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.

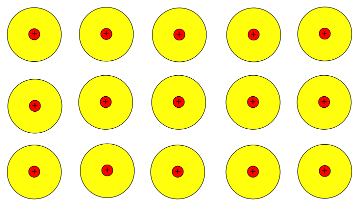


Figure 59: A simple model of a neutral material

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,

$$\mathbf{p} = \alpha \mathbf{E} \tag{7.1}$$

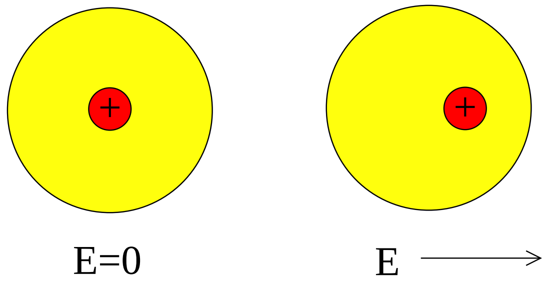


Figure 60: The polarisation of an atom

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.

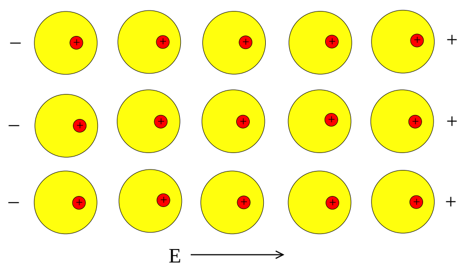


Figure 61: A polarised material

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

$$\begin{aligned}\nabla\cdot\mathbf{E} &= \frac{1}{\epsilon_0}(\rho_{\text{free}}+\rho_{\text{bound}})\\ &= \frac{1}{\epsilon_0}\left(\rho_{\text{free}}-\nabla\cdot\mathbf{P}\right)\end{aligned}$$

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

$$Q_{\text{bound}} = \frac{Q}{\epsilon_r} - Q = \frac{1-\epsilon_r}{\epsilon_r} Q = -\frac{\chi_e}{\epsilon_r} Q$$

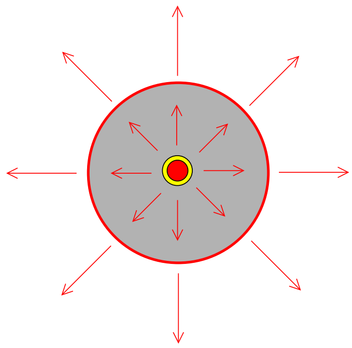


Figure 62: A polarised material

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:

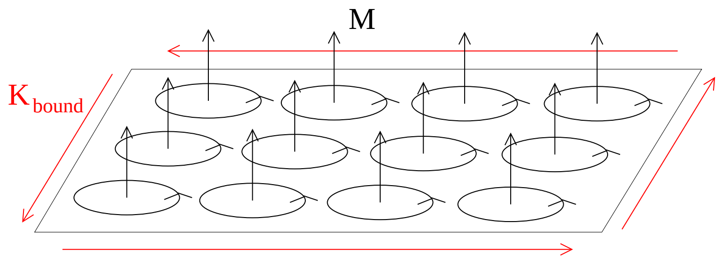
- *Ferromagnetism*: $\mathbf{M} \neq 0$ when $\mathbf{B} = 0$. Materials with this property are what you usually call "magnets". They're the things stuck to your fridge. The direction of $\mathbf{B}$ is from the south pole to the north. Only a few elements are ferromagnetic. The most familiar is Iron. 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

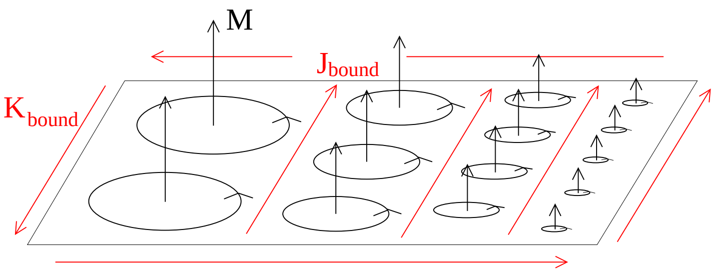
In the previous section, we saw that when a material is polarised, it results in bound charge. There is a similar story here. When a material becomes magnetised (at least in an anisotropic way), there will necessarily be regions in which there is a current. This is called the *bound current*.

