathematically, we do this using the Heaviside step-function

$$\Theta(t-t') = \begin{cases} 1 & t > t' \\ 0 & t < t' \end{cases}$$

We have

$$\delta\left((X-X')^2\right)\Theta\left(t-t'\right) = \frac{\delta(ct'-ct_{\rm ret})}{2|\mathbf{x}-\mathbf{x}'|} \tag{6.48}$$

The left-hand side is manifestly Lorentz invariant. The right-hand side doesn't look Lorentz invariant, but this formula tells us that it must be! Now we can make use of this to rewrite (6.46) in a way that the Lorentz covariance is obvious. It is

$$A_\mu(X) = \frac{\mu_0}{2\pi}\int d^4X'\; J_\mu(X')\,\delta\left((X-X')^2\right)\Theta\left(t-t'\right) \tag{6.49}$$

where the integration is now over spacetime, $d^4X' = c\,dt'\,d^3x'$. The combination of the delta-function and step-functions ensure that this integration is limited to the past light-cone of a point.

A Covariant Current

Next, we want a covariant expression for the current formed by a moving charged particle. We saw earlier that a particle tracing out a trajectory $\mathbf{y}(t)$ gives rise to a charge density (6.30) and current (6.31) given by

$$\rho(\mathbf{x},t) = q\,\delta^3(\mathbf{x}-\mathbf{y}(t)) \quad \text{and} \quad \mathbf{J}(\mathbf{x},t) = q\,\mathbf{v}(t)\,\delta^3(\mathbf{x}-\mathbf{y}(t)) \tag{6.50}$$

(We've changed notation from $\mathbf{r}(t)$ to $\mathbf{y}(t)$ to denote the trajectory of the particle). How can we write this in a manifestly covariant form?

We know from our course on Special Relativity that the best way to parametrise the worldline of a particle is by using its proper time τ. We'll take the particle to have trajectory $Y^\mu(\tau) = (ct(\tau), \mathbf{y}(\tau))$. Then the covariant form of the current is

$$J^\mu(X) = qc\int d\tau\;\dot{Y}^\mu(\tau)\,\delta^4(X^\nu - Y^\nu(\tau)) \tag{6.51}$$

It's not obvious that (6.51) is the same as (6.50). To see that it is, we can decompose the delta-function as

$$\delta^4(X^\nu - Y^\nu(\tau)) = \delta(ct - Y^0(\tau))\,\delta^3(\mathbf{x}-\mathbf{y}(\tau))$$

The first factor allows us to do the integral over $d\tau$, but at the expense of picking up a Jacobian-like factor $1/\dot{Y}^0$ from (6.47). We have

$$J^\mu = \frac{qc\dot{Y}^\mu}{\dot{Y}^0}\,\delta^3(\mathbf{x} - \mathbf{y}(\tau))$$

which does give us back the same expressions (6.50).

Covariant Liènard-Wierchert Potentials

We can now combine (6.49) and (6.51) to get the retarded potential,

$$\begin{aligned} A^\mu(X) &= \frac{\mu_0 qc}{4\pi}\int d^4X' \int d\tau\ \dot{Y}^\mu(\tau)\ \delta^4(X'^\nu - Y^\nu(\tau))\frac{\delta(ct' - ct_{\rm ret})}{|\mathbf{x} - \mathbf{x}'|} \\ &= \frac{\mu_0 qc}{4\pi}\int d\tau\ \dot{Y}^\mu(\tau)\ \frac{\delta(ct - Y^0(\tau) - |\mathbf{x} - \mathbf{y}(\tau)|}{|\mathbf{x} - \mathbf{y}(\tau)|} \end{aligned}$$

This remaining delta-function implicitly allows us to do the integral over proper time. Using (6.48) we can rewrite it as

$$\frac{\delta(ct - Y^0(\tau) - |\mathbf{x} - \mathbf{y}(\tau)|)}{2|\mathbf{x} - \mathbf{y}(\tau)|} = \delta(R(\tau)\cdot R(\tau))\,\Theta(R^0(\tau)) \tag{6.52}$$

where we're introduced the separation 4-vector

$$R^\mu = X^\mu - Y^\mu(\tau)$$

The delta-function and step-function in (6.52) pick out a unique value of the proper time that contributes to the gauge potential at point X. We call this proper time $\tau_\star$. It is the retarded time lying along a null direction, $R(\tau_\star)\cdot R(\tau_\star) = 0$. This should be thought of as the proper time version of our previous formula (6.36).

The form (6.52) allows us to do the integral over τ. But we still pick up a Jacobian-like factor from (6.47). This gives

$$\delta(R(\tau)\cdot R(\tau))\,\Theta(R^0(\tau)) = \frac{\delta(\tau - \tau_\star)}{2|R^\mu(\tau_\star)\dot{Y}^\mu(\tau_\star)|}$$

Putting all of this together gives our covariant form for the Liènard-Wierchert potential,

$$A^\mu(X) = \frac{\mu_0 qc}{4\pi}\,\frac{\dot{Y}^\mu(\tau_\star)}{|R^\nu(\tau_\star)\dot{Y}_\nu(\tau_\star)|}$$

This is our promised, compact expression. Expanding it out will give the previous results for the scalar (6.34) and vector (6.35) potentials. (To see this, you'll need to first show that $|R^\nu(\tau_\star)\dot{Y}^\nu(\tau_\star)| = c\gamma(\tau_\star)R(\tau_\star)(1 - \hat{\mathbf{R}}(\tau_\star)\cdot\mathbf{v}(\tau_\star)/c)$.)

The next step is to compute the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. This is what took us some time in Section 6.4.3. It turns out to be somewhat easier in the covariant approach. We need to remember that $\tau_\star$ is a function of X^μ. Then, we get

$$F_{\mu\nu} = \frac{\mu_0 qc}{4\pi}\left(\frac{\ddot{Y}_\nu(\tau_\star)}{|R^\rho(\tau_\star)\dot{Y}_\rho(\tau_\star)|}\frac{\partial\tau_\star}{\partial X^\mu} - \frac{\dot{Y}_\nu(\tau_\star)}{|R^\rho(\tau_\star)\dot{Y}_\rho(\tau_\star)|^2}\frac{\partial|R^\sigma(\tau_\star)\dot{Y}_\sigma(\tau_\star)|}{\partial X^\mu}\right) - (\mu \leftrightarrow \nu) \quad (6.53)$$

The simplest way to compute $\partial\tau_\star/\partial X^\mu$ is to start with $\eta_{\rho\sigma}R^\rho(\tau_\star)R^\sigma(\tau_\star) = 0$. Differentiating gives

$$\eta_{\rho\sigma}R^\rho(\tau_\star)\partial_\mu R^\sigma(\tau_\star) = \eta_{\rho\sigma}R^\rho(\tau_\star)\left(\delta^\sigma_\mu - \dot{Y}^\sigma(\tau_\star)\,\partial_\mu\tau_\star\right) = 0$$

Rearranging gives

$$\frac{\partial\tau_\star}{\partial X^\mu} = \frac{R_\mu(\tau_\star)}{R^\nu(\tau_\star)\dot{Y}_\nu(\tau_\star)}$$

For the other term, we have

$$\begin{aligned}\frac{\partial|R^\sigma(\tau_\star)\dot{Y}_\sigma(\tau_\star)|}{\partial X^\mu} &= \left(\delta^\sigma_\mu - \dot{Y}^\sigma(\tau_\star)\partial_\mu\tau_\star\right)\dot{Y}_\sigma(\tau_\star) + R^\sigma(\tau_\star)\ddot{Y}_\sigma(\tau_\star)\partial_\mu\tau_\star \\ &= \left(R^\sigma(\tau_\star)\ddot{Y}_\sigma(\tau_\star) + c^2\right)\partial_\mu\tau_\star + \dot{Y}_\mu(\tau_\star)\end{aligned}$$

where we've used $\dot{Y}^\mu\dot{Y}_\mu = c^2$. Using these in (6.53), we get our final expression for the field strength,

$$F_{\mu\nu}(X) = \frac{\mu_0 qc}{4\pi}\frac{1}{R^\rho\dot{Y}_\rho}\left[(-c^2 + R^\lambda\ddot{Y}_\lambda)\frac{R_\mu\dot{Y}_\nu - R_\nu\dot{Y}_\mu}{(R^\sigma\dot{Y}_\sigma)^2} + \frac{\ddot{Y}_\mu R_\nu - \ddot{Y}_\nu R_\mu}{R^\sigma\dot{Y}_\sigma}\right] \qquad (6.54)$$

This is the covariant field strength. It takes a little work to write this in terms of the component **E** and **B** fields but the final answer is, of course, given by (6.42) and (6.44) that we derived previously. Indeed, you can see the general structure in (6.54). The first term is proportional to velocity and goes as $1/R^2$; the second term is proportional to acceleration and goes as $1/R$.

6.4.5 Bremsstrahlung, Cyclotron and Synchrotron Radiation

To end our discussion, we derive the radiation due to some simple relativistic motion.

Power Radiated Again: Relativistic Larmor Formula

In Section 6.2.2, we derived the Larmor formula for the emitted power in the electric dipole approximation to radiation. In this section, we present the full, relativistic version of this formula.

We'll work with the expressions for the radiation fields $\mathbf{E}$ (6.42) and $\mathbf{B}$ (6.44). As previously, we consider only the radiative part of the electric and magnetic fields which drops off as $1/R$. The Poynting vector is

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{1}{\mu_0 c}\mathbf{E}\times(\hat{\mathbf{R}}\times\mathbf{E}) = \frac{1}{\mu_0 c}|\mathbf{E}|^2\hat{\mathbf{R}}$$

where all of these expressions are to be computed at the retarded time. The second equality follows from the relation (6.45), while the final equality follows because the radiative part of the electric field (6.42) is perpendicular to $\hat{\mathbf{R}}$. Using the expression (6.42), we have

$$\mathbf{S} = \frac{q^2}{16\pi^2\epsilon_0 c^3}\frac{|\hat{\mathbf{R}}\times[(\hat{\mathbf{R}}-\mathbf{v}/c)\times\mathbf{a}]|^2}{\kappa^6 R^2}\hat{\mathbf{R}}$$

with $\kappa = 1 - \hat{\mathbf{R}}\cdot\mathbf{v}/c$.

Recall that everything in the formula above is evaluated at the retarded time t', defined by $t' + R(t')/c = t$. This means, that the coordinates are set up so that we can integrate $\mathbf{S}$ over a sphere of radius R that surrounds the particle at its retarded time. However, there is a subtlety in computing the emitted power, associated to the Doppler effect. The energy emitted per unit time t is not the same as the energy emitted per unit time t'. They differ by the factor $dt/dt' = \kappa$. The power emitted per unit time t', per solid angle $d\Omega$, is

$$\frac{d\mathcal{P}}{d\Omega} = \kappa R^2\,\mathbf{S}\cdot\hat{\mathbf{R}} = \frac{q^2}{16\pi^2\epsilon_0 c^3}\frac{|\hat{\mathbf{R}}\times[(\hat{\mathbf{R}}-\mathbf{v}/c)\times\mathbf{a}]|^2}{\kappa^5} \tag{6.55}$$

To compute the emitted power, we must integrate this expression over the sphere. This is somewhat tedious. The result is given by

$$\mathcal{P} = \frac{q^2}{6\pi\epsilon_0 c^3}\gamma^4\left(a^2 + \frac{\gamma^2}{c^2}(\mathbf{v}\cdot\mathbf{a})^2\right) \tag{6.56}$$

This is the relativistic version of the Larmor formula (6.18). (There is a factor of 2 difference when compared to (6.20) because the former equation was time averaged). We now apply this to some simple examples.

Bremsstrahlung

Suppose a particle is travelling in a straight line, with velocity $\mathbf{v}$ parallel to acceleration $\mathbf{a}$. The most common situation of this type occurs when a particle decelerates. In this case, the emitted radiation is called *bremsstrahlung*, German for "braking radiation".

We'll sit at some point $\mathbf{x}$, at which the radiation reaches us from the retarded point on the particle's trajectory $\mathbf{r}(t')$. As before, we define $\mathbf{R}(t') = \mathbf{x} - \mathbf{r}(t')$. We introduce the angle θ, defined by

$$\hat{\mathbf{R}} \cdot \mathbf{v} = v \cos\theta$$

Because the $\mathbf{v} \times \mathbf{a}$ term in (6.55) vanishes, the angular dependence of the radiation is rather simple in this case. It is given by

$$\frac{d\mathcal{P}}{d\Omega} = \frac{q^2 a^2}{16\pi\epsilon_0 c^3} \frac{\sin^2\theta}{(1-(v/c)\cos\theta)^5}$$

For $v \ll c$, the radiation is largest in the direction $\theta \approx \pi/2$, perpendicular to the direction of travel. But, at relativistic speeds, $v \to c$, the radiation is beamed in the forward direction in two lobes, one on either side of the particle's trajectory. The total power emitted is (6.56) which, in this case, simplifies to

$$\mathcal{P} = \frac{q^2\gamma^6 a^2}{6\pi\epsilon_0 c^3}$$

Cyclotron and Synchrotron Radiation

Suppose that the particle travels in a circle, with $\mathbf{v} \cdot \mathbf{a} = 0$. We'll pick axes so that $\mathbf{a}$ is aligned with the x-axis and $\mathbf{v}$ is aligned with the z-axis. Then we write

$$\hat{\mathbf{R}} = \sin\theta\cos\phi\hat{\mathbf{x}} + \sin\theta\sin\phi\hat{\mathbf{y}} + \cos\theta\hat{\mathbf{z}}$$

After a little algebra, we find that the angular dependence of the emitted radiation is

$$\frac{d\mathcal{P}}{d\Omega} = \frac{q^2 a^2}{16\pi\epsilon_0 c^3} \frac{1}{(1-(v/c)\cos\theta)^3} \left(1 - \frac{\sin^2\theta\cos^2\phi}{\gamma^2(1-(v/c)\cos\theta)^2}\right)$$

At non-relativistic speeds, $v \ll c$, the angular dependence takes the somewhat simpler form $(1-\sin^2\theta\cos^2\phi)$. In this limit, the radiation is referred to as *cyclotron radiation*. In contrast, in the relativistic limit $v \to c$, the radiation is again beamed mostly in the forwards direction. This limit is referred to as *synchrotron radiation*. The total emitted power (6.56) is this time given by

$$\mathcal{P} = \frac{q^2\gamma^4 a^2}{6\pi\epsilon_0 c^3}$$

Note that the factors of γ differ from the case of linear acceleration.

7. Electromagnetism in Matter

Until now, we've focussed exclusively on electric and magnetic fields in vacuum. We end this course by describing the behaviour of electric and magnetic fields inside materials, whether solids, liquids or gases.

The materials that we would like to discuss are insulators which, in this context, are usually called *dielectrics*. These materials are the opposite of conductors: they don't have any charges that are free to move around. Moreover, they are typically neutral so that – at least when averaged – the charge density vanishes: $\rho = 0$. You might think that such neutral materials can't have too much effect on electric and magnetic fields. But, as we will see, things are more subtle and interesting.

7.1 Electric Fields in Matter

The fate of electric fields inside a dielectric depends on the microscopic make-up of the material. We going to work only with the simplest models. We'll consider our material to be constructed from a lattice of neutral atoms. Each of these atoms consists of a positively charged nuclei, surrounded by a negatively charged cloud of electrons. A cartoon of this is shown in the figure; the nucleus is drawn in red, the cloud of electrons in yellow.

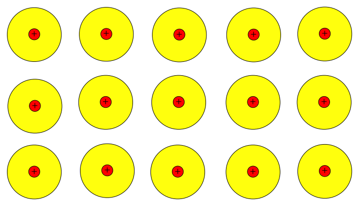


Figure 59: A simple model of a neutral material

Suppose that electric field $\mathbf{E}$ is applied to this material. What happens? Although each atom is neutral, its individual parts are not. This results in an effect called *polarisation*: the positively charged nucleus gets pushed a little in the direction of $\mathbf{E}$; the negatively charged cloud gets pushed a little in the opposite direction. (This is not to be confused with the orientation of the electromagnetic wave which also has the name "polarisation").

The net effect is that the neutral atom gains an electric dipole moment. Recall from Section 2 that two equal and opposite charges, $+q$ and $-q$, separated by a distance $\mathbf{d}$, have an electric dipole $\mathbf{p} = q\mathbf{d}$. By convention, $\mathbf{p}$ points from the negative charge to the positive charge.

It turns out that in most materials, the induced electric dipole is proportional to the electric field,

$$\mathbf{p} = \alpha\mathbf{E} \tag{7.1}$$

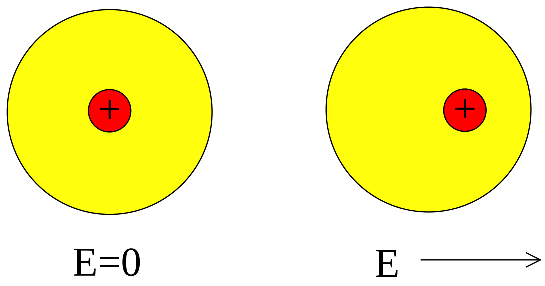


Figure 60: The polarisation of an atom

The proportionality factor α is called the *atomic polarisability*. Because $\mathbf{p}$ points from negative to positive charge, it points in the same direction as $\mathbf{E}$. The electric field will also result in higher multipole moments of the atoms. (For example, the cloud of electrons will be distorted). We will ignore these effects.

A Simple Model for Atomic Polarisability

Here's a simple model which illustrates how the relationship (7.1) arises. It also gives a ball-park figure for the value of the atomic polarisability α. Consider a nucleus of charge $+q$, surrounded by a spherical cloud of electrons of radius a. We'll take this cloud to have uniform charge density. If we just focus on the electron cloud for now, the electric field it produces was computed in Section 2: it rises linearly inside the cloud, before dropping off as $1/r^2$ outside the cloud. Here we're interested in the linearly increasing behaviour inside

$$\mathbf{E}_{\text{cloud}} = \frac{1}{4\pi\epsilon_0}\frac{qr}{a^3}\,\hat{\mathbf{r}} \qquad\qquad (r < a) \tag{7.2}$$

In the absence of an external field, the nucleus feels the field due to the cloud and sits at $r = 0$. Now apply an external electric field $\mathbf{E}$. The nucleus will be displaced to sit at a point where $\mathbf{E} + \mathbf{E}_{\text{cloud}} = 0$. In other words, it will be displaced by

$$\mathbf{r} = \frac{4\pi\epsilon_0 a^3}{q}\mathbf{E} \quad \Rightarrow \quad \mathbf{p} = q\mathbf{r} = 4\pi\epsilon_0 a^3\,\mathbf{E}$$

This gives the simple expression $\alpha = 4\pi\epsilon_0 a^3$. This isn't too far off the experimentally measured values. For example, for hydrogen $\alpha/4\pi\epsilon_0 \approx 0.7 \times 10^{-30}\ m^3$ which, from the above formula, suggests that the size of the cloud is around $a \sim 10^{-10}\ m$.