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 $\mathbf{J}_{\text{bound}}$ that we've discussed above, and the current $\mathbf{J}_{\text{free}}$ from freely flowing electrons that we were implicitly talking. In Section 3, we were implicitly talking about $\mathbf{J}_{\text{free}}$ when we discussed currents. Ampère's law does not distinguish between these two currents; the magnetic field receives contributions from both.

$$\begin{aligned}\nabla \times \mathbf{B} &= \mu_0(\mathbf{J}_{\text{free}} + \mathbf{J}_{\text{bound}}) \\ &= \mu_0 \mathbf{J}_{\text{free}} + \mu_0 \nabla \times \mathbf{M}\end{aligned}$$

We define the *magnetising field*, $\mathbf{H}$ as

$$\mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M} \tag{7.11}$$

This obeys

$$\nabla \times \mathbf{H} = \mathbf{J}_{\text{free}} \tag{7.12}$$

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.

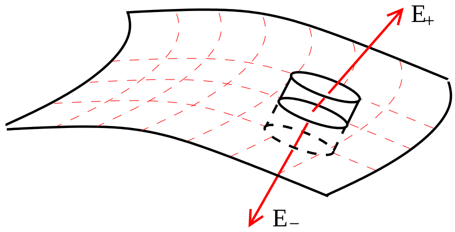


Figure 63: The normal component of the electric field is discontinuous

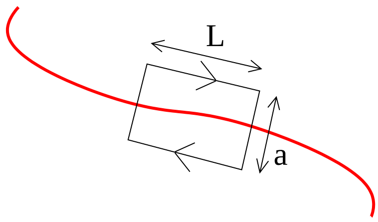


Figure 64: The tangential component of the electric field is continuous.

For example, consider the Gaussian pillbox shown in the left-hand figure above. Integrating the Maxwell equation $\nabla \cdot \mathbf{D} = \rho_{\text{free}}$ 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.

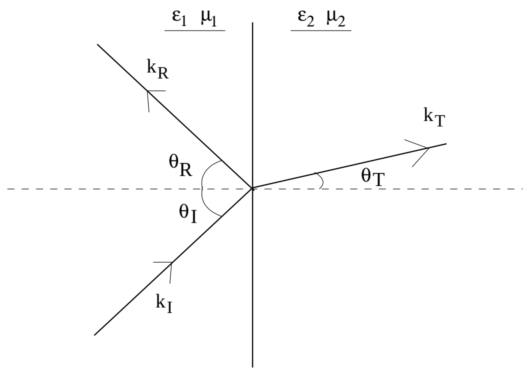


Figure 65: Incident, reflected and transmitted waves in a dielectric interface.

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,

$$\mathbf{E} = \begin{cases} \mathbf{E}_{\rm inc} + \mathbf{E}_{\rm ref} & x < 0 \\ \mathbf{E}_{\rm trans} & x > 0 \end{cases}$$

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*.

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).

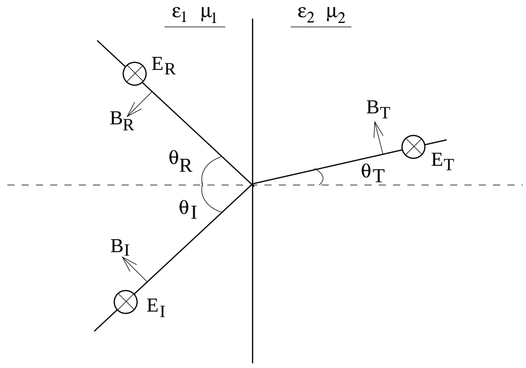


Figure 66: Incident, reflected and transmitted waves with normal polarisation.

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.

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).