7.1.1 Polarisation

We've learnt that applying an electric field to a material causes each atom to pick up a dipole moment. We say that the material is *polarised*. The *polarisation* $\mathbf{P}$ is defined to be the average dipole moment per unit volume. If n is the density of atoms, each with dipole moment $\mathbf{p}$, then we can write

$$\mathbf{P} = n\mathbf{p} \tag{7.3}$$

We've actually dodged a bullet in writing this simple equation and evaded a subtle, but important, point. Let me try to explain. Viewed as a function of spatial position, the dipole moment $\mathbf{p}(\mathbf{r})$ is ridiculously complicated, varying wildly on distances comparable to the atomic scale. We really couldn't care less about any of this. We just want the average dipole moment, and that's what the equation above captures. But we do care if the average dipole moment varies over large, macroscopic distances. For example, the density n may be larger in some parts of the solid than others. And, as we'll see, this is going to give important, physical effects. This means that we don't want to take the average of $\mathbf{p}(\mathbf{r})$ over the whole solid since this would wash out all variations. Instead, we just want to average over small distances, blurring out any atomic messiness, but still allowing $\mathbf{P}$ to depend on $\mathbf{r}$ over large scales. The equation $\mathbf{P} = n\mathbf{p}$ is supposed to be shorthand for all of this. Needless to say, we could do a better job of defining $\mathbf{P}$ if forced to, but it won't be necessary in what follows.

The polarisation of neutral atoms is not the only way that materials can become polarised. One simple example is water. Each H_2O molecule already carries a dipole moment. (The oxygen atom carries a net negative charge, with each hydrogen carrying a positive charge). However, usually these molecules are jumbled up in water, each pointing in a different direction so that the dipole moments cancel out and the polarisation is $\mathbf{P} = 0$. This changes if we apply an electric field. Now the dipoles all want to align with the electric field, again leading to a polarisation.

In general, the polarisation $\mathbf{P}$ can be a complicated function of the electric field $\mathbf{E}$. However, most materials it turns out that $\mathbf{P}$ is proportional to $\mathbf{E}$. Such materials are called *linear dielectrics*. They have

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E} \tag{7.4}$$

where χ_e is called the *electric susceptibility*. It is always positive: $\chi_e > 0$. Our simple minded computation of atomic polarisability above gave such a linear relationship, with $\epsilon_0 \chi_e = n\alpha$.

The reason why most materials are linear dielectrics follows from some simple dimensional analysis. Any function that has $\mathbf{P}(\mathbf{E} = 0) = 0$ can be Taylor expanded as a linear term + quadratic + cubic and so on. For suitably small electric fields, the linear term always dominates. But how small is small? To determine when the quadratic and higher order terms become important, we need to know the relevant scale in the problem. For us, this is the scale of electric fields inside the atom. But these are huge. In most situations, the applied electric field leading to the polarisation is a tiny perturbation and the linear term dominates. Nonetheless, from this discussion it should be clear that we do expect the linearity to fail for suitably high electric fields.

There are other exceptions to linear dielectrics. Perhaps the most striking exception are materials for which $\mathbf{P} \neq 0$ even in the absence of an electric field. Such materials – which are not particularly common – are called *ferroelectric*. For what it's worth, an example is $BaTiO_3$.

Bound Charge

Whatever the cause, when a material is polarised there will be regions in which there is a build up of electric charge. This is called *bound charge* to emphasise the fact that it's not allowed to move and is arising from polarisation effects. Let's illustrate this with a simple example before we describe the general case. Let's go back to our lattice of neutral atoms. As we've seen, in the presence of an electric field they become polarised, as shown in the figure. However, as long as the polarisation is uniform, so $\mathbf{P}$ is constant, there is no net charge in the middle of the material: averaged over many atoms, the total charge remains the same. The only place that there is a net build up of charge is on the surface. In contrast, if $\mathbf{P}(\mathbf{r})$ is not constant, there will also be regions in the middle that have excess electric charge.

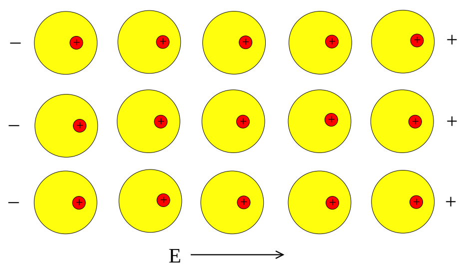


Figure 61: A polarised material

To describe this, recall that the electric potential due to each dipole $\mathbf{p}$ is

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \mathbf{r}}{r^3}$$

(We computed this in Section 2). Integrating over all these dipoles, we can write the potential in terms of the polarisation,

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int_V d^3r' \, \frac{\mathbf{P}(\mathbf{r}') \cdot (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}$$

We then have the following manipulations.,

$$\begin{aligned}\phi(\mathbf{r}) &= \frac{1}{4\pi\epsilon_0}\int_V d^3r'\;\mathbf{P}(\mathbf{r}')\cdot\nabla'\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)\\ &= \frac{1}{4\pi\epsilon_0}\int_S d\mathbf{S}\cdot\frac{\mathbf{P}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{4\pi\epsilon_0}\int_V d^3r'\;\frac{\nabla'\cdot\mathbf{P}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\end{aligned}$$

where S is the boundary of V. But both of these terms have a very natural interpretation. The first is the kind of potential that we would get from a surface charge,

$$\sigma_{\text{bound}} = \mathbf{P}\cdot\hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is the normal to the surface S. The second term is the kind of potential that we would get from a charge density of the form,

$$\rho_{\text{bound}}(\mathbf{r}) = -\nabla\cdot\mathbf{P}(\mathbf{r}) \tag{7.5}$$

This matches our intuition above. If the polarisation $\mathbf{P}$ is constant then we only find a surface charge. But if $\mathbf{P}$ varies throughout the material then this can lead to non-vanishing charge density sitting inside the material.

7.1.2 Electric Displacement

We learned in our first course that the electric field obeys Gauss' law

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0}$$

This is a fundamental law of Nature. It doesn't change just because we're inside a material. But, from our discussion above, we see that there's a natural way to separate the electric charge into two different types. There is the bound charge ρ_{bound} that arises due to polarisation. And then there is anything else. This could be some electric impurities that are stuck in the dielectric, or it could be charge that is free to move because our insulator wasn't quite as good an insulator as we originally assumed. The only important thing is that this other charge does not arise due to polarisation. We call this extra charge *free charge*, ρ_{free}. Gauss' law reads

$$\begin{aligned}\nabla\cdot\mathbf{E} &= \frac{1}{\epsilon_0}(\rho_{\text{free}}+\rho_{\text{bound}})\\ &= \frac{1}{\epsilon_0}(\rho_{\text{free}}-\nabla\cdot\mathbf{P})\end{aligned}$$

We define the *electric displacement*,

$$\mathbf{D} = \epsilon_0\mathbf{E}+\mathbf{P} \tag{7.6}$$

This obeys

$$\nabla \cdot \mathbf{D} = \rho_{\text{free}} \tag{7.7}$$

That's quite nice. Gauss' law for the displacement involves only the free charge; any bound charge arising from polarisation has been absorbed into the definition of $\mathbf{D}$.

For linear dielectrics, the polarisation is given by (7.4) and the displacement is proportional to the electric field. We write

$$\mathbf{D} = \epsilon \mathbf{E}$$

where $\epsilon = \epsilon_0(1+\chi_e)$ is the called the *permittivity* of the material. We see that, for linear dielectrics, things are rather simple: all we have to do is replace ϵ_0 with ϵ everywhere. Because $\epsilon > \epsilon_0$, it means that the electric field will be decreased. We say that it is *screened* by the bound charge. The amount by which the electric field is reduced is given by the dimensionless *relative permittivity* or *dielectric constant*,

$$\epsilon_r = \frac{\epsilon}{\epsilon_0} = 1 + \chi_e$$

For gases, ϵ_r is very close to 1. (It differs at one part in 10^{-3} or less). For water, $\epsilon_r \approx 80$.

An Example: A Dielectric Sphere

As a simple example, consider a sphere of dielectric material of radius R. We'll place a charge Q at the centre. This gives rise to an electric field which polarises the sphere and creates bound charge. We want to understand the resulting electric field $\mathbf{E}$ and electric displacement $\mathbf{D}$.

The modified Gauss' law (7.7) allows us to easily compute $\mathbf{D}$ using the same kind of methods that we used in Section 2. We have

$$\mathbf{D} = \frac{Q}{4\pi r^2}\hat{\mathbf{r}} \qquad\qquad (r < R)$$

where the condition $r < R$ means that this holds inside the dielectric. The electric field is then given by

$$\mathbf{E} = \frac{Q}{4\pi \epsilon r^2}\hat{\mathbf{r}} = \frac{Q/\epsilon_r}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}} \qquad\qquad (r < R) \tag{7.8}$$

This is what we'd expect from a charge Q/ϵ_r placed at the origin. The interpretation of this is that there is the bound charge gathers at the origin, screening the original charge Q. This bound charge is shown as the yellow ring in the figure surrounding the original charge in red. The amount of bound charge is simply the difference

$$Q_{\text{bound}} = \frac{Q}{\epsilon_r} - Q = \frac{1-\epsilon_r}{\epsilon_r}Q = -\frac{\chi_e}{\epsilon_r}Q$$

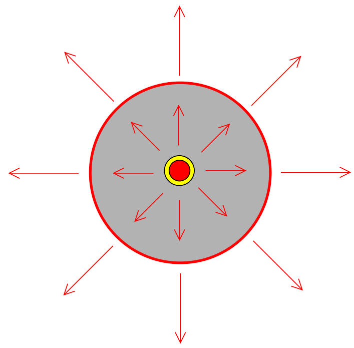


Figure 62: A polarised material

This bound charge came from the polarisation of the sphere. But the sphere is a neutral object which means that total charge on it has to be zero. To accomplish this, there must be an equal, but opposite, charge on the surface of the sphere. This is shown as the red rim in the figure. This surface charge is given by

$$4\pi R^2 \sigma_{\text{bound}} = -Q_{\text{bound}} = \frac{\epsilon_r - 1}{\epsilon_r}Q$$

We know from our first course that such a surface charge will lead to a discontinuity in the electric field. And that's exactly what happens. Inside the sphere, the electric field is given by (7.8). Meanwhile outside the sphere, Gauss' law knows nothing about the intricacies of polarisation and we get the usual electric field due to a charge Q,

$$\mathbf{E} = \frac{Q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}} \qquad\qquad (r > R)$$

At the surface $r = R$ there is a discontinuity,

$$\mathbf{E}\cdot\hat{\mathbf{r}}|_+ - \mathbf{E}\cdot\hat{\mathbf{r}}|_- = \frac{Q}{4\pi\epsilon_0 R^2} - \frac{Q}{4\pi\epsilon R^2} = \frac{\sigma_{\text{bound}}}{\epsilon_0}$$

which is precisely the expected discontinuity due to surface charge.

7.2 Magnetic Fields in Matter

Electric fields are created by charges; magnetic fields are created by currents. We learned in our first course that the simplest way to characterise any localised current distribution is through a *magnetic dipole moment* $\mathbf{m}$. For example, a current I moving in a planar loop of area A with normal $\hat{\mathbf{n}}$ has magnetic dipole moment,

$$\mathbf{m} = IA\hat{\mathbf{n}}$$

The resulting long-distance gauge field and magnetic field are

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\mathbf{r}}{r^3} \quad\Rightarrow\quad \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}}{r^3}\right)$$

The basic idea of this section is that current loops, and their associated dipole moments, already exist inside materials. They arise through two mechanisms:

- Electrons orbiting the nucleus carry angular momentum and act as magnetic dipole moments.
- Electrons carry an intrinsic spin. This is purely a quantum mechanical effect. This too contributes to the magnetic dipole moment.

In the last section, we defined the polarisation $\mathbf{P}$ to be the average dipole moment per unit volume. In analogy, we define the *magnetisation* $\mathbf{M}$ to be the average magnetic dipole moment per unit volume. Just as in the polarisation case, here "average" means averaging over atomic distances, but keeping any macroscopic variations of the polarisation $\mathbf{M}(\mathbf{r})$. It's annoyingly difficult to come up with simple yet concise notation for this. I'll choose to write,

$$\mathbf{M}(\mathbf{r}) = n\langle\mathbf{m}(\mathbf{r})\rangle$$

where n is the density of magnetic dipoles (which can, in principle, also depend on position) and the notation $\langle\cdot\rangle$ means averaging over atomic distance scales. In most (but not all) materials, if there is no applied magnetic field then the different atomic dipoles all point in random directions. This means that, after averaging, $\langle\mathbf{m}\rangle = 0$ when $\mathbf{B} = 0$. However, when a magnetic field is applied, the dipoles line up. The magnetisation typically takes the form $\mathbf{M} \propto \mathbf{B}$. We're going to use a slightly strange notation for the proportionality constant. (It's historical but, as we'll see, it turns out to simplify a later equation)

$$\mathbf{M} = \frac{1}{\mu_0}\frac{\chi_m}{1+\chi_m}\mathbf{B} \tag{7.9}$$

where χ_m is the *magnetic susceptibility*. The magnetic properties of materials fall into three different categories. The first two are dictated by the sign of χ_m:

- *Diamagnetism*: $-1 < \chi_m < 0$. The magnetisation of diamagnetic materials points in the opposite direction to the applied magnetic field. Most metals are diamagnetic, including copper and gold. Most non-metallic materials are also diamagnetic including, importantly, water with $\chi_m \approx -10^{-5}$. This means, famously, that frogs are also diamagnetic. Superconductors can be thought of as "perfect" diamagnets with $\chi_m = -1$.
- *Paramagnetism*: $\chi_m > 0$. In paramagnets, the magnetisation points in the same direction as the field. There are a number of paramagnetic metals, including Tungsten, Cesium and Aluminium.

We see that the situation is already richer than what we saw in the previous section. There, the polarisation takes the form $\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}$ with $\chi_e > 0$. In contrast, χ_m can have either sign. On top of this, there is another important class of material that don't obey (7.9). These are ferromagnets:

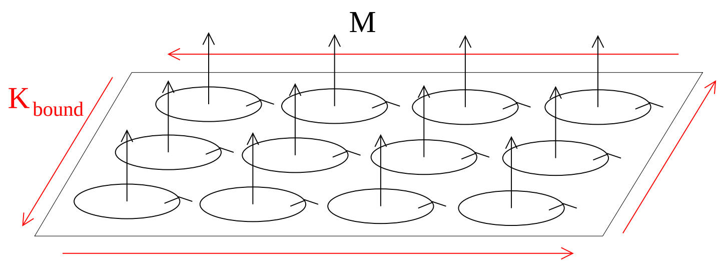
- *Ferromagnetism*: $\mathbf{M} \neq 0$ when $\mathbf{B} = 0$. Materials with this property are what you usually call "magnets". They're the things stuck to your fridge. The direction of $\mathbf{B}$ is from the south pole to the north. Only a few elements are ferromagnetic. The most familiar is Iron. Nickel and Cobalt are other examples.

In this course, we won't describe the microscopic effects that cause these different magnetic properties. They all involve quantum mechanics. (Indeed, the Bohr-van Leeuwen theorem says magnetism *can't* happen in a classical world — see the lecture notes on *Classical Dynamics*). A number of mechanisms for paramagetism and diamagnetism in metals are described in the lecture notes on *Statistical Physics*.

7.2.1 Bound Currents

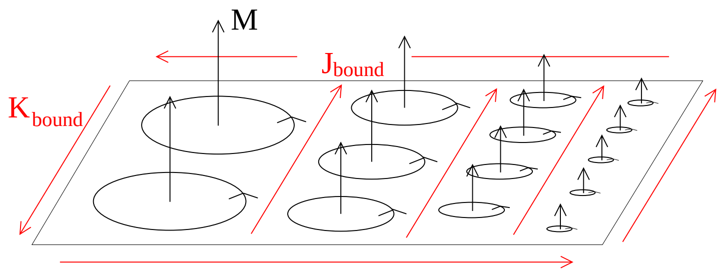
In the previous section, we saw that when a material is polarised, it results in bound charge. There is a similar story here. When a material becomes magnetised (at least in an anisotropic way), there will necessarily be regions in which there is a current. This is called the *bound current*.

Let's first give an intuitive picture for where these bound currents appear from. Consider a bunch of equal magnetic dipoles arranged uniformly on a plane like this:



The currents in the interior region cancel out and we're left only with a surface current around the edge. In Section 3, we denoted a surface current as $\mathbf{K}$. We'll follow this notation and call the surface current arising from a constant, internal magnetisation $\mathbf{K}_{\text{bound}}$.

Now consider instead a situation where the dipoles are arranged on a plane, but have different sizes. We'll put the big ones to the left and the small ones to the right, like this:



In this case, the currents in the interior no longer cancel. As we can see from the picture, they go into the page. Since $\mathbf{M}$ is out of the page, and we've arranged things so that $\mathbf{M}$ varies from left to right, this suggests that $\mathbf{J}_{\rm bound} \sim \nabla \times \mathbf{M}$.

Let's now put some equations on this intuition. We know that the gauge potential due to a magnetic dipole is

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

Integrating over all dipoles, and doing the same kinds of manipulations that we saw for the polarisations, we have

$$\begin{aligned}
\mathbf{A}(\mathbf{r}) &= \frac{\mu_0}{4\pi}\int_V d^3r'\;\frac{\mathbf{M}(\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \\
&= \frac{\mu_0}{4\pi}\int_V d^3r'\;\mathbf{M}(\mathbf{r}')\times\nabla'\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right) \\
&= -\frac{\mu_0}{4\pi}\int_S d\mathbf{S}'\times\frac{\mathbf{M}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\mu_0}{4\pi}\int_V d^3r'\frac{\nabla\times\mathbf{M}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}
\end{aligned}$$

Again, both of these terms have natural interpretation. The first can be thought of as due to a surface current

$$\mathbf{K}_{\rm bound} = \mathbf{M}\times\hat{\mathbf{n}}$$

where $\hat{n}$ is normal to the surface. The second term is the bound current in the bulk of the material. We can compare its form to the general expression for the Biot-Savart law that we derived in Section 3,

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int d^3r'\;\frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}$$

We see that the bound current is given by

$$\mathbf{J}_{\rm bound} = \nabla\times\mathbf{M} \tag{7.10}$$

as expected from our intuitive description above. Note that the bound current is a steady current, in the sense that it obeys $\nabla\cdot\mathbf{J}_{\rm bound} = 0$.

7.2.2 Ampère's Law Revisited

Recall that Ampère's law describes the magnetic field generated by static currents. We've now learned that, in a material, there can be two contributions to a current: the bound current $\mathbf{J}_{\rm bound}$ that we've discussed above, and the current $\mathbf{J}_{\rm free}$ from freely flowing electrons that we were implicitly talking. In Section 3, we were implicitly talking about $\mathbf{J}_{\rm free}$ when we discussed currents. Ampère's law does not distinguish between these two currents; the magnetic field receives contributions from both.

$$\begin{aligned}\nabla\times\mathbf{B} &= \mu_0(\mathbf{J}_{\rm free}+\mathbf{J}_{\rm bound})\\ &= \mu_0\mathbf{J}_{\rm free}+\mu_0\nabla\times\mathbf{M}\end{aligned}$$

We define the *magnetising field*, $\mathbf{H}$ as

$$\mathbf{H}=\frac{1}{\mu_0}\mathbf{B}-\mathbf{M} \tag{7.11}$$

This obeys

$$\nabla\times\mathbf{H}=\mathbf{J}_{\rm free} \tag{7.12}$$

We see that the field $\mathbf{H}$ plays a similar role to the electric displacement $\mathbf{D}$; the effect of the bound currents have been absorbed into $\mathbf{H}$, so that only the free currents contribute. Note, however, that we can't quite forget about $\mathbf{B}$ entirely, since it obeys $\nabla\cdot\mathbf{B}=0$. In contrast, we don't necessarily have "$\nabla\cdot\mathbf{H}=0$". Rather annoyingly, in a number of books $\mathbf{H}$ is called the magnetic field and $\mathbf{B}$ is called the magnetic induction. But this is stupid terminology so we won't use it.