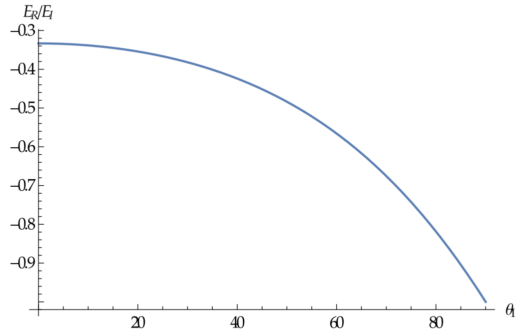


Figure 67: The reflected field with normal polarisation

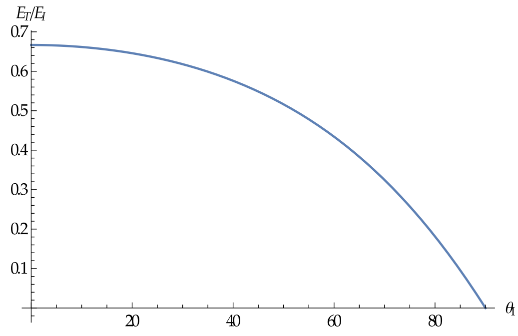


Figure 68: The transmitted field with normal polarisation

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.

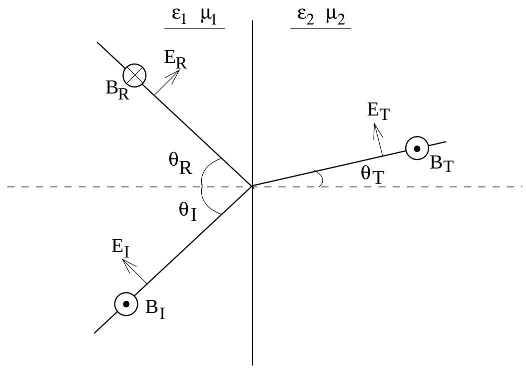


Figure 69: Incident, reflected and transmitted waves with parallel polarisation.

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

$$B_I - B_R = B_T \qquad \Rightarrow \qquad \frac{E_I - E_R}{v_1} = \frac{E_T}{v_2}$$

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 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 below.

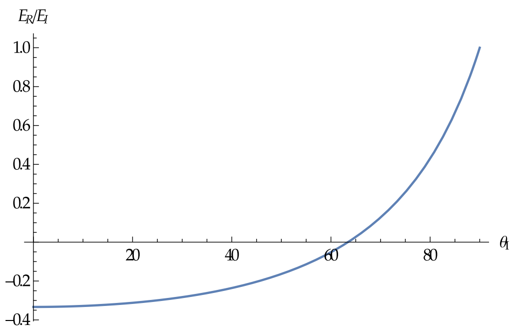


Figure 70: The reflected field with parallel polarisation

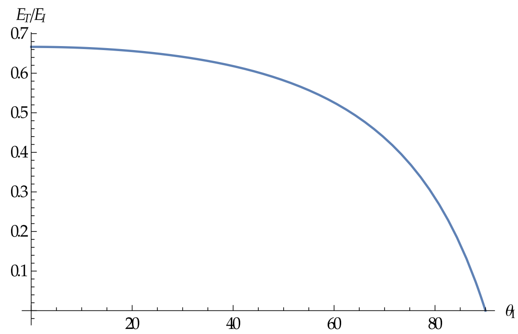


Figure 71: The transmitted field with parallel polarisation

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, $\theta_B \approx 56^\circ$.

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.

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

$$m\ddot{\mathbf{r}} = -\frac{q}{m}\mathbf{E}(t) - m\omega_0^2\mathbf{r} + m\gamma\dot{\mathbf{r}} \tag{7.28}$$

Solutions to this describe the atomic cloud oscillating about the nucleus.