For diamagnets or paramagnets, the magnetisation is linear in the applied magnetic field $\mathbf{B}$ and we can write

$$\mathbf{B}=\mu\mathbf{H}$$

A little algebra shows that $\mu=\mu_0(1+\chi_m)$. It is called the *permeability*. For most materials, μ differs from μ_0 only by 1 part in 10^5 or so. Finally, note that the somewhat strange definition (7.9) leaves us with the more sensible relationship between $\mathbf{M}$ and $\mathbf{H}$,

$$\mathbf{M}=\chi_m\mathbf{H}$$

7.3 Macroscopic Maxwell Equations

We've seen that the presence of bound charge and bound currents in matter can be absorbed into the definitions of $\mathbf{D}$ and $\mathbf{H}$. This allowed us to present versions of Gauss' law (7.7) and Ampère's law (7.12) which feature only the free charges and free currents. These equations hold for electrostatic and magnetostatic situations respectively. In this section we explain how to reformulate Maxwell's equations in matter in more general, time dependent, situations.

Famously, when fields depend on time there is an extra term required in Ampère's law. However, there is also an extra term in the expression (7.10) for the bound current. This arises because the bound charge, $\rho_{\rm bound}$, no longer sits still. It moves. But although it moves, it must still be locally conserved which means that it should satisfy a continuity equation

$$\nabla \cdot \mathbf{J}_{\rm bound} = -\frac{\partial \rho_{\rm bound}}{\partial t}$$

From our earlier analysis (7.5), we can express the bound charge in terms of the polarisation: $\rho_{\rm bound} = -\nabla \cdot \mathbf{P}$. Including both this contribution and the contribution (7.10) from the magnetisation, we have the more general expression for the bound current

$$\mathbf{J}_{\rm bound} = \nabla \times \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t}$$

Let's see how we can package the Maxwell equation using this notation. We're interested in the extension to Ampère's law which reads

$$\begin{aligned}\nabla \times \mathbf{B} - \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} &= \mu_0 \mathbf{J}_{\rm free} + \mu_0 \mathbf{J}_{\rm bound} \\ &= \mu_0 \mathbf{J}_{\rm free} + \mu_0 \nabla \times \mathbf{M} + \mu_0 \frac{\partial \mathbf{P}}{\partial t}\end{aligned}$$

As before, we can use the definition of $\mathbf{H}$ in (7.11) to absorb the magnetisation term. But we can also use the definition of $\mathbf{D}$ to absorb the polarisation term. We're left with the Maxwell equation

$$\nabla \times \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} = \mathbf{J}_{\mathbf{free}}$$

The Macroscopic Maxwell Equations

Let's gather together everything we've learned. Inside matter, the four Maxwell equations become

$$\begin{aligned}\nabla \cdot \mathbf{D} = \rho_{\rm free} \quad &\text{and} \quad \nabla \times \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} = \mathbf{J}_{\rm free} \\ \nabla \cdot \mathbf{B} = 0 \quad &\text{and} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\end{aligned} \tag{7.13}$$

There are the *macroscopic Maxwell equations*. Note that half of them are written in terms of the original $\mathbf{E}$ and $\mathbf{B}$ while the other half are written in terms of $\mathbf{D}$ and $\mathbf{H}$. Before we solve them, we need to know the relationships between these quantities. In the simplest, linear materials, this can be written as

$$\mathbf{D} = \epsilon \mathbf{E} \quad \text{and} \quad \mathbf{B} = \mu \mathbf{H}$$

Doesn't all this look simple! The atomic mess that accompanies most materials can simply be absorbed into two constants, the permittivity ϵ and the permeability μ. Be warned, however: things are not always as simple as they seem. In particular, we'll see in Section 7.5 that the permittivity ϵ is not as constant as we're pretending.

7.3.1 A First Look at Waves in Matter

We saw earlier how the Maxwell equations give rise to propagating waves, travelling with speed c. We call these waves "light". Much of our interest in this section will be on what becomes of these waves when we work with the macroscopic Maxwell equations. What happens when they bounce off different materials? What really happens when they propagate through materials?

Let's start by looking at the basics. In the absence of any free charge or currents, the macroscopic Maxwell equations (7.13) become

$$\begin{aligned} \nabla \cdot \mathbf{D} = 0 \quad &\text{and} \quad \nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} \\ \nabla \cdot \mathbf{B} = 0 \quad &\text{and} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \end{aligned} \tag{7.14}$$

which should be viewed together with the relationships $\mathbf{D} = \epsilon \mathbf{E}$ and $\mathbf{B} = \mu \mathbf{H}$. But these are of exactly the same form as the Maxwell equations in vacuum. Which means that, at first glance, the propagation of waves through a medium works just like in vacuum. All we have to do is replace $\epsilon_0 \to \epsilon$ and $\mu_0 \to \mu$. By the same sort of manipulations that we used in Section 4.3, we can derive the wave equations

$$\frac{1}{v^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} - \nabla^2 \mathbf{E} = 0 \quad \text{and} \quad \frac{1}{v^2}\frac{\partial^2 \mathbf{H}}{\partial t^2} - \nabla^2 \mathbf{H} = 0$$

The only difference from what we saw before is that the speed of propagation is now given by

$$v^2 = \frac{1}{\epsilon \mu}$$

This is less than the speed in vacuum: $v^2 \leq c^2$. It's common to define the *index of refraction*, n, as

$$n = \frac{c}{v} \geq 1 \tag{7.15}$$

In most materials, $\mu \approx \mu_0$. In this case, the index of refraction is given in terms of the dielectric constant as

$$n \approx \sqrt{\epsilon_r}$$

The monochromatic, plane wave solutions to the macroscopic wave equations take the familiar form

$$\mathbf{E} = \mathbf{E}_0\, e^{i(\mathbf{k}\cdot\mathbf{x}+\omega t)} \quad \text{and} \quad \mathbf{B} = \mathbf{B}_0\, e^{i(\mathbf{k}\cdot\mathbf{x}+\omega t)}$$

where the dispersion relation is now given by

$$\omega^2 = v^2 k^2$$

The polarisation vectors must obey $\mathbf{E}_0 \cdot \mathbf{k} = \mathbf{B}_0 \cdot \mathbf{k} = 0$ and

$$\mathbf{B}_0 = \frac{\hat{\mathbf{k}} \times \mathbf{E}_0}{v}$$

Boundary Conditions

In what follows, we're going to spend a lot of time bouncing waves off various surfaces. We'll typically consider an interface between two dielectric materials with different permittivities, ϵ_1 and ϵ_2. In this situation, we need to know how to patch together the fields on either side.

Let's first recall the boundary conditions that we derived in Sections 2 and 3. In the presence of surface charge, the electric field normal to the surface is discontinuous, while the electric field tangent to the surface is continuous. For magnetic fields, it's the other way around: in the presence of a surface current, the magnetic field normal to the surface is continuous while the magnetic field tangent to the surface is discontinuous.

What happens with dielectrics? Now we have two options of the electric field, $\mathbf{E}$ and $\mathbf{D}$, and two options for the magnetic field, $\mathbf{B}$ and $\mathbf{H}$. They can't both be continuous because they're related by $\mathbf{D} = \epsilon\mathbf{E}$ and $\mathbf{B} = \mu\mathbf{H}$ and we'll be interested in situation where ϵ (and possibly μ) are different on either side. Nonetheless, we can use the same kind of computations that we saw previously to derive the boundary conditions. Roughly, we get one boundary condition from each of the Maxwell equations.

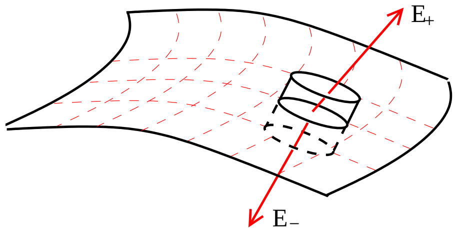


Figure 63: The normal component of the electric field is discontinuous

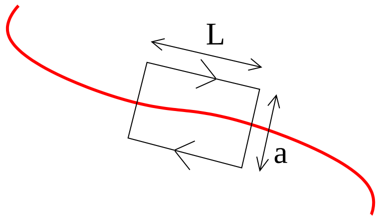


Figure 64: The tangential component of the electric field is continuous.

For example, consider the Gaussian pillbox shown in the left-hand figure above. Integrating the Maxwell equation $\nabla \cdot \mathbf{D} = \rho_{\rm free}$ tells us that the normal component of $\mathbf{D}$ is discontinuous in the presence of surface charge,

$$\hat{\mathbf{n}} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \sigma \tag{7.16}$$

where $\hat{\mathbf{n}}$ is the normal component pointing from 1 into 2. Here σ refers only to the free surface charge. It does not include any bound charges. Similarly, integrating $\nabla \cdot \mathbf{B} = 0$ over the same Gaussian pillbox tells us that the normal component of the magnetic field is continuous,

$$\hat{\mathbf{n}} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0 \tag{7.17}$$

To determine the tangential components, we integrate the appropriate field around the loop shown in the right-hand figure above. By Stoke's theorem, this is going to be equal to the integral of the curl of the field over the bounding surface. This tells us what the appropriate field is: it's whatever appears in the Maxwell equations with a curl. So if we integrate $\mathbf{E}$ around the loop, we get the result

$$\hat{\mathbf{n}} \times (\mathbf{E}_2 - \mathbf{E}_1) = 0 \tag{7.18}$$

Meanwhile, integrating $\mathbf{H}$ around the loop tells us the discontinuity condition for the magnetic field

$$\hat{\mathbf{n}} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{K} \tag{7.19}$$

where $\mathbf{K}$ is the surface current.

7.4 Reflection and Refraction

We're now going to shine light on something and watch how it bounces off. We did something very similar in Section 4.3, where the light reflected off a conductor. Here, we're going to shine the light from one dielectric material into another. These two materials will be characterised by the parameters ϵ_1, μ_1 and ϵ_2, μ_2. We'll place the interface at $x = 0$, with "region one" to the left and "region two" to the right.

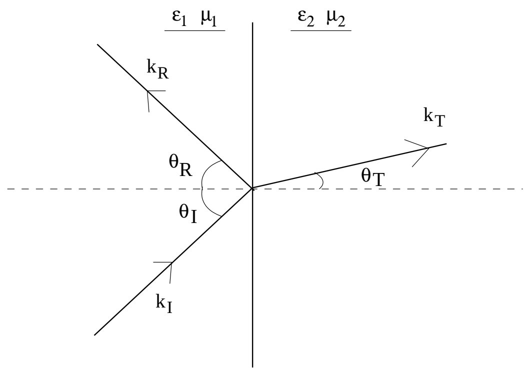


Figure 65: Incident, reflected and transmitted waves in a dielectric interface.

We send in an incident wave from region one towards the interface with a frequency ω_I and wavevector $\mathbf{k}_I$,

$$\mathbf{E}_{\rm inc} = \mathbf{E}_I \, e^{i(\mathbf{k}_I\cdot\mathbf{x}-\omega_I t)}$$

where

$$\mathbf{k}_I = k_I \cos\theta_I \,\hat{\mathbf{x}} + k_I \sin\theta_I \,\hat{\mathbf{z}}$$

When the wave hits the interface, two things can happen. It can be reflected, or it can pass through to the other region. In fact, in general, both of these things will happen. The reflected wave takes the general form,

$$\mathbf{E}_{\rm ref} = \mathbf{E}_R \, e^{i(\mathbf{k}_R\cdot\mathbf{x}-\omega_R t)}$$

where we've allowed for the possibility that the amplitude, frequency, wavevector and polarisation all may change. We will write the reflected wavevector as

$$\mathbf{k}_R = -k_R \cos\theta_R \,\hat{\mathbf{x}} + k_R \sin\theta_R \,\hat{\mathbf{z}}$$

Meanwhile, the part of the wave that passes through the interface and into the second region is the transmitted wave which takes the form,

$$\mathbf{E}_{\rm trans} = \mathbf{E}_T \, e^{i(\mathbf{k}_T\cdot\mathbf{x}-\omega_T t)}$$

with

$$\mathbf{k}_T = k_T \cos\theta_T \,\hat{\mathbf{x}} + k_T \sin\theta_T \,\hat{\mathbf{z}} \tag{7.20}$$

Again, we've allowed for the possibility that all the different properties of the wave could differ from the incoming wave. The electric field then takes the general form,

$$\mathbf{E} = \begin{cases} \mathbf{E}_{\text{inc}} + \mathbf{E}_{\text{ref}} & x < 0 \\ \mathbf{E}_{\text{trans}} & x > 0 \end{cases}$$

All of this is summarised in the figure.

We want to impose the matching conditions (7.16), (7.18), (7.19) and (7.17), with no surface charges and no surface currents. To start, we need the phase factors to be equal for all time. This means that we must have

$$\omega_I = \omega_R = \omega_T \tag{7.21}$$

and

$$\mathbf{k}_I \cdot \mathbf{x} = \mathbf{k}_R \cdot \mathbf{x} = \mathbf{k}_T \cdot \mathbf{x} \quad \text{at } x = 0 \tag{7.22}$$

This latter condition tells us that all of the wavevectors lie in the (x, z)-plane because $\mathbf{k}_I$ originally lay in this plane. It further imposes the equality of the $\hat{\mathbf{z}}$ components of the wavevectors:

$$k_I \sin\theta_I = k_R \sin\theta_R = k_T \sin\theta_T \tag{7.23}$$

But, in each region, the frequency and wavenumbers are related, through the dispersion relation, to the speed of the wave. In region 1, we have $\omega_I = v_1 k_I$ and $\omega_R = v_1 k_R$ which, using (7.21) and (7.23), tells us that

$$\theta_I = \theta_R$$

This is the familiar law of reflection.

Meanwhile, in region 2 we have $\omega_T = v_2 k_T$. Now (7.21) and (7.23) tell us that

$$\frac{\sin\theta_I}{v_1} = \frac{\sin\theta_T}{v_2}$$

In terms of the refractive index $n = c/v$, this reads

$$n_1 \sin\theta_I = n_2 \sin\theta_T \tag{7.24}$$

This is the law of refraction, known as *Snell's law*.

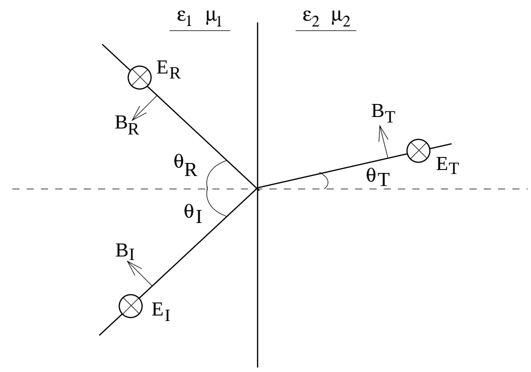


Figure 66: Incident, reflected and transmitted waves with normal polarisation.

7.4.1 Fresnel Equations

There's more information to be extracted from this calculation: we can look at the amplitudes of the reflected and transmitted waves. As we now show, this depends on the polarisation of the incident wave. There are two cases:

Normal Polarisation:

When the direction of $\mathbf{E}_I = E_I\hat{\mathbf{y}}$ is normal to the (x, z)-plane of incidence, it's simple to check that the electric polarisation of the other waves must lie in the same direction: $\mathbf{E}_R = E_T\hat{\mathbf{y}}$ and $\mathbf{E}_T = E_T\hat{\mathbf{y}}$. This situation, shown in Figure 66, is sometimes referred to as *s-polarised* (because the German word for normal begins with s).
The matching condition (7.18) requires

$$E_I + E_R = E_T$$

Meanwhile, as we saw in (7.16), the magnetic fields are given by $\mathbf{B} = (\hat{\mathbf{k}} \times \mathbf{E})/v$. The matching condition (7.19) then tells us that

$$B_I \cos\theta_I - B_R \cos\theta_R = B_T \cos\theta_T \quad \Rightarrow \quad \frac{E_I - E_R}{v_1}\cos\theta_I = \frac{E_T}{v_2}\cos\theta_T$$

With a little algebra, we can massage these conditions into the expressions,

$$\frac{E_R}{E_I} = \frac{n_1\cos\theta_I - n_2\cos\theta_T}{n_1\cos\theta_I + n_2\cos\theta_T} \quad \text{and} \quad \frac{E_T}{E_I} = \frac{2n_1\cos\theta_I}{n_1\cos\theta_I + n_2\cos\theta_T} \tag{7.25}$$

These are the *Fresnel equations* for normal polarised light. We can then use Snell's law (7.24) to get the amplitudes in terms of the refractive indices and the incident angle θ_I.

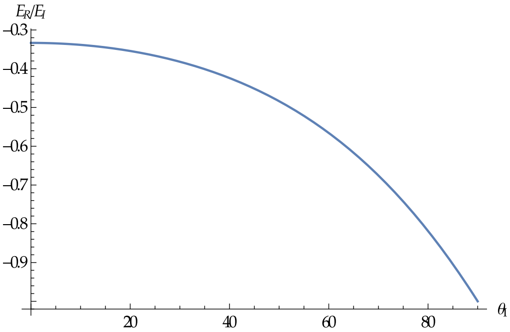


Figure 67: The reflected field with normal polarisation

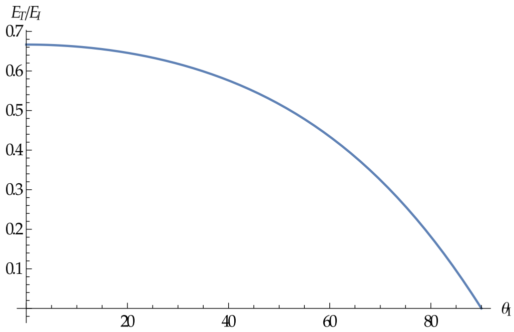


Figure 68: The transmitted field with normal polarisation

The most common example is if region 1 contains only air, with $n_1 \approx 1$, and region 2 consists of some transparent material. (For example, glass which has $n_2 \approx 1.5$). The normalised reflected and transmitted fields are plotted in the figures below for $n_1 = 1$ and $n_2 = 2$, with θ_I plotted in degrees along the horizontal axis).
Note that the horizontal axis are different; negative for the reflected wave, positive for the transmitted wave. In particular, when $\theta = 90°$, the whole wave is reflected and nothing is transmitted.

Parallel Polarisation:

The case in which the electric field lies lies within the (x, z)-plane of incidence is sometimes referred to as *p-polarised* (because the English word for parallel begins with p). It is shown in Figure 69. Of course, we still require $\mathbf{E}_I \cdot \mathbf{k} = 0$, which means that

$$\mathbf{E}_I = -E_I \sin\theta_I\,\hat{\mathbf{x}} + E_I \cos\theta_I\,\hat{\mathbf{z}}$$

with similar expressions for $\mathbf{E}_R$ and $\mathbf{E}_T$. The magnetic field now lies in the $\pm\hat{\mathbf{y}}$ direction. The matching condition (7.18) equates the components of the electric field tangential to the surface. This means

$$E_I \cos\theta_I + E_R \cos\theta_R = E_T \cos\theta_T$$

while the matching condition (7.19) for the components of magnetic field tangent to the surface gives

$$B_I - B_R = B_T \quad\quad \Rightarrow \quad\quad \frac{E_I - E_R}{v_1} = \frac{E_T}{v_2}$$

where the minus sign for B_R can be traced to the fact that the direction of the $\mathbf{B}$ field (relative to $\mathbf{k}$) points in the opposite direction after a reflection. These two conditions

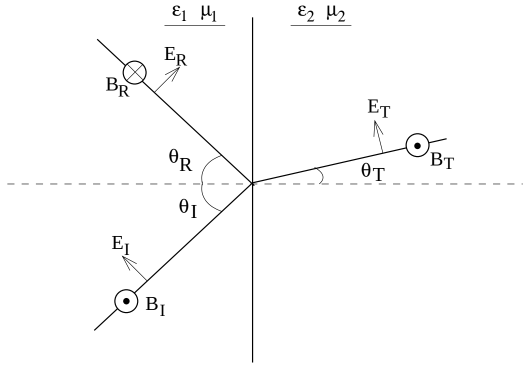


Figure 69: Incident, reflected and transmitted waves with parallel polarisation.

can be written as

$$\frac{E_R}{E_I} = \frac{n_1 \cos\theta_T - n_2 \cos\theta_I}{n_1 \cos\theta_T + n_2 \cos\theta_I} \quad \text{and} \quad \frac{E_T}{E_I} = \frac{2n_1 \cos\theta_I}{n_1 \cos\theta_T + n_2 \cos\theta_I} \tag{7.26}$$

These are the *Fresnel equations* for parallel polarised light. Note that when the incident wave is normal to the surface, so both $\theta_I = \theta_T = 0$, the amplitudes for the normal (7.25) and parallel (7.26) polarisations coincide. But in general, they are different.

We can again plot the reflected and transmitted amplitudes in the case $n_1 = 1$ and $n_2 = 2$, shown in the figure.

Brewster's Angle

We can see from the left-hand figure that something interesting happens in the case of parallel polarisation. There is an angle for which there is no reflected wave. Everything gets transmitted. This is called the *Brewster Angle*, θ_B. It occurs when $n_1 \cos\theta_T = n_2 \cos\theta_I$. Of course, we also need to obey Snell's law (7.24). These two conditions are only satisfied when $\theta_I + \theta_T = \pi/2$. The Brewster angle is given by

$$\tan\theta_B = \frac{n_2}{n_1}$$

For the transmission of waves from air to glass, $\theta_B \approx 56°$.

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.

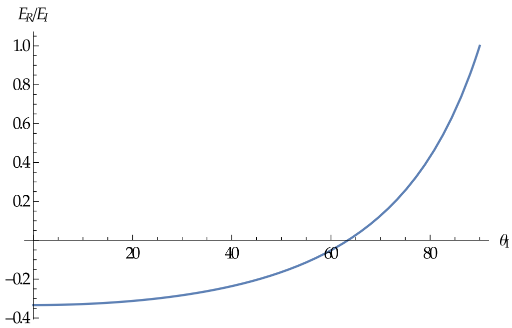


Figure 70: The reflected field with parallel polarisation

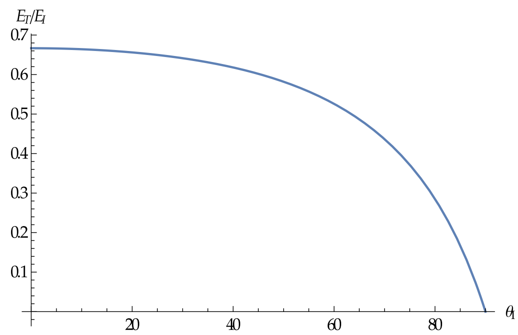


Figure 71: The transmitted field with parallel polarisation

7.4.2 Total Internal Reflection

Let's return to Snell's law (7.24) that tells us the angle of refraction,

$$\sin\theta_T = \frac{n_1}{n_2}\sin\theta_I$$

But there's a problem with this equation: if $n_2 > n_1$ then the right-hand side can be greater that one, in which case there are no solutions. This happens at the critical angle of incidence, θ_C, defined by

$$\sin\theta_C = \frac{n_2}{n_1}$$

For example, if light is moving from glass, into air, then $\theta_C \approx 42^\circ$. At this angle, and beyond, there is no transmitted wave. Everything is reflected. This is called *total internal reflection*. It's what makes diamonds sparkle and makes optical fibres to work.

Here our interest is not in jewellery, but rather in a theoretical puzzle about how total internal reflection can be consistent. After all, we've computed the amplitude of the transmitted electric field in (7.25) and (7.26) and it's simple to check that it doesn't vanish when $\theta_I = \theta_C$. What's going on?

The answer lies back in our expression for the transmitted wavevector $\mathbf{k}_T$ which we decomposed in (7.20) using geometry. The matching condition (7.22) tells us that $\mathbf{k}_T\cdot\hat{\mathbf{y}} = 0$ and

$$\mathbf{k}_T\cdot\hat{\mathbf{z}} = \mathbf{k}_I\cdot\hat{\mathbf{z}} = \frac{\omega_I}{v_1}\sin\theta_I$$

But, from the matching of frequencies (7.21), we know that $\omega_I = \omega_T \equiv \omega$. We also know that the magnitude of the transmitted wavevector is given by $|\mathbf{k}_T|^2 = \omega^2/v_2^2$. But this means that the component of the wavevector in the $\hat{\mathbf{x}}$ direction of propagation must be

$$\mathbf{k}_T \cdot \hat{\mathbf{x}} = \pm\sqrt{|\mathbf{k}_T|^2 - (\mathbf{k}_T \cdot \hat{\mathbf{z}})^2} = \pm\frac{\omega}{v_2}\sqrt{1 - \frac{v_2^2 \sin^2\theta_I}{v_1^2}} = \pm\frac{\omega}{v_2}\sqrt{1 - \frac{n_1^2 \sin^2\theta_I}{n_2^2}}$$

We see that when $n_1 \sin\theta_I / n_2 > 1$, the $\hat{\mathbf{x}}$ component of the wavevector is imaginary! We'll write $\mathbf{k}_T \cdot \hat{\mathbf{x}} = \pm i\omega\alpha/v_2$. An imaginary wavevector sounds strange, but it's very simple to interpret: we simply substitute it into our wave solution to find

$$\mathbf{E}_{\text{trans}} = \mathbf{E}_T\, e^{(i\mathbf{k}_T\cdot\hat{\mathbf{z}} - \omega t)}\, e^{\mp\omega\alpha x/v_2} \qquad x > 0$$

Picking the minus sign in the exponent gives the physically sensible solution which decays as we move into region 2. We see that beyond the critical angle θ_C, there is no propagating wave in region 2. Instead it is replaced by a decaying solution. This is called an *evanescent wave*.

As we'll now see, the idea that the wavevector can be imaginary is very useful in a many other circumstances.

7.5 Dispersion

The dielectric constant $\epsilon_r = \epsilon/\epsilon_0$ is poorly named. It is not constant. This is because, in the presence of time-dependent electric fields, the permittivity typically depends on the frequency: $\epsilon = \epsilon(\omega)$. In this section, we will first provide a simple model to understand why this is the case and what form of $\epsilon(\omega)$ we should expect. We'll then move on to see the consequences of this frequency dependence.

7.5.1 Atomic Polarisability Revisited

In Section 7.1, we introduced a simple model for electric polarisability. This treats the atom as a point-like nucleus with charge q, surrounded by a cloud of electrons which we treat as a solid ball of radius a with uniform charge density. It's obviously a daft model for the atom, but it will do for our purposes.

Suppose that the centre of the electron cloud is displaced by a distance r. (You can equivalently think of the nucleus as displaced by the same distance in the opposite direction). We previously computed the restoring force (7.2) which acts on cloud,

$$\mathbf{F}_{\text{cloud}} = -\frac{q^2}{4\pi\epsilon_0 a^3}\mathbf{r} = -m\omega_0^2\mathbf{r}$$

In the final equality, we've introduced the mass m of the cloud and defined the quantity ω_0 which we will call the *resonant frequency*.

In Section 7.1, we just looked at the equilibrium configuration of the electron cloud. Here, instead, we want to subject the atom to a time-dependent electric field $\mathbf{E}(t)$. In this situation, the electron cloud also feels a damping force

$$\mathbf{F}_{\text{damping}} = -m\gamma\dot{\mathbf{r}} \tag{7.27}$$

for some constant coefficient γ. You might find it strange to see such a friction term occurring for an atomic system. After all, we usually learn that friction is the effect of averaging over many many atoms. The purpose of this term is to capture the fact that the atom can lose energy, either to surrounding atoms or emitted electromagnetic radiation (which we'll learn more about in Section 6). If we now apply a time dependent electric field $\mathbf{E}(t)$ to this atom, the equation of motion for the displacement it

$$m\ddot{\mathbf{r}} = -q\mathbf{E}(t) - m\omega_0^2\mathbf{r} + m\gamma\dot{\mathbf{r}} \tag{7.28}$$

Solutions to this describe the atomic cloud oscillating about the nucleus.

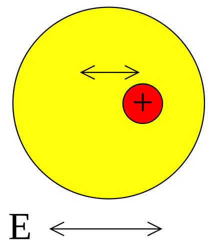


Figure 72:

The time dependent electric field will be of the wave form that we've seen throughout these lectures: $\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$. However, the atom is tiny. In particular, it is small compared to the wavelength of (at least) visible light, meaning $ka \ll 1$. For this reason, we can ignore the fact that the phase oscillates in space and work with an electric field of the form $\mathbf{E}(t) = \mathbf{E}_0 e^{-i\omega t}$. Then (7.28) is the equation for a forced, damped harmonic oscillator. We search for solutions to (7.28) of the form $\mathbf{r}(t) = \mathbf{r}_0 e^{-i\omega t}$. (In the end we will take the real part). The solution is

$$\mathbf{r}_0 = -\frac{q\mathbf{E}_0}{m}\frac{1}{-\omega^2 + \omega_0^2 - i\gamma\omega}$$

This gives the atomic polarisability $\mathbf{p} = \alpha\mathbf{E}$, where

$$\alpha = \frac{q^2/m}{-\omega^2 + \omega_0^2 - i\gamma\omega} \tag{7.29}$$

As promised, the polarisability depends on the frequency. Moreover, it is also complex. This has the effect that the polarisation of the atom is not in phase with the oscillating electric field.

Because the polarisability is both frequency dependent and complex, the permittivity $\epsilon(\omega)$ will also be both frequency dependent and complex. (In the simplest settings, they are related by $\epsilon(\omega) = \epsilon_0 + n\alpha(\omega)$ where n is the density of atoms). We'll now see the effect this has on the propagation of electromagnetic waves through materials.

7.5.2 Electromagnetic Waves Revisited

To start, we'll consider a general form of the permittivity $\epsilon(\omega)$ which is both frequency dependent and complex; we'll return to the specific form arising from the polarisability (7.29) later. In contrast, we will assume that the magnetic thing μ is both constant and real, which turns out to be a good approximation for most materials. This means that we have

$$\mathbf{D} = \epsilon(\omega)\mathbf{E} \quad \text{and} \quad \mathbf{B} = \mu\mathbf{H}$$

We'll look for plane wave solutions, so that the electric and magnetic fields takes the form

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}(\omega)\, e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \quad \text{and} \quad \mathbf{B}(\mathbf{x}, t) = \mathbf{B}(\omega)\, e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$$

Maxwell's equations in matter were given in (7.14). The first two simply tell us

$$\begin{aligned} \nabla\cdot\mathbf{D} = 0 \qquad &\Rightarrow \qquad \epsilon(\omega)\mathbf{k}\cdot\mathbf{E}(\omega) = 0 \\ \nabla\cdot\mathbf{B} = 0 \qquad &\Rightarrow \qquad \mathbf{k}\cdot\mathbf{B}(\omega) = 0 \end{aligned}$$

These are the statements that the electric and magnetic fields remain transverse to the direction of propagation. (In fact there's a caveat here: if $\epsilon(\omega) = 0$ for some frequency ω, then the electric field need not be transverse. This won't affect our discussion here, but we will see an example of this when we turn to conductors in Section 7.6). Meanwhile, the other two equations are

$$\begin{aligned} \nabla\times\mathbf{H} = \frac{\partial\mathbf{D}}{\partial t} \qquad &\Rightarrow \qquad \mathbf{k}\times\mathbf{B}(\omega) = -\mu\epsilon(\omega)\omega\mathbf{E}(\omega) \\ \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \qquad &\Rightarrow \qquad \mathbf{k}\times\mathbf{E}(\omega) = \omega\mathbf{B}(\omega) \end{aligned} \tag{7.30}$$

We do the same manipulation that we've seen before: look at $\mathbf{k}\times(\mathbf{k}\times\mathbf{E})$ and use the fact that $\mathbf{k}\cdot\mathbf{E} = 0$. This gives us the *dispersion* relation

$$\mathbf{k}\cdot\mathbf{k} = \mu\epsilon(\omega)\omega^2 \tag{7.31}$$

We need to understand what this equation is telling us. In particular, $\epsilon(\omega)$ is typically complex. This, in turn, means that the wavevector $\mathbf{k}$ will also be complex. To be specific, we'll look at waves propagating in the z-direction and write $\mathbf{k} = k\hat{\mathbf{z}}$. We'll write the real and imaginary parts as

$$\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega) \quad \text{and} \quad k = k_1 + ik_2$$

Then the dispersion relation reads

$$k_1 + ik_2 = \omega\sqrt{\mu}\sqrt{\epsilon_1 + i\epsilon_2} \tag{7.32}$$

and the electric field takes the form

$$\mathbf{E}(\mathbf{x},t) = \mathbf{E}(\omega)\, e^{-k_2 z}\, e^{i(k_1 z - \omega t)} \tag{7.33}$$

We now see the consequence of the imaginary part of $\epsilon(\omega)$; it causes the amplitude of the wave to decay as it extends in the z-direction. This is also called *attenuation*. The real part, k_1, determines the oscillating part of the wave. The fact that ϵ depends on ω means that waves of different frequencies travel with different speeds. We'll discuss shortly the ways of characterising these speeds.

The magnetic field is

$$\mathbf{B}(\omega) = \frac{k}{\omega}\,\hat{\mathbf{z}} \times \mathbf{E}(\omega) = \frac{|k|e^{i\phi}}{\omega}\,\hat{\mathbf{z}} \times \mathbf{E}(\omega)$$

where $\phi = \tan^{-1}(k_2/k_1)$ is the phase of the complex wavenumber k. This is the second consequence of a complex permittivity $\epsilon(\omega)$; it results in the electric and magnetic fields oscillating out of phase. The profile of the magnetic field is

$$\mathbf{B}(\mathbf{x},t) = \frac{|k|}{\omega}(\hat{\mathbf{z}} \times \mathbf{E}(\omega))\, e^{-k_2 z}\, e^{i(k_1 z - \omega t + \phi)} \tag{7.34}$$

As always, the physical fields are simply the real parts of (7.33) and (7.34), namely

$$\begin{aligned}
\mathbf{E}(\mathbf{x},t) &= \mathbf{E}(\omega)\, e^{-k_2 z}\, \cos(k_1 z - \omega t) \\
\mathbf{B}(\mathbf{x},t) &= \frac{|k|}{\omega}(\hat{\mathbf{z}} \times \mathbf{E}(\omega))\, e^{-k_2 z}\, \cos(k_1 z - \omega t + \phi)
\end{aligned}$$

To recap: the imaginary part of ϵ means that $k_2 \neq 0$. This has two effects: it leads to the damping of the fields, and to the phase shift between $\mathbf{E}$ and $\mathbf{B}$.

Measures of Velocity

The other new feature of $\epsilon(\omega)$ is that it depends on the frequency ω. The dispersion relation (7.31) then immediately tells us that waves of different frequencies travel at different speeds. There are two, useful characterisations of these speeds. The *phase velocity* is defined as

$$v_p = \frac{\omega}{k_1}$$

As we can see from (7.33) and (7.34), a wave of a fixed frequency ω propagates with phase velocity $v_p(\omega)$.

Waves of different frequency will travel with different phase velocities v_p. This means that for wave pulses, which consist of many different frequencies, different parts of the wave will travel with different speeds. This will typically result in a change of shape of the pulse as it moves along. We'd like to find a way to characterise the speed of the whole pulse. The usual measure is the *group velocity*, defined as

$$v_g = \frac{d\omega}{dk_1}$$

where we've inverted (7.31) so that we're now viewing frequency as a function of (real) wavenumber: $\omega(k_1)$.

To see why the group velocity is a good measure of the speed, let's build a pulse by superposing lots of waves of different frequencies. To make life simple, we'll briefly set $\epsilon_2 = 0$ and $k_1 = k$ for now so that we don't have to think about damping effects. Then, focussing on the electric field, we can build a pulse by writing

$$\mathbf{E}(\mathbf{x}, t) = \int \frac{dk}{2\pi}\, \mathbf{E}(k) e^{i(kz-\omega t)}$$

Suppose that our choice of wavepacket $\mathbf{E}(k)$ is heavily peaked around some fixed wavenumber k_0. Then we can expand the exponent as

$$\begin{aligned} kz - \omega(k)t &\approx kz - \omega(k_0)t - \left.\frac{d\omega}{dk}\right|_{k_0} (k - k_0)t \\ &= -[\omega(k_0) + v_g(k_0)]t + k[z - v_g(k_0)t] \end{aligned}$$

The first term is just a constant oscillation in time; the second term is the one of interest. It tells us that the peak of the wave pulse is moving to the right with approximate speed $v_g(k_0)$.

Following (7.15), we also define the index of refraction

$$n(\omega) = \frac{c}{v_p(\omega)}$$

This allows us to write a relation between the group and phase velocities:

$$\frac{1}{v_g} = \frac{dk_1}{d\omega} = \frac{d}{d\omega}\left(\frac{n\omega}{c}\right) = \frac{1}{v_p} + \frac{\omega}{c}\frac{dn}{\omega}$$

Materials with $dn/d\omega > 0$ have $v_g < v_p$; this is called *normal dispersion.* Materials with $dn/d\omega < 0$ have $v_g > v_p$; this is called *anomalous dispersion.*

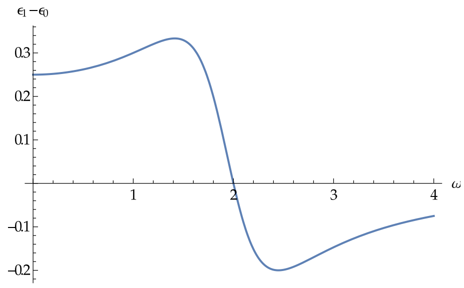


Figure 73: The real part of the permittivity, $\epsilon_1 - \epsilon_0$

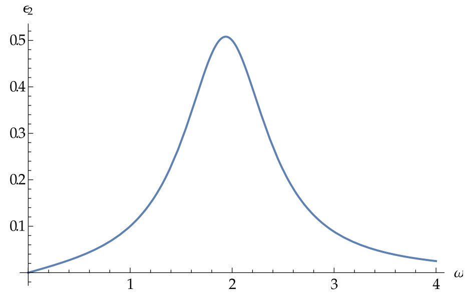


Figure 74: The imaginary part of the permittivity, ϵ_2

7.5.3 A Model for Dispersion

Let's see how this story works for our simple model of atomic polarisability $\alpha(\omega)$ given in (7.29). The permittivity is $\epsilon(\omega) = \epsilon_0 + n\alpha(\omega)$ where n is the density of atoms. The real and imaginary parts $\epsilon = \epsilon_1 + i\epsilon_2$ are

$$\begin{aligned} \epsilon_1 &= \epsilon_0 - \frac{nq^2}{m}\frac{\omega^2 - \omega_0^2}{(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2} \\ \epsilon_2 &= \frac{nq^2}{m}\frac{\gamma\omega}{(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2} \end{aligned}$$

These functions look like this: (These particular plots are made with $\gamma = 1$ and $\omega_0 = 2$ and $nq^2/m = 1$).
The real part is an even function: it has a maximum at $\omega = \omega_0 - \gamma/2$ and a minimum at $\omega = \omega_0 + \gamma/2$, each offset from the resonant frequency by an amount proportional to the damping γ. The imaginary part is an odd function; it has a maximum at $\omega = \omega_0$, the resonant frequency of the atom. The width of the imaginary part is roughly $\gamma/2$.

A quantity that will prove important later is the *plasma frequency*, ω_p. This is defined as

$$\omega_p^2 = \frac{nq^2}{m\epsilon_0} \tag{7.35}$$

We'll see the relevance of this quantity in Section 7.6. But for now it will simply be a useful combination that appears in some formulae below.

The dispersion relation (7.32) tells us

$$k_1^2 - k_2^2 + 2ik_1k_2 = \omega^2\mu(\epsilon_1 + i\epsilon_2)$$

Equating real and imaginary parts, we have

$$
\begin{aligned}
k_1 &= \pm\omega\sqrt{\mu}\left(\frac{1}{2}\sqrt{\epsilon_1^2+\epsilon_2^2}+\frac{1}{2}\epsilon_1\right)^{1/2} \\
k_2 &= \pm\omega\sqrt{\mu}\left(\frac{1}{2}\sqrt{\epsilon_1^2+\epsilon_2^2}-\frac{1}{2}\epsilon_1\right)^{1/2}
\end{aligned} \tag{7.36}
$$

To understand how light propagates through the material, we need to look at the values of k_1 and k_2 for different values of the frequency. There are three different types of behaviour.

Transparent Propagation: Very high or very low frequencies

The most straightforward physics happens when $\epsilon_1 > 0$ and $\epsilon_1 \gg \epsilon_2$. For our simple model, this ocurs when $\omega < \omega_0 - \gamma/2$ or when $\omega > \omega_\star$, the value at which $\epsilon_1(\omega_\star) = 0$.

Expanding to leading order, we have

$$
k_1 \approx \pm\omega\sqrt{\mu\epsilon_1} \quad \text{and} \quad k_2 \approx \pm\omega\sqrt{\frac{\mu\epsilon_2^2}{4\epsilon_1}} = \left(\frac{\epsilon_2}{2\epsilon_1}\right)k_1 \ll k_1
$$

Because $k_2 \ll k_1$, the damping is small. This means that the material is transparent at these frequencies.

There's more to this story. For the low frequencies, $\epsilon_1 > \epsilon_0 + nq^2/m\omega_0^2$. This is the same kind of situation that we dealt with in Section 7.3. The phase velocity $v_p < c$ in this regime. For high frequencies, however, $\epsilon_1 < \epsilon_0$; in fact, $\epsilon_1(\omega) \to \epsilon_1$ from below as $\omega \to \infty$. This means that $v_p > c$ in this region. This is nothing to be scared of! The plane wave is already spread throughout space; it's not communicating any information faster than light. Instead, pulses propagate at the group velocity, v_g. This is less than the speed of light, $v_g < c$, in both high and low frequency regimes.

Resonant Absorption: $\omega \approx \omega_0$

Resonant absorption occurs when $\epsilon_2 \gg |\epsilon_1|$. In our model, this phenomenon is most pronounced when $\omega_0 \gg \gamma$ so that the resonant peak of ϵ_2 is sharp. Then for frequencies close to the resonance, $\omega \approx \omega_0 \pm \gamma/2$, we have

$$
\epsilon_1 \approx \epsilon_0 \quad \text{and} \quad \epsilon_2 \approx \frac{nq^2}{m}\frac{1}{\omega_0\gamma} = \epsilon_0\left(\frac{\omega_p}{\omega_0}\right)^2\frac{\omega_0}{\gamma}
$$

We see that we meet the requirement for resonant absorption if we also have $\omega_p \gtrsim \omega_0$. When $\epsilon_2 \gg |\epsilon_1|$, we can expand (7.36) to find

$$k_1 \approx k_2 \approx \pm\omega\sqrt{\frac{\mu\epsilon_2}{2}}$$

The fact that $k_2 \approx k_1$ means that the wave decays very rapidly: it has effectively disappeared within just a few wavelengths of propagation. This is because the frequency of the wave is tuned to coincide with the natural frequency of the atoms, which easily become excited, absorbing energy from the wave.

<u>Total Reflection:</u>

The third region of interest occurs when $\epsilon_1 < 0$ and $|\epsilon_1| \gg \epsilon_2$. In our model, it is roughly for frequencies $\omega_0 + \gamma/2 < \omega < \omega_\star$. Now, the expansion of (7.36) gives

$$k_1 \approx \pm\omega\sqrt{\mu}\left(\frac{1}{2}|\epsilon_1| + \frac{1}{4}\frac{\epsilon_2^2}{|\epsilon_1|} + \frac{1}{2}\epsilon_1 + \ldots\right)^{1/2} \approx \pm\omega\frac{\epsilon_2}{2}\sqrt{\frac{\mu}{|\epsilon_1|}}$$

and

$$k_2 \approx \pm\omega\sqrt{\mu|\epsilon_1|} = \frac{|\epsilon_1|}{2\epsilon_2}k_1 \gg k_1$$

Now the wavenumber is almost pure imaginary. The wave doesn't even manage to get a few wavelengths before it decays. It's almost all gone before it even travels a single wavelength.

We're not tuned to the resonant frequency, so this time the wave isn't being absorbed by the atoms. Instead, the applied electromagnetic field is almost entirely cancelled out by the induced electric and magnetic fields due to polarisation.

7.5.4 Causality and the Kramers-Kronig Relation

Throughout this section, we used the relationship between the polarisation $\mathbf{p}$ and applied electric field $\mathbf{E}$. In frequency space, this reads

$$\mathbf{p}(\omega) = \alpha(\omega)\mathbf{E}(\omega) \tag{7.37}$$

Relationships of this kind appear in many places in physics. The polarisability $\alpha(\omega)$ is an example of a *response function*. As their name suggests, such functions tell us how some object – in this case $\mathbf{p}$ – respond to a change in circumstance – in this case, the application of an electric field.

There is a general theory around the properties of response functions[5]. The most important fact follows from causality. The basic idea is that if we start off with a vanishing electric field and turn it on only at some fixed time, $t_\star$, then the polarisation shouldn't respond to this until after $t_\star$. This sounds obvious. But how is it encoded in the mathematics?

The causality properties are somewhat hidden in (7.37) because we're thinking of the electric field as oscillating at some fixed frequency, which implicitly means that it oscillates for all time. If we want to turn the electric field on and off in time, when we need to think about superposing fields of lots of different frequencies. This, of course, is the essence of the Fourier transform. If we shake the electric field at lots of different frequencies, its time dependence is given by

$$\mathbf{E}(t) = \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \mathbf{E}(\omega)\, e^{-i\omega t}$$

where, if we want $\mathbf{E}(t)$ to be real, we should take $\mathbf{E}(-\omega) = \mathbf{E}(\omega)^\star$. Conversely, for a given time dependence of the electric field, the component at some frequency ω is given by the inverse Fourier transform,

$$\mathbf{E}(\omega) = \int_{-\infty}^{+\infty} dt\; \mathbf{E}(t)\, e^{i\omega t}$$

Let's now see what this tells us about the time dependence of the polarisation $\mathbf{p}$. Using (7.37), we have

$$\begin{aligned}\mathbf{p}(t) &= \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \mathbf{p}(\omega)\, e^{-i\omega t} \\ &= \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \alpha(\omega) \int_{-\infty}^{+\infty} dt'\; \mathbf{E}(t')\, e^{-i\omega(t-t')} \\ &= \int_{-\infty}^{+\infty} dt'\; \tilde{\alpha}(t-t')\mathbf{E}(t') \end{aligned} \tag{7.38}$$

where, in the final line, we've introduced the Fourier transform of the polarisability,

$$\tilde{\alpha}(t) = \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\, \alpha(\omega)\, e^{-i\omega t} \tag{7.39}$$

(Note that I've been marginally inconsistent in my notation here. I've added the tilde above $\tilde{\alpha}$ to stress that this is the Fourier transform of $\alpha(\omega)$ even though I didn't do the same to $\mathbf{p}$ and $\mathbf{E}$).

[5] You can learn more about this in the *Response Functions* section of the lectures on *Kinetic Theory.*

Equation (7.38) relates the time dependence of $\mathbf{p}$ to the time dependence of the electric field $\mathbf{E}$. It's telling us that the effect isn't immediate; the polarisation at time t depends on what the electric field was doing at all times t'. But now we can state the requirement of causality: the response function must obey

$$\tilde{\alpha}(t) = 0 \quad \text{for} \quad t < 0$$

Using (7.39), we can translate this back into a statement about the response function in frequency space. When $t < 0$, we can perform the integral over ω by completing the contour in the upper-half plane as shown in the figure. Along the extra semi-circle, the exponent is $-i\omega t \to -\infty$ for $t < 0$, ensuring that this part of the integral vanishes. By the residue theorem, the integral is just given by the sum of residues inside the contour. If we want $\alpha(t) = 0$ for $t < 0$, we need there to be no poles. In other words, we learn that

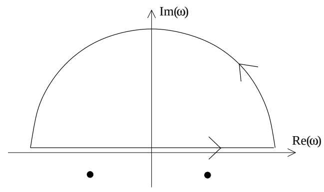


Figure 75:

$$\alpha(\omega) \text{ is analytic for } \operatorname{Im}\omega > 0$$

In contrast, $\alpha(\omega)$ can have poles in the lower-half imaginary plane. For example, if you look at our expression for the polarisability in (7.29), you can see that there are two poles at $\omega = -i\gamma/2 \pm \sqrt{\omega_0^2 - \gamma^2/4}$. Both lie in the lower-half of the complex ω plane.

The fact that α is analytic in the upper-half plane means that there is a relationship between its real and imaginary parts. This is called the Kramers-Kronig relation. Our task in this section is to derive it. We start by providing a few general mathematical statements about complex integrals.

A Discontinuous Function

First, consider a general function $\rho(\omega)$. We'll ask that $\rho(\omega)$ is meromorphic, meaning that it is analytic apart from at isolated poles. But, for now, we won't place any restrictions on the position of these poles. (We will shortly replace $\rho(\omega)$ by $\alpha(\omega)$ which, as we've just seen, has no poles in the upper half plane). We can define a new function $f(\omega)$ by the integral,

$$f(\omega) = \frac{1}{i\pi} \int_a^b \frac{\rho(\omega')}{\omega' - \omega} d\omega' \tag{7.40}$$

Here the integral is taken along the interval $\omega' \in [a, b]$ of the real line. However, when ω also lies in this interval, we have a problem because the integral diverges at $\omega' = \omega$.

To avoid this, we can simply deform the contour of the integral into the complex plane, either running just above the singularity along $\omega' + i\epsilon$ or just below the singularity along $\omega' - i\epsilon$. Alternatively (in fact, equivalently) we could just shift the position of the singularity to $\omega \to \omega \mp \epsilon$. In both cases we just skim by the singularity and the integral is well defined. The only problem is that we get different answers depending on which way we do things. Indeed, the difference between the two answers is given by Cauchy's residue theorem,

$$\frac{1}{2}[f(\omega + i\epsilon) - f(\omega - i\epsilon)] = \rho(\omega) \tag{7.41}$$

The difference between $f(\omega+i\epsilon)$ and $f(\omega-i\epsilon)$ means that the function $f(\omega)$ is discontinuous across the real axis for $\omega \in [a, b]$. If $\rho(\omega)$ is everywhere analytic, this discontinuity is a branch cut.

We can also define the average of the two functions either side of the discontinuity. This is usually called the *principal value*, and is denoted by adding the symbol $\mathcal{P}$ before the integral,

$$\frac{1}{2}[f(\omega + i\epsilon) + f(\omega - i\epsilon)] \equiv \frac{1}{i\pi}\mathcal{P}\int_a^b \frac{\rho(\omega')}{\omega' - \omega} d\omega' \tag{7.42}$$

We can get a better handle on the meaning of this principal part if we look at the real and imaginary pieces of the denominator in the integrand $1/[\omega' - (\omega \pm i\epsilon)]$,

$$\frac{1}{\omega' - (\omega \pm i\epsilon)} = \frac{\omega' - \omega}{(\omega' - \omega)^2 + \epsilon^2} \pm \frac{i\epsilon}{(\omega' - \omega)^2 + \epsilon^2} \tag{7.43}$$

The real and imaginary parts of this function are shown in the figures.

We can isolate the real part by taking the sum of $f(\omega + i\epsilon)$ and $f(\omega - i\epsilon)$ in (7.42). It can be thought of as a suitably cut-off version of $1/(\omega' - \omega)$. It's as if we have deleted an small segment of this function lying symmetrically about divergent point ω and replaced it with a smooth function going through zero. This is the usual definition of the principal part of an integral.

Similarly, the imaginary part can be thought of as a regularised delta-function. As $\epsilon \to 0$, it tends towards a delta function, as expected from (7.41).

Kramers-Kronig

Let's now apply this discussion to our polarisability response function $\alpha(\omega)$. We'll be interested in the integral

$$\frac{1}{i\pi}\oint_C d\omega' \frac{\alpha(\omega')}{\omega' - \omega} \qquad \omega \in \mathbf{R} \tag{7.44}$$

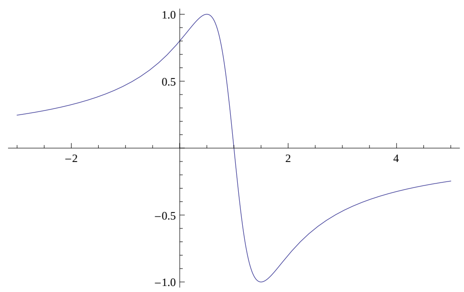


Figure 76: The real part of the function plotted with $\omega' = 1$ and $\epsilon = 0.5$.

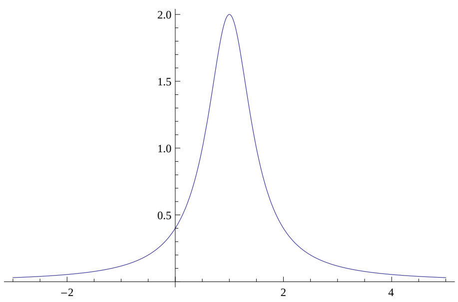


Figure 77: The imaginary part of the function plotted with $\omega' = 1$ and $\epsilon = 0.5$

where the contour C skims just above the real axis, before closing at infinity in the upper-half plane. We'll need to make one additional assumption: that $\alpha(\omega)$ falls off faster than $1/|\omega|$ at infinity. If this holds, the integral is the same as we consider in (7.40) with $[a, b] \to [-\infty, +\infty]$. Indeed, in the language of the previous discussion, the integral is $f(\omega - i\epsilon)$, with $\rho = \alpha$.

We apply the formulae (7.41) and (7.42). It gives

$$f(\omega - i\epsilon) = \frac{1}{i\pi}\mathcal{P}\left[\int_{-\infty}^{+\infty} d\omega'\, \frac{\alpha(\omega')}{\omega' - \omega}\right] - \alpha(\omega)$$

But we know the integral in (7.44) has to be zero since $\alpha(\omega)$ has no poles in the upper-half plane. This means that $f(\omega - i\epsilon) = 0$, or

$$\alpha(\omega) = \frac{1}{i\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega'\, \frac{\alpha(\omega')}{\omega' - \omega}$$

The important part for us is that factor of "i" sitting in the denominator. Taking real and imaginary parts, we learn that

$$\mathrm{Re}\,\alpha(\omega) = \mathcal{P}\int_{-\infty}^{+\infty} \frac{d\omega'}{\pi}\frac{\mathrm{Im}\,\alpha(\omega')}{\omega' - \omega}$$

and

$$\mathrm{Im}\,\alpha(\omega) = -\mathcal{P}\int_{-\infty}^{+\infty} \frac{d\omega'}{\pi}\frac{\mathrm{Re}\,\alpha(\omega')}{\omega' - \omega}$$

These are the *Kramers-Kronig* relations. They follow from causality alone and tell us that the imaginary part of the response function is determined in terms of the real part, and vice-versa. However, the relationship is not local in frequency space: you need to know $\mathrm{Re}\,\alpha(\omega)$ for all frequencies in order to reconstruct $\mathrm{Im}\,\alpha(\omega)$ for any single frequency.

7.6 Conductors Revisited

Until now, we've only discussed electromagnetic waves propagating through insulators. (Or, dielectrics to give them their fancy name). What happens in conductors where electric charges are free to move? We met a cheap model of a conductor in Section 2.4, where we described them as objects which screen electric fields. Here we'll do a slightly better job and understand how this happens dynamically.

7.6.1 The Drude Model

The Drude model is simple. Really simple. It describes the electrons moving in a conductor as billiard-balls, bouncing off things. The electrons have mass m, charge q and velocity $\mathbf{v} = \dot{\mathbf{r}}$. We treat them classically using $F = ma$; the equation of motion is

$$m\frac{d\mathbf{v}}{dt} = q\mathbf{E} - \frac{m}{\tau}\mathbf{v} \tag{7.45}$$

The force is due to an applied electric field $\mathbf{E}$, together with a linear friction term. This friction term captures the effect of electrons hitting things, whether the background lattice of fixed ions, impurities, or each other. (Really, these latter processes should be treated in the quantum theory but we'll stick with a classical treatment here). The coefficient τ is called the *scattering time*. It should be thought of as the average time that the electron travels before it bounces off something. For reference, in a good metal, $\tau \approx 10^{-14}\,s$. (Note that this friction term is the same as (7.27) that we wrote for the atomic polarisability, although the mechanisms behind it may be different in the two cases).

We start by applying an electric field which is constant in space but oscillating in time

$$\mathbf{E} = \mathbf{E}(\omega)e^{-i\omega t}$$

This can be thought of as applying an AC voltage to a conductor. We look for solutions of the form

$$\mathbf{v} = \mathbf{v}(\omega)\,e^{-i\omega t}$$

Plugging this into (7.45) gives

$$\left(-i\omega + \frac{1}{\tau}\right)\mathbf{v}(\omega) = \frac{q}{m}\mathbf{E}(\omega)$$

The current density is $\mathbf{J} = nq\mathbf{v}$, where n is the density of charge carriers, so the solution tells us that

$$\mathbf{J}(\omega) = \sigma(\omega)\mathbf{E}(\omega) \tag{7.46}$$

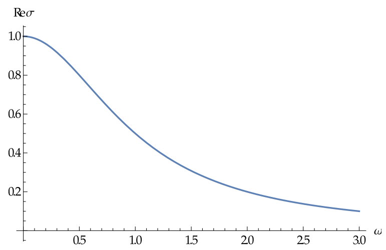


Figure 78: The real, dissipative part of the conductivity

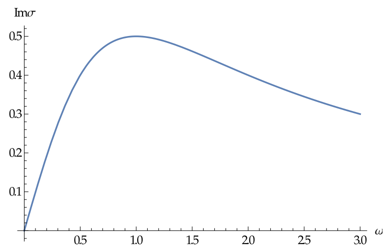


Figure 79: The imaginary, reactive part of the conductivity

This, of course, is Ohm's law. The proportionality constant $\sigma(\omega)$ depends on the frequency and is given by

$$\sigma(\omega) = \frac{\sigma_{\rm DC}}{1 - i\omega\tau} \tag{7.47}$$

It is usually referred to as the *optical conductivity*. In the limit of vanishing frequency, $\omega = 0$, it reduces to the *DC conductivity*,

$$\sigma_{\rm DC} = \frac{nq^2\tau}{m}$$

The DC conductivity is real and is inversely related to the *resistivity* $\rho = 1/\sigma_{DC}$. In contrast, the optical conductivity is complex. Its real and imaginary parts are given by

$$\mathrm{Re}\,\sigma(\omega) = \frac{\sigma_{\rm DC}}{1+\omega^2\tau^2} \quad \text{and} \quad \mathrm{Im}\,\sigma(\omega) = \frac{\sigma_{\rm DC}\,\omega\tau}{1+\omega^2\tau^2}$$

These are plotted for $\sigma_{DC} = 1$ and $\tau = 1$:

The conductivity is complex simply because we're working in Fourier space. The real part tells us about the dissipation of energy in the system. The bump at low frequencies, $\omega \sim 1/\tau$, is referred to as the *Drude peak*. The imaginary part of the conductivity tells us about the response of the system. (To see how this is relevant note that, in the Fourier ansatz, the velocity is related to the position by $\mathbf{v} = \dot{\mathbf{r}} = -i\omega\mathbf{r}$). At very large frequencies, $\omega\tau \gg 1$, the conductivity becomes almost purely imaginary, $\sigma(\omega) \sim i/\omega\tau$. This should be thought of as the conductivity of a free particle; you're shaking it so fast that it turns around and goes the other way before it's had the chance to hit something.

Although we derived our result (7.47) using a simple, Newtonian model of free electrons, the expression for the conductivity itself is surprisingly robust. In fact, it survives just about every subsequent revolution in physics; the development of quantum mechanics and Fermi surfaces, the presence of lattices and Bloch waves, even interactions between electrons in a framework known as Landau's Fermi liquid model. In all of these, the optical conductivity (7.47) remains the correct answer[6]. (This is true, at least, at low frequencies, At very high frequencies other effects can come in and change the story).

7.6.2 Electromagnetic Waves in Conductors

Let's now ask our favourite question: how do electromagnetic waves move through a material? The macroscopic Maxwell equations (7.14) that we wrote before assumed that there were no free charges or currents around. Now we're in a conductor, we need to include the charge density and current terms on the right-hand side:

$$\begin{aligned} \nabla\cdot\mathbf{D} &= \rho \qquad \text{and} \qquad \nabla\times\mathbf{H} = \mathbf{J} + \frac{\partial\mathbf{D}}{\partial t} \\ \nabla\cdot\mathbf{B} &= 0 \qquad \text{and} \qquad \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \end{aligned} \tag{7.48}$$

It's important to remember that here ρ refers only to the free charge. (We called it ρ_{free} in Section 7.1). We can still have bound charge in conductors, trapped around the ions of the lattice, but this has already been absorbed in the definition of $\mathbf{D}$ which is given by

$$\mathbf{D} = \epsilon(\omega)\mathbf{E}$$

Similarly, the current $\mathbf{J}$ is due only to the free charge.

We now apply a spatially varying, oscillating electromagnetic field, using the familiar ansatz,

$$\mathbf{E}(\mathbf{x},t) = \mathbf{E}(\omega)e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \quad \text{and} \quad \mathbf{B}(\mathbf{x},t) = \mathbf{B}(\omega)e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \tag{7.49}$$

At this point, we need to make an do something that isn't obviously allowed: we will continue to use Ohm's law (7.46), even in the presence of a varying electric field, so that

$$\mathbf{J}(\mathbf{x},t) = \sigma(\omega)\mathbf{E}(\omega)e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \tag{7.50}$$

This looks dubious; we derived Ohm's law by assuming that the electric field was the same everywhere in space. Why do we now get to use it when the electric field varies?

[6] As an extreme example, the conductivity of the horizon of certain black holes can be computed in general relativity. Even here, the result at low frequency is given by the simple Drude formula (7.47)! Details can be found in Gary Horowitz, Jorge Santos and David Tong, "*Optical Conductivity with Holographic Lattices*, arXiv:1204.0519.

For this to be valid, we need to assume that over the time scales τ, relevant in the derivation of Ohm's law, the electric field is more or less constant. This will be true if the wavelength of the electric field, $\lambda = 2\pi/|\mathbf{k}|$ is greater than the distance travelled by the electrons between collisions. This distance, known as the *mean free path*, is given by $l = \langle v \rangle \tau$, where v is the average speed. In most metals, $l \approx 10^{-7}\ m$. (This is around 1000 lattice spacings; to understand how it can be so large requires a quantum treatment of the electrons). This means that we should be able to trust (7.50) for wavelengths $\lambda \gtrsim l \approx 10^{-7}\ m$, which is roughly around the visible spectrum.

The continuity equation $\nabla \cdot \mathbf{J} + d\rho/dt = 0$ tells us that if the current oscillates, then the charge density must as well. In Fourier space, the continuity equation becomes

$$\rho = \frac{\mathbf{k}\cdot\mathbf{J}}{\omega} = \frac{\sigma(\omega)}{\omega}\,\mathbf{k}\cdot\mathbf{E}(\omega)\, e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)} \tag{7.51}$$

We can now plug these ansatze into the Maxwell equations (7.48). We also need $\mathbf{B} = \mu\mathbf{H}$ where, as previously, we'll take μ to be independent of frequency. We have

$$\begin{aligned} \nabla\cdot\mathbf{D} = \rho \quad &\Rightarrow \quad i\left(\epsilon(\omega) + i\frac{\sigma(\omega)}{\omega}\right)\mathbf{k}\cdot\mathbf{E}(\omega) = 0 \\ \nabla\cdot\mathbf{B} = 0 \quad &\Rightarrow \quad \mathbf{k}\cdot\mathbf{B}(\omega) = 0 \end{aligned} \tag{7.52}$$

As before, these tell us that the electric and magnetic fields are transverse to the direction of propagation. Although, as we mentioned previously, there is a caveat to this statement: if we can find a frequency for which $\epsilon(\omega)+i\sigma(\omega)/\omega = 0$ then longitudinal waves are allowed for the electric field. We will discuss this possibility in Section 7.6.3. For now focus on the transverse fields $\mathbf{k}\cdot\mathbf{E} = \mathbf{k}\cdot\mathbf{B} = 0$.

The other two equations are

$$\begin{aligned} \nabla\times\mathbf{H} = \mathbf{J} + \frac{\partial\mathbf{D}}{\partial t} \quad &\Rightarrow \quad i\mathbf{k}\times\mathbf{B}(\omega) = -i\mu\omega\left(\epsilon(\omega) + i\frac{\sigma(\omega)}{\omega}\right)\mathbf{E}(\omega) \\ \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad &\Rightarrow \quad \mathbf{k}\times\mathbf{E}(\omega) = \omega\mathbf{B}(\omega) \end{aligned}$$

The end result is that the equations governing waves in a conductor take exactly the same form as those derived in (7.30) governing waves in an insulator. The only difference is that we have to make the substitution

$$\epsilon(\omega) \;\longrightarrow \epsilon^{\text{eff}}(\omega) = \epsilon(\omega) + i\frac{\sigma(\omega)}{\omega}$$

This means that we can happily import our results from Section 7.5. In particular, the dispersion relation is given by

$$\mathbf{k}\cdot\mathbf{k} = \mu\epsilon^{\text{eff}}(\omega)\omega^2 \tag{7.53}$$

Let's now see how this extra term affects the physics, assuming that the optical conductivity takes the Drude form

$$\sigma(\omega) = \frac{\sigma_{DC}}{1 - i\omega\tau}$$

Low Frequencies

At frequencies that are low compared to the scattering time, $\omega\tau \ll 1$, we have $\sigma(\omega) \approx \sigma_{DC}$. This means that the real and imaginary parts of $\epsilon^{\rm eff}$ are

$$\epsilon^{\rm eff} = \epsilon_1^{\rm eff} + i\epsilon_2^{\rm eff} \approx \epsilon_1 + i\left(\epsilon_2 + \frac{\sigma_{DC}}{\omega}\right) \tag{7.54}$$

For sufficiently small ω, we always have $\epsilon_2^{\rm eff} \gg \epsilon_1^{\rm eff}$. This is the regime that we called *resonant absorption* in Section 7.5. The physics here is the same; no waves can propagate through the conductor; all are absorbed by the mobile electrons.

In this regime, the effective dielectric constant is totally dominated by the contribution from the conductivity and is almost pure imaginary: $\epsilon^{\rm eff} \approx i\sigma_{DC}/\omega$. The dispersion relation (7.53) then tells us that the wavenumber is

$$k = k_1 + ik_2 = \sqrt{i\mu\omega\sigma_{DC}} = \sqrt{\frac{\mu\omega\sigma_{DC}}{2}}\,(1+i)$$

So $k_1 = k_2$. This means that, for a wave travelling in the z-direction, so $\mathbf{k} = k\hat{\mathbf{z}}$, the electric field takes the form

$$\mathbf{E}(z,t) = \mathbf{E}(\omega)e^{-z/\delta}\,e^{i(k_1 z - \omega t)}$$

where

$$\delta = \frac{1}{k_2} = \sqrt{\frac{2}{\mu\omega\sigma_{DC}}}$$

The distance δ is called the *skin depth*. It is the distance that electromagnetic waves will penetrate into a conductor. Not that as $\omega \to 0$, the waves get further and further.

The fact that $k_1 = k_2$ also tells us, through (7.34), that the electric and magnetic fields oscillate $\pi/4$ out of phase. (The phase difference is given by $\tan\phi = k_2/k_1$). Finally, the magnitudes of the ratio of the electric and magnetic field amplitudes are given by

$$\frac{|\mathbf{B}(\omega)|}{|\mathbf{E}(\omega)|} = \frac{k}{\omega} = \sqrt{\frac{\mu\sigma_{DC}}{\omega}}$$

As $\omega \to 0$, we see that more and more of the energy lies in the magnetic, rather than electric, field.

High Frequencies

Let's now look at what happens for high frequencies. By this, we mean both $\omega\tau \gg 1$, so that $\sigma(\omega) \approx i\sigma_{\rm DC}/\omega\tau$ and $\omega \gg \omega_0$ so that $\epsilon(\omega) \approx \epsilon_0$. Now the effective permittivity is more or less real,

$$\epsilon^{\rm eff}(\omega) \approx \epsilon_0 - \frac{\sigma_{\rm DC}}{\omega^2\tau} = \epsilon_0\left(1 - \frac{\omega_p^2}{\omega^2}\right) \tag{7.55}$$

where we are are using the notation of the plasma frequency $\omega_p^2 = nq^2/m\epsilon_0$ that we introduced in (7.35). What happens next depends on the sign of $\epsilon^{\rm eff}$:

- $\omega > \omega_p$: At these high frequencies, $\epsilon^{\rm eff} > 0$ and k is real. This is the regime of transparent propagation. We see that, at suitably high frequencies, conductors become transparent. The dispersion relation is $\omega^2 = \omega_p^2 + c^2k^2$.
- $\omega < \omega_p$: This regime only exists if $\omega_p > \omega_0, 1/\tau$. (This is usually the case). Now $\epsilon^{\rm eff} < 0$ so k is purely imaginary. This is the regime of total reflection; no wave can propagate inside the conductor.

We see that the plasma frequency ω_p sets the lower-limit for when waves can propagate through a conductor. For most metals, $\omega_p^{-1} \approx 10^{-16}s$ with a corresponding wavelength of $\lambda_p \approx 3 \times 10^{-10}\ m$. This lies firmly in the ultraviolet, meaning that visible light is reflected. This is why most metals are shiny. (Note, however, that this is smaller than the wavelength that we needed to really trust (7.50); you would have to work harder to get a more robust derivation of this effect).

There's a cute application of this effect. In the upper atmosphere of the Earth, many atoms are ionised and the gas acts like a plasma with $\omega_p \approx 2\pi \times 9\ MHz$. Only electromagnetic waves above this frequency can make it through. This includes FM radio waves. But, in contrast, AM radio waves are below this frequency and bounce back to Earth. This is why you can hear AM radio far away. And why aliens can't.

7.6.3 Plasma Oscillations

We noted in (7.52) that there's a get out clause in the requirement that the electric field is transverse to the propagating wave. The Maxwell equation reads

$$\nabla \cdot \mathbf{D} = \rho \quad \Rightarrow \quad i\left(\epsilon(\omega) + i\frac{\sigma(\omega)}{\omega}\right)\mathbf{k}\cdot\mathbf{E}(\omega) = 0$$

Which means that we can have $\mathbf{k}\cdot\mathbf{E} \neq 0$ as long as $\epsilon^{\rm eff}(\omega) = \epsilon(\omega) + i\sigma(\omega)/\omega = 0$.

We could try to satisfy this requirement at low frequencies where the effective permittivity is given by (7.54). Since we typically have $\epsilon_1 \gg \epsilon_2$ in this regime, this is approximately

$$\epsilon^{\text{eff}}(\omega) \approx \epsilon_1 + i\frac{\sigma_{\text{DC}}}{\omega}$$

Which can only vanish if we take the frequency to be purely imaginary,

$$\omega = -i\frac{\sigma_{\text{DC}}}{\epsilon_1}$$

This is easy to interpret. Plugging it into the ansatz (7.49), we have

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}(\omega)\, e^{i\mathbf{k}\cdot\mathbf{x}}\, e^{-\sigma_{\text{DC}}t/\epsilon_1}$$

which is telling us that if you try to put such a low frequency longitudinal field in a conductor then it will decay in time $\sim \epsilon_1/\sigma_{\text{DC}}$. This is not the solution we're looking for.

More interesting is what happens at high frequencies, $\omega \gg 1/\tau, \omega_0$, where the effective permittivity is given by (7.55). It vanishes at $\omega = \omega_p$:

$$\epsilon^{\text{eff}}(\omega_p) \approx 0$$

Now we can have a new, propagating solution in which $\mathbf{B} = 0$, while $\mathbf{E}$ is parallel to $\mathbf{k}$. This is a longitudinal wave. It is given by

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}(\omega_p) e^{i(\mathbf{k}\cdot\mathbf{x} - \omega_p t)}$$

By the relation (7.51), we see that for these longitudinal waves the charge density is also oscillating,

$$\rho(\mathbf{x}, t) = \mathbf{k} \cdot \mathbf{E}(\omega_p) e^{i(\mathbf{k}\cdot\mathbf{x} - \omega_p t)}$$

These are called *plasma oscillations*.

Note that, while the frequency of oscillation is always ω_p, the wavenumber k can be anything. This slightly strange state of affairs is changed if you take into account thermal motion of the electrons. This results in an electron pressure which acts as a restoring force on the plasma, inducing a non-trivial dispersion relation. When quantised, the resulting particles are called *plasmons*.

7.6.4 Dispersion Relations in Quantum Mechanics

So far we've derived a number of dispersion relations for various wave excitations. In all cases, these become particle excitations when we include quantum mechanics.

The paradigmatic example is the way light waves are comprised of *photons*. These are massless particles with energy E and momentum $\mathbf{p}$ given by

$$E = \hbar\omega \quad \text{and} \quad \mathbf{p} = \hbar\mathbf{k} \tag{7.56}$$

With this dictionary, the wave dispersion relation becomes the familiar energy-momentum relation for massless particles that we met in our special relativity course,

$$\omega = kc \quad \Rightarrow \quad E = pc$$

The relationships (7.56) continue to hold when we quantise any other dispersion relation. However, one of the main lessons of this section is that both the wavevector and frequency can be complex. These too have interpretations after we quantise. A complex k means that the wave dies away quickly, typically after some boundary. In the quantum world, this just means that the particle excitations are confined close to the boundary. Meanwhile, an imaginary ω means that the wave dies down over time. In the quantum world, the imaginary part of ω has the interpretation as the lifetime of the particle.

7.7 Charge Screening

Take a system in which charges are free to move around. To be specific, we'll talk about a metal but everything we say could apply to any plasma. Then take another charge and place it at a fixed location in the middle of the system. This could be, for example, an impurity in the metal. What happens?

The mobile charges will be either attracted or repelled by the impurity. If the impurity has positive charge, the mobile, negatively charged electrons will want to cluster around it. The charge of these electrons acts to cancel out the charge of the impurity so that, viewed from afar, the region around the impurity will appear to have greatly reduced charge. There is a similar story if the charge of the impurity is negative; now the electrons are repelled, exposing the lattice of positively charged ions that lies underneath. Once again, the total charge of a region around the impurity will be greatly reduced. This is the phenomenon of *charge screening*.

Our goal here is to understand more quantitatively how this happens and, in particular, how the effective charge of the impurity changes as we move away from it. As we'll see, ultimately quantum effects will result in some rather surprising behaviour. I should mention that, unlike other parts of these notes, this section will need results from both quantum mechanics and statistical mechanics.

7.7.1 Classical Screening: The Debye-Hückel model

We'll start by looking at a simple classical model for charge screening which will give us some intuition for what's going on. Our metal consists of a mobile gas of electrons, each of charge q. These are described by a charge density $\rho(\mathbf{r})$. In the absence of any impurity, we would have $\rho(\mathbf{r}) = \rho_0$, some constant.

The entire metal is neutral. The charges of the mobile electrons are cancelled by the charges of the ions that they leave behind, fixed in position in the crystal lattice. Instead of trying to model this lattice with any accuracy, we'll simply pretend that it has a uniform, constant charge density $-\rho_0$, ensuring that the total system is neutral. This very simple toy model sometimes goes by the toy name of *jellium*.

Now we introduce the impurity by placing a fixed charge Q at the origin. We want to know how the electron density $\rho(\mathbf{r})$ responds. The presence of the impurity sets up an electric field, with the electrostatic potential $\phi(\mathbf{r})$ fixed by Gauss' law

$$\nabla^2\phi = -\frac{1}{\epsilon_0}\left(Q\delta^3(\mathbf{r}) - \rho_0 + \rho(\mathbf{r})\right) \tag{7.57}$$

Here the $-\rho_0$ term is due to the uniform background charge, while $\rho(\mathbf{r})$ is due to the electron density. It should be clear that this equation alone is not enough to solve for both $\rho(\mathbf{r})$ and $\phi(\mathbf{r})$. To make progress, we need to understand more about the forces governing the charge distribution $\rho(\mathbf{r})$. This sounds like it might be a difficult problem. However, rather than approach it as a problem in classical mechanics, we do something clever: we import some tools from statistical mechanics[7].

We place our system at temperature T. The charge density $\rho(\mathbf{r})$ will be proportional to the probability of finding a charge q at position $\mathbf{r}$. If we assume that there are no correlations between the electrons, this is just given by the Bolzmann distribution. The potential energy needed to put a charge q at position $\mathbf{r}$ is simply $q\phi(\mathbf{r})$ so we have

$$\rho(\mathbf{r}) = \rho_0\, e^{-q\phi(\mathbf{r})/k_BT} \tag{7.58}$$

where the normalisation ρ_0 is fixed by assuming that far from the impurity $\phi(\mathbf{r}) \to 0$ and the system settles down to its original state.

[7]See the lecture notes on *Statistical Physics*. The Debye-Hückel model was described in Section 2.6 of these notes.

The result (7.58) is a very simple solution to what looks like a complicated problem. Of course, in part this is the beauty of statistical mechanics. But there is also an important approximation that has gone into this result: we assume that a given electron feels the average potential produced by all the others. We neglect any fluctuations around this average. This is an example of the *mean field approximation*, sometimes called the *Hartree approximation*. (We used the same kind of trick in the *Statistical Physics* notes when we first introduced the Ising model).

For suitably large temperatures, we can expand the Boltzmann distribution and write

$$\rho(\mathbf{r}) \approx \rho_0 \left(1 - \frac{q\phi(\mathbf{r})}{k_B T} + \dots\right)$$

Substituting this into Gauss' law (7.57) then gives

$$\left(\nabla^2 - \frac{1}{\lambda_D^2}\right)\phi(\mathbf{r}) = -\frac{Q}{\epsilon_0}\delta^3(\mathbf{r})$$

where λ_D is called the *Debye screening length* (we'll see why shortly) and is given by

$$\lambda_D^2 = \frac{k_B T \epsilon_0}{q^2 n_0} \tag{7.59}$$

We've written this in terms of the number density n_0 of electrons instead of the charge density $\rho_0 = qn_0$. The solution to this equation is

$$\phi(\mathbf{r}) = \frac{Q\, e^{-r/\lambda_D}}{4\pi\epsilon_0 r} \tag{7.60}$$

This equation clearly shows the screening phenomenon that we're interested in. At short distances $r \ll \lambda_D$, the electric field due to the impurity doesn't look very much different from the familiar Coulomb field. But at larger distances $r \gg \lambda_D$, the screening changes the potential dramatically and it now dies off exponentially quickly rather than as a power-law. Note that the electrons become less efficient at screening the impurity as the temperature increases. In contrast, if we take this result at face value, it looks as if they can screen the impurity arbitrarily well at low temperatures. But, of course, the classical description of electrons is not valid at low temperatures. Instead we need to turn to quantum mechanics.

7.7.2 The Dielectric Function

Before we look at quantum versions of screening, it's useful to first introduce some new terminology. Let's again consider introducing an impurity into the system, this time with some fixed charge distribution $\rho^{\rm ext}(\mathbf{r})$, where "ext" stands for "external". We know that, taken on its own, this will induce a background electric field with potential

$$\nabla^2\phi^{\rm ext} = -\frac{\rho^{\rm ext}}{\epsilon_0}$$

But we also know that the presence of the impurity will affect the charge distribution of the mobile electrons. We'll call $\rho^{\rm ind}(\mathbf{r}) = \rho(\mathbf{r}) - \rho_0$ the "induced charge". We know that the actual electric field will be given by the sum of $\rho^{\rm ext}$ and $\rho^{\rm ind}$,

$$\nabla^2\phi = -\frac{1}{\epsilon_0}\left(\rho^{\rm ext}(\mathbf{r}) + \rho^{\rm ind}(\mathbf{r})\right)$$

This set-up is very similar to our discussion in Section 7.1 when we first introduced the idea of polarisation $\mathbf{P}$ and the electric displacement $\mathbf{D}$. In that case, we were interested in insulators and the polarisation described the response of bound charge to an applied electric field. Now we're discussing conductors and the polarisation should be thought of as the response of the mobile electrons to an external electric field. In other words, $\nabla\cdot\mathbf{P} = -\rho^{\rm ind}$. (Compare this to (7.5) for an insulator). Meanwhile, the electric displacement $\mathbf{D}$ is the electric field that you apply to the material, as opposed to $\mathbf{E}$ which is the actual electric field inside the material. In the present context, that means

$$\mathbf{E} = -\nabla\phi \quad \text{and} \quad \mathbf{D} = -\epsilon_0\nabla\phi^{\rm ext}$$

When we first introduced $\mathbf{E}$ and $\mathbf{D}$, we defined the relationship between them to be simply $\mathbf{D} = \epsilon\mathbf{E}$, where ϵ is the permittivity. Later, in Section 7.5, we realised that ϵ could depend on the frequency of the applied electric field. Now we're interested in static situations, so there's no frequency, but the electric fields vary in space. Therefore we shouldn't be surprised to learn that ϵ now depends on the wavelength, or wavevector, of the electric fields.

It's worth explaining a little more how this arises. The first thing we could try is to relate $\mathbf{E}(\mathbf{r})$ and $\mathbf{D}(\mathbf{r})$. The problem is that this relationship is not local in space. An applied electric field $\mathbf{D}(\mathbf{r})$ will move charges far away which, in turn, will affect the electric field $\mathbf{E}(\mathbf{r})$ far away. This means that, in real space, the relationship between $\mathbf{D}$ and $\mathbf{E}$ takes the form,

$$\mathbf{D}(\mathbf{r}) = \int d^3r'\ \epsilon(\mathbf{r}-\mathbf{r}')\mathbf{E}(\mathbf{r}') \tag{7.61}$$

The quantity $\epsilon(\mathbf{r}-\mathbf{r}')$ is known as the dielectric response function. It depends only on the difference $\mathbf{r}-\mathbf{r}'$ because the underlying system is translationally invariant. This relationship looks somewhat simpler if we Fourier transform and work in momentum space. We write

$$\mathbf{D}(\mathbf{k}) = \int d^3r\; e^{-i\mathbf{k}\cdot\mathbf{r}}\,\mathbf{D}(\mathbf{r}) \quad\Leftrightarrow\quad \mathbf{D}(\mathbf{r}) = \int \frac{d^3k}{(2\pi)^3}\; e^{i\mathbf{k}\cdot\mathbf{r}}\,\mathbf{D}(\mathbf{k})$$

and similar expressions for other quantities. (Note that we're using the notation in which the function and its Fourier transform are distinguished only by their argument). Taking the Fourier transform of both sides of (7.61), we have

$$\mathbf{D}(\mathbf{k}) = \int d^3r\; e^{-i\mathbf{k}\cdot\mathbf{r}}\mathbf{D}(\mathbf{r}) = \int d^3r \int d^3r'\; e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}\epsilon(\mathbf{r}-\mathbf{r}')\, e^{-i\mathbf{k}\cdot\mathbf{r}'}\mathbf{E}(\mathbf{r}')$$

But this final expression is just the product of two Fourier transforms. This tells us that we have the promised expression

$$\mathbf{D}(\mathbf{k}) = \epsilon(\mathbf{k})\mathbf{E}(\mathbf{k})$$

The quantity $\epsilon(\mathbf{k})$ is called the *dielectric function*. The constant permittivity that we first met in Section 7.1 is simply given by $\epsilon(\mathbf{k}\to 0)$.

In what follows, we'll work with the potentials ϕ and charge densities ρ, rather than $\mathbf{D}$ and $\mathbf{E}$. The dielectric function is then defined as

$$\phi^{\rm ext}(\mathbf{k}) = \epsilon(\mathbf{k})\phi(\mathbf{k}) \tag{7.62}$$

We write $\phi = \phi^{\rm ext} + \phi^{\rm ind}$, where

$$-\nabla^2\phi^{\rm ind} = \frac{\rho^{\rm ind}}{\epsilon_0} \quad\Rightarrow\quad k^2\phi^{\rm ind}(\mathbf{k}) = \frac{\rho^{\rm ind}(\mathbf{k})}{\epsilon_0}$$

Rearranging (7.62) then gives us an expression for the dielectric function in terms of the induced charge $\rho^{\rm ind}$ and the total electrostatic potential ϕ.

$$\epsilon(\mathbf{k}) = 1 - \frac{1}{\epsilon_0 k^2}\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} \tag{7.63}$$

This will turn out to be the most useful form in what follows.

Debye-Hückel Revisited

So far, we've just given a bunch of definitions. They'll be useful moving forward, but first let's see how we can recover the results of the Debye-Hückel model using this machinery. We know from (7.58) how the induced charge $\rho^{\rm ind}$ is related to the electrostatic potential,

$$\rho^{\rm ind}(\mathbf{r}) = \rho_0 \left(e^{-q\phi(\mathbf{r})/k_BT} - 1\right) \approx -\frac{q\rho_0\phi(\mathbf{r})}{k_BT} + \ldots \qquad (7.64)$$

To leading order, we then also get a linear relationship between the Fourier components,

$$\rho^{\rm ind}(\mathbf{k}) \approx -\frac{q\rho_0}{k_BT}\phi(\mathbf{k})$$

Substituting this into (7.63) gives us an expression for the dielectric function,

$$\epsilon(\mathbf{k}) = 1 + \frac{k_D^2}{k^2} \qquad (7.65)$$

where $k_D^2 = q\rho_0/\epsilon_0 k_BT = 1/\lambda_D^2$, with λ_D the Debye screening length that we introduced in (7.59).

Let's now see the physics that's encoded in the dielectric function. Suppose that we place a point charge at the origin. We have

$$\phi^{\rm ext}(\mathbf{r}) = \frac{Q}{4\pi\epsilon_0 r} \quad \Rightarrow \quad \phi^{\rm ext}(\mathbf{k}) = \frac{Q}{\epsilon_0 k^2}$$

Then, using the form of the dielectric function (7.65), the resulting electrostatic potential ϕ is given by

$$\phi(\mathbf{k}) = \frac{\phi^{\rm ext}(\mathbf{k})}{\epsilon(\mathbf{k})} = \frac{Q}{\epsilon_0(k^2 + k_D^2)}$$

We need to do the inverse Fourier transform of $\phi(\mathbf{k})$ to find $\phi(\mathbf{r})$. Let's see how to do it; we have

$$\phi(\mathbf{r}) = \int \frac{d^3k}{(2\pi)^3}\, e^{i\mathbf{k}\cdot\mathbf{r}}\phi(\mathbf{k}) = \frac{Q}{(2\pi)^3\epsilon_0}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta\, \sin\theta \int_0^\infty dk\, \frac{k^2}{k^2+k_D^2}\, e^{ikr\cos\theta}$$

where, in the second equality, we've chosen to work in spherical polar coordinates in which the k_z axis is aligned with $\mathbf{r}$, so that $\mathbf{k}\cdot\mathbf{r} = kr\cos\theta$. We do the integrals over the two angular variables, to get

$$\begin{aligned}
\phi(\mathbf{r}) &= \frac{Q}{(2\pi)^2\epsilon_0}\int_0^\infty dk\, \frac{k^2}{k^2+k_D^2}\,\frac{2\sin kr}{kr} \\
&= \frac{Q}{(2\pi)^2\epsilon_0 r}\int_{-\infty}^{\infty} dk\, \frac{k\sin kr}{k^2+k_D^2} \\
&= \frac{Q}{2\pi\epsilon_0 r}\,\mathrm{Re}\left[\int_{-\infty}^{+\infty}\frac{dk}{2\pi i}\,\frac{ke^{ikr}}{k^2+k_D^2}\right]
\end{aligned}$$

We compute this last integral by closing the contour in the upper-half plane with $k \to +i\infty$, picking up the pole at $k = +ik_D$. This gives our final answer for the electrostatic potential,

$$\phi(\mathbf{r}) = \frac{Q\, e^{-r/\lambda_D}}{4\pi\epsilon_0 r}$$

That's quite nice: we see that the dielectric function (7.65) contains the same physics (7.60) that we saw earlier in the direct computation of classical electrostatic screening. We could also compute the induced charge density to find

$$\rho^{\text{ind}}(\mathbf{r}) = -\frac{Qe^{-r/\lambda_D}}{4\pi\lambda_D^2 r}$$

which agrees with (7.64).

But the dielectric function $\epsilon(\mathbf{k})$ contains more information: it tells us how the system responds to each Fourier mode of an externally placed charge density. This means that we can use it to compute the response to any shape $\rho^{\text{ext}}(\mathbf{r})$.

Here, for example, is one very simple bit of physics contained in $\epsilon(\mathbf{k})$. In the limit $k \to 0$, we have $\epsilon(k) \to \infty$. This means that, in the presence of any constant, applied electric field $\mathbf{D}$, the electric field inside the material will be $\mathbf{E} = \mathbf{D}/\epsilon = 0$. But you knew this already: it's the statement that you can't have electric fields inside conductors because the charges will always move to cancel it. More generally, classical conductors will effectively screen any applied electric field which doesn't vary much on distances smaller than λ_D.

7.7.3 Thomas-Fermi Theory

The Debye-Hückel result describes screening by classical particles. But, as we lower the temperature, we know that quantum effects become important. Our first pass at this is called the *Thomas-Fermi approximation.* It's basically the same idea that we used in the Debye-Hückel approach, but with the probability determined by the Fermi-Dirac distribution rather than the classical Boltzmann distribution.

We work in the grand canonical ensemble, with temperature T and chemical potential μ. Recall that the probability of finding a fermion in a state $|k\rangle$ with energy E_k is given by the Fermi-Dirac distribution

$$f(\mathbf{k}) = \frac{1}{e^{(E_k-\mu)/k_BT} + 1} \tag{7.66}$$

The chemical potential μ is determined by the requirement that the equilibrium charge density is $\rho(\mu) = \rho_0$, where

$$\rho(\mu) = g_s \int \frac{d^3k}{(2\pi)^3} \, \frac{q}{e^{(E_k - \mu)/k_BT} + 1} \tag{7.67}$$

Here g_s is the spin degeneracy factor which we usually take to be $g_s = 2$.

Let's now place the external charge density $\rho^{\rm ext}(\mathbf{r})$ in the system. The story is the same as we saw before: the mobile charges move, resulting in an induced charge density $\rho^{\rm ind}(\mathbf{r})$, and a total electrostatic potential $\phi(\mathbf{r})$. The Thomas-Fermi approximation involves working with the new probability distribution

$$f(\mathbf{k}, \mathbf{r}) = \frac{1}{e^{(E_k + q\phi(\mathbf{r}) - \mu)/k_BT} + 1} \tag{7.68}$$

This can be thought of as either changing the energy to $E = E_k + q\phi(\mathbf{r})$ or, alternatively, allowing for a spatially varying chemical potential $\mu \rightarrow \mu - q\phi(\mathbf{r})$.

The first thing to say about the probability distribution (7.68) is that it doesn't make any sense! It claims to be the probability for a state with momentum $\mathbf{k}$ and position $\mathbf{r}$, yet states in quantum mechanics are, famously, not labelled by both momentum and position at the same time! So what's going on? We should think of (7.68) as an approximation that is valid when $\phi(\mathbf{r})$ is very slowly varying compared to any microscopic length scales. Then we can look in a patch of space where $\phi(\mathbf{r})$ is roughly constant and apply (7.68). In a neighbouring patch of space we again apply (7.68), now with a slightly different value of $\phi(\mathbf{r})$. This idea of *local equilibrium* underlies the Thomas-Fermi (and, indeed, the Debye-Hückel) approximations.

Let's see how this works in practice. The spatially dependent charge density is now given by

$$\rho(\mathbf{r}; \mu) = g_s \int \frac{d^3k}{(2\pi)^3} \, \frac{q}{e^{(E_k + q\phi(\mathbf{r}) - \mu)/k_BT} + 1} \tag{7.69}$$

We're interested in computing the induced charge density $\rho^{\rm ind}(\mathbf{r}) = \rho(\mathbf{r}) - \rho_0$. Combining (7.69) and (7.67), we have

$$\rho^{\rm ind}(\mathbf{r}) = g_s \int \frac{d^3k}{(2\pi)^3} \left[\frac{q}{e^{(E_k + q\phi(\mathbf{r}) - \mu)/k_BT} + 1} - \frac{q}{e^{(E_k - \mu)/k_BT} + 1} \right]$$

But we can rewrite this using the notation of (7.67) simply as

$$\rho^{\rm ind}(\mathbf{r}) = \rho\left(\mu - q\phi(\mathbf{r})\right) - \rho(\mu) \approx -q\phi(\mathbf{r}) \, \frac{\partial \rho(\mu)}{\partial \mu}$$

where, in the last step, we have Taylor expanded the function which is valid under the assumption that $q\phi(\mathbf{r}) \ll \mu$. But this immediately gives us an expression for the dielectric function using (7.63),

$$\epsilon(\mathbf{k}) = 1 + \frac{\partial\rho}{\partial\mu}\frac{q}{\epsilon_0 k^2}$$

We're almost there. We still need to figure out what $\partial\rho/\partial\mu$ is. This is particularly easy if we work at $T = 0$, where we can identify the chemical potential μ with the Fermi energy: $\mu = E_F$. In this case, the Fermi-Dirac distribution is a step function and the total charge density is simply given by

$$\rho(E_F) = q\int_0^{E_F} dE\, g(E)$$

where $g(E)$ is the density of states (we'll remind ourselves what form the density of states takes below). We learn that $\partial\rho/\partial E_F = g(E_F)$ and the dielectric function is given by

$$\epsilon(\mathbf{k}) = 1 + \frac{q^2 g(E_F)}{\epsilon_0 k^2} \tag{7.70}$$

Note that the functional form of $\epsilon(\mathbf{k})$ is exactly the same as we saw in the classical case (7.65). The only thing that's changed is the coefficient of the $1/k^2$ term which, as we saw before, determines the screening length. Let's look at a simple example.

A Simple Example

For non-relativistic particles, the energy is given by $E = \hbar^2 k^2/2m$. In three spatial dimensions, the density of states is given by[8]

$$g(E) = g_s \frac{1}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}$$

This is kind of a mess, but there's a neater way to write $g(E_F)$. (This neater way will also allow for a simple comparison to the Debye screening length as well). At zero temperature, the total charge density is

$$\rho_0 = q\int_0^{E_F} dE\ g(E)$$

[8]See the lecture notes on *Statistical Physics* for details on how to compute the density of states. The $g(E)$ we use here differs slightly from that presented in the Statistical Physics lectures because it does not include an overall volume factor. This is because we want to compute the number density of particles rather than the total number of particles.

Using this, we have

$$g(E_F) = \frac{3}{2q}\frac{\rho_0}{E_F}$$

and we can write the dielectric function as

$$\epsilon(\mathbf{k}) = 1 + \frac{k_{TF}^2}{k^2}$$

where $k_{TF}^2 = 3q\rho_0/2\epsilon_0 E_F$. This is our expression for the Thomas-Fermi screening length $\lambda_{TF} = 1/k_{TF}$.

It's instructive to compare this screening length with the classical Debye length λ_D. We have

$$\frac{\lambda_D^2}{\lambda_{TF}^2} = \frac{2}{3}\frac{T}{T_F}$$

where $T_F = k_B E_F$ is the Fermi temperature. The classical analysis can only be trusted at temperature $T \gg T_F$ where $\lambda_D \gg \lambda_{TF}$. But, for metals, the Fermi temperature is hot; something like $10^4, K$. This means that, at room temperature, $T \ll T_F$ and our quantum result above (which, strictly speaking, was only valid at $T = 0$) is a good approximation. Here $\lambda_D \ll \lambda_{TF}$. The upshot is that quantum mechanics acts to increase the screening length beyond that suggested by classical physics.

7.7.4 Lindhard Theory

The Thomas-Fermi approximation is straightforward, but it relies crucially on the potential $\phi(\mathbf{r})$ varying only over large scales. However, as we will now see, the most interesting physics arises due to variations of $\phi(\mathbf{r})$ over small scales (or, equivalently, large $\mathbf{k}$). For this we need to work harder.

The key idea is to go back to basics where, here, basics means quantum mechanics. Before we add the impurity, the energy eigenstates are plane waves $|\mathbf{k}\rangle$ with energy $E(k) = \hbar^2 k^2/2m$. To determine the dielectric function (7.63), we only need to know how the mobile charge density $\rho(\mathbf{r})$ changes in the presence of a potential $\phi(\mathbf{r})$. We can do this by considering a small perturbation to the Hamiltonian of the form

$$\Delta H = q\phi(\mathbf{r})$$

The energy eigenstate that is labelled by $\mathbf{k}$ now shifts. We call the new state $|\psi(\mathbf{k})\rangle$. Ultimately, our goal is to compute the induced charge density. For an electron in state $|\psi(\mathbf{k})\rangle$, the probabilty of finding it at position $\mathbf{r}$ is simply $|\langle\mathbf{r}|\psi(\mathbf{k})\rangle|^2$. Which means that, for this state, the change in the density is $|\langle\mathbf{r}|\psi(\mathbf{k})\rangle|^2 - |\langle\mathbf{r}|\mathbf{k}\rangle|^2$. The induced charge density $\rho^{\rm ind}(\mathbf{r})$ is obtained by summing over all such states, weighted with the Fermi-Dirac distribution function. We have

$$\rho^{\rm ind}(\mathbf{r}) = qg_s \int \frac{d^3k}{(2\pi)^3}\; f(k)\,\left[|\langle\mathbf{r}|\psi(\mathbf{k})\rangle|^2 - |\langle\mathbf{r}|\mathbf{k}\rangle|^2\right]$$

where $f(k)$ is the Fermi-Dirac distribution (7.66) and we've remembered to include the spin degeneracy factor $g_s = 2$. To make progress, we need to get to work computing the overlap of states.

To first order in perturbation theory, the new energy eigenstate is given by

$$|\psi(\mathbf{k})\rangle = |\mathbf{k}\rangle + \int \frac{d^3k'}{(2\pi)^3}\frac{\langle\mathbf{k}'|\Delta H|\mathbf{k}\rangle}{E(k)-E(k')}\,|\mathbf{k}'\rangle$$

Keeping only terms linear in ΔH, we can expand this out to read

$$|\langle\mathbf{r}|\psi(\mathbf{k})\rangle|^2 - |\langle\mathbf{r}|\mathbf{k}\rangle|^2 = \int \frac{d^3k'}{(2\pi)^3}\left[\langle\mathbf{r}|\mathbf{k}\rangle\frac{\langle\mathbf{k}|\Delta H|\mathbf{k}'\rangle}{E(k)-E(k')}\langle\mathbf{k}'|\mathbf{r}\rangle + \langle\mathbf{k}|\mathbf{r}\rangle\frac{\langle\mathbf{k}'|\Delta H|\mathbf{k}\rangle}{E(k)-E(k')}\langle\mathbf{r}|\mathbf{k}'\rangle\right]$$

But we have expressions for each of these matrix elements. Of course, the plane waves take the form $\langle\mathbf{r}|\mathbf{k}\rangle = e^{i\mathbf{k}\cdot\mathbf{r}}$, while the matrix elements of the perturbed Hamiltonian are

$$\langle\mathbf{k}'|q\phi(\mathbf{r})|\mathbf{k}\rangle = \int d^3r d^3r'\; e^{i(\mathbf{k}\cdot\mathbf{r}-\mathbf{k}'\cdot\mathbf{r}')}\langle\mathbf{r}'|q\phi(\mathbf{r})|\mathbf{r}\rangle = q\phi(\mathbf{k}-\mathbf{k}')$$

In other words, it gives the Fourier transform of the electrostatic potential. Putting this together, we arrive at an integral expression for the induced charge,

$$\rho^{\rm ind}(\mathbf{r}) = q^2 g_s\int \frac{d^3k}{(2\pi)^3}\frac{d^3k'}{(2\pi)^3}\; f(k)\;\left[\frac{e^{-i(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}}\phi(\mathbf{k}-\mathbf{k}')}{E(k)-E(k')} + \frac{e^{-i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}\phi(\mathbf{k}'-\mathbf{k})}{E(k)-E(k')}\right]$$

Of course, what we really want for the dielectric function (7.63) is the Fourier transform of the induced charge,

$$\rho^{\rm ind}(\mathbf{k}) = \int d^3r\, e^{-i\mathbf{k}\cdot\mathbf{r}}\rho^{\rm ind}(\mathbf{r})$$

Thankfully, doing the $\int d^3r$ integral gives rise to a delta-function which simplifies our life rather than complicating it. Performing some relabelling of dummy integration variables, we have

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s\int \frac{d^3k'}{(2\pi)^3}\,f(k')\,\left[\frac{1}{E(k')-E(|\mathbf{k}'-\mathbf{k}|)} + \frac{1}{E(k')-E(|\mathbf{k}+\mathbf{k}'|)}\right] \qquad (7.71)$$

These two terms are more similar than they look. If we change the dummy integration variable in the first term to $\mathbf{k}' \to \mathbf{k}' + \mathbf{k}$ then we can write

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{f(|\mathbf{k}+\mathbf{k}'|) - f(k')}{E(|\mathbf{k}+\mathbf{k}'|) - E(k')} \tag{7.72}$$

The left-hand side is exactly what we want. The right-hand side is an integral. It's not too hard to do this integral, but let's first check that this result gives something sensible.

Thomas-Fermi Revisited

Let's first see how we can recover the Thomas-Fermi result for the dielectric function. Recall that the Thomas-Fermi approximation was only valid when the potential $\phi(\mathbf{r})$, and hence the induced charge $\rho^{\rm ind}(\mathbf{r})$, vary slowly over large distances. In the present context, this means it is valid at small k. But here we can simply Taylor expand the numerator and denominator of (7.72).

$$E(|\mathbf{k}+\mathbf{k}'|) - E(k') \approx \frac{\partial E}{\partial \mathbf{k}'} \cdot \mathbf{k}$$

$$\text{and} \qquad f(|\mathbf{k}+\mathbf{k}'|) - f(k') \approx \frac{\partial f}{\partial E}\frac{\partial E}{\partial \mathbf{k}'} \cdot \mathbf{k}$$

So we have

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{\partial f}{\partial E} = q^2 \int dE\; g(E) \frac{\partial f}{\partial E}$$

where the last step is essentially the definition of the density of states $g(E)$. But at $T = 0$, the Fermi-Dirac distribution $f(E)$ is just a step function, and $\partial f/\partial E = -\delta(E - E_F)$. So at $T = 0$, we get

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g(E_F) \quad \Rightarrow \quad \epsilon(\mathbf{k}) = 1 + \frac{q^2 g(E_F)}{\epsilon_0 k^2}$$

which we recognise as the Thomas-Fermi result (7.70) that we derived previously.

The Lindhard Function

While the Thomas-Fermi approximation suffices for variations over large scales and small k, our real interest here is in what happens at large k. As we will now show, quantum mechanics gives rise to some interesting features in the screening when impurities have structure on scales of order $\sim 1/k_F$ where k_F is the Fermi-wavevector. For this, we need to go back to the Lindhard result

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{f(|\mathbf{k}+\mathbf{k}'|) - f(k')}{E(|\mathbf{k}+\mathbf{k}'|) - E(k')}$$

Our task is to do this integral properly.

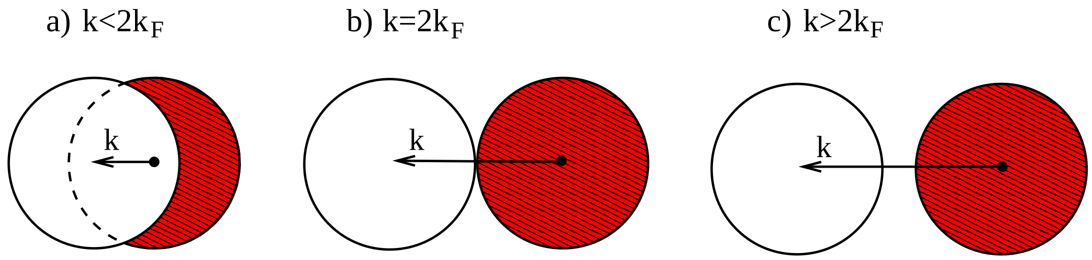


Figure 80: The two Fermi surfaces in momentum space. The integration region Σ is shown shaded in red for a) $k < 2k_F$, b) $k = 2k_F$ and c) $k > 2k_F$.

Let's firstly get a sense for what the integrand looks like. We'll work at $T = 0$, so the Fermi-Dirac distribution function $f(k)$ is a step function with

$$f(k) = \begin{cases} 1 & k < k_F \\ 0 & k > k_F \end{cases}$$

This makes the integral much easier. All the subtleties now come from figuring out which region in momentum space gives a non-vanishing contribution. The filled states associated to $f(k')$ form a ball in momentum space of radius k_F, centered at the origin. Meanwhile, the filled states associated to $f(|\mathbf{k}' + \mathbf{k}|)$ form a ball in momentum space of radius k_F centered at $\mathbf{k}' = -\mathbf{k}$. These are shown in a number of cases in Figure 80. Because the integral comes with a factor of $f(|\mathbf{k} + \mathbf{k}'|) - f(k')$, it gets contributions only from states that are empty in one ball but filled in the other. We call this region Σ; it is the shaded red region shown in the figures. There is a also a mirror region in the other ball that also contributes to the integral, but this simply gives an overall factor of 2. So we have

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = 2q^2 g_s \int_\Sigma \frac{d^3k'}{(2\pi)^3} \frac{1}{E(|\mathbf{k} + \mathbf{k}'|) - E(k')}$$

The important physics lies in the fact that the nature of Σ changes as we vary k. For $k < 2k_F$, Σ is a crescent-shaped region as shown in Figure 80a. But for $k \geq 2k_F$, Σ is the whole Fermi ball as shown in Figures 80b and 80c.

We'll work with non-relativistic fermions with $E = \hbar^2 k^2/2m$. While the graphical picture above will be useful to get intuition for the physics, to do the integral it's actually simpler to return to the form (7.71). At zero temperature, we have

$$\begin{aligned}\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} &= q^2 g_s \int_{k\le k_F} \frac{d^3k'}{(2\pi)^3}\left[\frac{1}{-k^2+2\mathbf{k}\cdot\mathbf{k}'}+\frac{1}{-k^2-2\mathbf{k}\cdot\mathbf{k}'}\right] \\ &= -q^2 g_s \frac{2m}{\hbar^2}\int_{k'\le k_F}\frac{d^3k'}{(2\pi)^3}\,\frac{2}{k^2-2\mathbf{k}'\cdot\mathbf{k}}\end{aligned}$$

where the two terms double-up because rotational symmetry ensures that the physics is invariant under $\mathbf{k}\to-\mathbf{k}$. Now the integration domain remains fixed as we vary $\mathbf{k}$, with the graphical change of topology that we saw above buried in the integrand. For $k\le 2k_F$, the denominator in the integrand can vanish. This reflects the fact that transitions between an occupied and unoccupied state with the same energy are possible. It corresponds to the situation depicted in Figure 80a. But for $k>2k_F$, the denominator is always positive. This corresponds to the situation shown in Figure 80c.

To proceed, we work in polar coordinates for $\mathbf{k}'$ with the z-axis aligned with $\mathbf{k}$. We have

$$\begin{aligned}\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} &= -\frac{4mq^2g_s}{(2\pi)^2\hbar^2}\int_0^\pi d\theta\,\sin\theta\int_0^{k_F}dk'\,\frac{k'^2}{k^2-2kk'\cos\theta} \\ &= \frac{2mq^2g_s}{(2\pi)^2\hbar^2}\frac{1}{k}\int_0^{k_F}dk'\,k'\log\left|\frac{k^2+2kk'}{k^2-2kk'}\right|\end{aligned}$$

But this is now an integral that we can do; the general form is

$$\int dy\;y\log\left(\frac{ay+b}{-ay+b}\right)=\frac{by}{a}+\frac{1}{2}\left(y^2-\frac{b^2}{a^2}\right)\log\left(\frac{ay+b}{-ay+b}\right)$$

We then have

$$\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} = -\frac{2mq^2g_s}{(2\pi)^2\hbar^2}\frac{1}{k}\left[\frac{kk_F}{2}+\frac{1}{2}\left(k_F^2-\frac{k^2}{4}\right)\log\left|\frac{2kk_F+k^2}{-2kk_F+k^2}\right|\right]$$

This gives our final expression, known as the *Lindhard dielectric function*,

$$\epsilon(k)=1+\frac{k_{TF}^2}{k^2}\,F\left(\frac{k}{2k_F}\right)$$

where all the constants that we gathered along our journey sit in $k_{TF}^2=q^2g(E_F)/\epsilon_0=g_sq^2mk_F/2\pi^2\hbar^2\epsilon_0$. This is the Thomas-Fermi wave result that we saw previously, but now it is dressed by the function

$$F(x)=\frac{1}{2}+\frac{1-x^2}{4x}\log\left|\frac{x+1}{x-1}\right|$$

At small k we have $F(x\to 0)=1$ and we recover the Thomas-Fermi result.

For variations on very small scales, we're interested in the large k regime where $x \to \infty$ and $F(x) \to 1/3x^2$. (You have to go to third order in the Taylor expansion of the log to see this!). This means that on small scales we have

$$\epsilon(k) \to 1 + \frac{4k_{TF}^2 k_F^2}{3k^4}$$

However, the most interesting physics occurs at near $k = 2k_F$.

7.7.5 Friedel Oscillations

We saw above that there's a qualitative difference in the accessible states when $k < 2k_F$ and $k > 2k_F$. Our goal is to understand what this means for the physics. The dielectric function itself is nice and continuous at $k = 2k_F$, with $F(x = 1) = 1/2$. However, it is not smooth: the derivative of the dielectric function suffers a logarithmic singularity,

$$F'(x \to 1^+) \to \frac{1}{2} \log\left(\frac{x-1}{2}\right)$$

This has an important consequence for the screening of a point charge.

As we saw in Section 7.7.2, a point charge gives rise to the external potential

$$\phi^{\rm ext}(\mathbf{k}) = \frac{Q}{\epsilon_0 k^2}$$

and, after screening, the true potential is $\phi(\mathbf{k}) = \phi^{\rm ext}(\mathbf{k})/\epsilon(\mathbf{k})$. However, the Fourier transform back to real space is now somewhat complicated. It turns out that it's easier to work directly with the induced charge density $\rho^{\rm ind}(\mathbf{r})$. From the definition of the dielectric function (7.63), the induced charge density in the presence of a point charge $\phi^{\rm ext}(\mathbf{k}) = Q/\epsilon_0 k^2$ is given by,

$$\rho^{\rm ind}(\mathbf{k}) = -Q\frac{\epsilon(\mathbf{k}) - 1}{\epsilon(\mathbf{k})}$$

where, for $k \approx 2k_F$, we have

$$\frac{\epsilon(\mathbf{k}) - 1}{\epsilon(\mathbf{k})} = \frac{k_{TF}^2}{8k_F^2}\left(1 + \frac{k - 2k_F}{2k_F}\log\left(\frac{k-2k_F}{4k_F}\right) + \ldots\right) \tag{7.73}$$

Now we want to Fourier transform this back to real space. We repeat the steps that we took in Section 7.7.2 for the Debye-Hückel model to get

$$\rho^{\rm ind}(\mathbf{r}) = -Q\int \frac{d^3k}{(2\pi)^3}\, e^{i\mathbf{k}\cdot\mathbf{r}} \left(\frac{\epsilon(\mathbf{k}) - 1}{\epsilon(\mathbf{k})}\right) = -\frac{Q}{2\pi^2}\frac{1}{r}\int_0^\infty dk\ \left(\frac{k\epsilon(\mathbf{k}) - k}{\epsilon(\mathbf{k})}\right)\sin kr$$

At this stage, it's useful if we integrate by parts twice. We have

$$\rho^{\rm ind}(\mathbf{r}) = \frac{Q}{2\pi^2}\frac{1}{r^3}\int_0^\infty dk\ \frac{d^2}{dk^2}\left(\frac{k\epsilon(\mathbf{k}) - k}{\epsilon(\mathbf{k})}\right)\sin kr$$

Of course, the Fourier integral requires us to know $\epsilon(\mathbf{k})$ at all values of k, rather than just around $k = 2k_F$. Suppose, however, that we're interested in the behaviour a long way from the point charge. At large r, the $\sin kr$ factor oscillates very rapidly with k, ensuring that the induced charge at large distances is essentially vanishing. This was responsible for the exponential behaviour of the screening that we saw in both the Debye-Hückel and Thomas-Fermi models. However, at $k = 2k_F$ the other factor in the integrand diverges,

$$\frac{d^2}{dk^2}\left(\frac{k\epsilon(\mathbf{k}) - k}{\epsilon(\mathbf{k})}\right) \approx \frac{k_{TF}^2}{4k_F^2}\frac{1}{k - 2k_F}$$

This will now give rise to a long-range contribution. Therefore, if we only care about this long-distance behaviour, we need only integrate over some small interval I about $k = 2k_F$,

$$\begin{aligned}\rho^{\rm ind}(\mathbf{r}) &\approx \frac{Qk_{TF}^2}{8\pi^2 k_F^2}\frac{1}{r^3}\int_I dk\ \frac{\sin kr}{k - 2k_F} \\ &= \frac{Qk_{TF}^2}{8\pi^2 k_F^2}\frac{1}{r^3}\int_I dk\ \left[\frac{\cos(2k_F r)\sin((q - 2k_F)r)}{k - 2k_F} + \frac{\sin(2k_F r)\cos((k - 2k_F)r)}{k - 2k_F}\right]\end{aligned}$$

where we've used a little trigonometry. The second term above vanishes on parity grounds (contributions from either side of $k = k_F$ cancel). We can approximate the first term by extending the range of the integral to all k (because, as we've just argued, the main contribution comes from the interval I anyway). Using $\int_{-\infty}^{+\infty} dx\ \sin x/x = \pi$, we get our final expression for the long-distance charge density induced by a point charge,

$$\rho^{\rm ind}(\mathbf{r}) \approx \frac{Qk_{TF}^2}{8\pi k_F^2}\frac{\cos(2k_F r)}{r^3} \tag{7.74}$$

We learn that the effect of the Fermi surface is to dramatically change the screening of electric charge. Instead of the usual exponential screening, we instead find a power-law fall off, albeit weaker than the Coulomb force in vacuum (i.e. $1/r^3$ instead of $1/r$). Moreover, the sign of the induced charge oscillates. These are called *Friedel oscillations*. They provide a very visual way to see the edge of the Fermi surface. This figure shows some Friedel oscillations on a two-dimensional surface[9]. You can make out a bright

[9]The figure is taken from *Direct Observation of Friedel Oscillations around Incorporated* $\mathrm{Si_{Ga}}$ *Dopants in GaAs by Low-Temperature Scanning Tunneling Microscopy* by M van der Wielen, A van Roij and H. van Kempen, Physical Review Letters 76, 7 (1996).

central region, surrounded by a black ring, surrounded by another white ring. This corresponds to a Fermi wavelength of around $\lambda_F \sim 10^{-8}\,m$.

Heuristically, what's going on here is that the wavefunction of the electrons has a finite size.. At zero temperature, the states with lowest energy have wavelength $\lambda = 1/k_F$. . These modes enthusiastically cluster around the impurity, keen to reduce its charge but, unaware of their own cumbersome nature, end up overscreening. Other electrons have to then respond to undo the damage and the story is then repeated, over exuberance piled upon over exuberance. The end result is a highly inefficient screening mechanism and the wonderful rippling patterns of charge that are seen in scanning tunnelling microscopes.

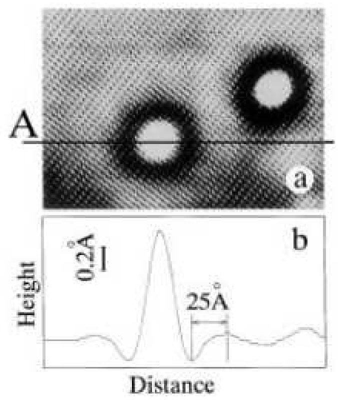


Figure 81: Friedel oscillations in GaAs doped with Silicon.