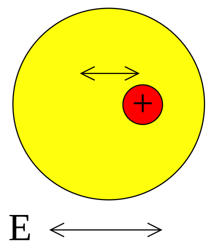


Figure 72:

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

$$\begin{aligned} \nabla\cdot\mathbf{D} = 0 \quad &\Rightarrow \quad \epsilon(\omega)\mathbf{k}\cdot\mathbf{E}(\omega) = 0 \\ \nabla\cdot\mathbf{B} = 0 \quad &\Rightarrow \quad \mathbf{k}\cdot\mathbf{B}(\omega) = 0 \end{aligned}$$

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

$$\begin{aligned} \nabla\times\mathbf{H} = \frac{\partial\mathbf{D}}{\partial t} \quad &\Rightarrow \quad \mathbf{k}\times\mathbf{B}(\omega) = -\mu\epsilon(\omega)\omega\mathbf{E}(\omega) \\ \nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \quad &\Rightarrow \quad \mathbf{k}\times\mathbf{E}(\omega) = \omega\mathbf{B}(\omega) \end{aligned} \tag{7.30}$$

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*.

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

$$\epsilon_1 = \epsilon_0 - \frac{nq^2}{m}\frac{\omega^2 - \omega_0^2}{(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2}$$

$$\epsilon_2 = \frac{nq^2}{m}\frac{\gamma\omega}{(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2}$$

These functions look like this: (These particular plots are made with $\gamma = 1$ and $\omega_0 = 2$ and $nq^2/m = 1$).

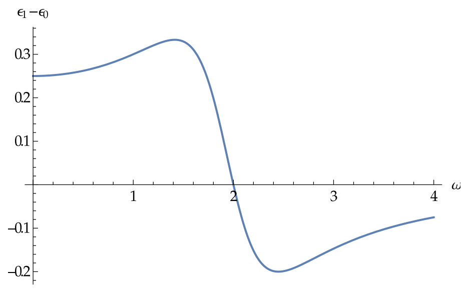


Figure 73: The real part of the permittivity, $\epsilon_1 - \epsilon_0$

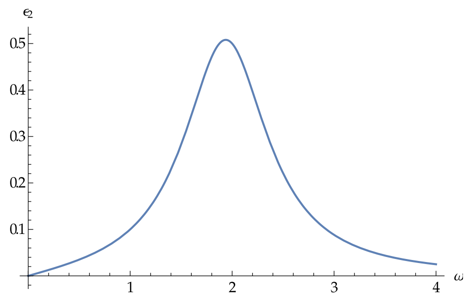


Figure 74: The imaginary part of the permittivity, ϵ_2

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

$$k_1 \approx k_2 \approx \pm\omega\sqrt{\frac{\mu\epsilon_2^2}{2}}$$

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.

Total Reflection:

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[4]. 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 $\mathbf{E}$).

[4] 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

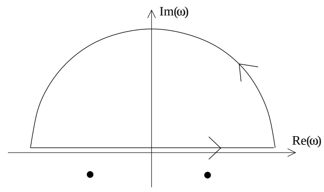


Figure 75:

$$\alpha(\omega) \text{ is analytic for } \operatorname{Im}\omega > 0$$

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 look like this:

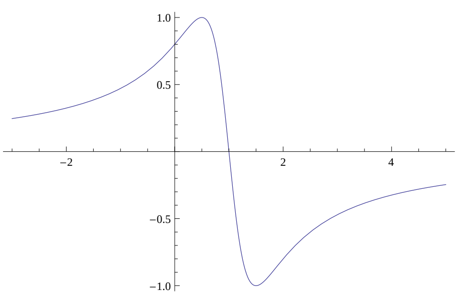


Figure 76: The real part of the function plotted with $\omega' = 1$ and $\epsilon = 0.5$.

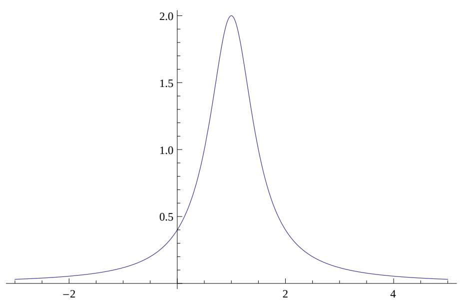


Figure 77: The imaginary part of the function plotted with $\omega' = 1$ and $\epsilon = 0.5$

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

$$\frac{1}{i\pi}\oint_C d\omega' \frac{\alpha(\omega')}{\omega' - \omega} \qquad \omega \in \mathbf{R} \tag{7.44}$$

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

$$\operatorname{Re}\alpha(\omega) = \mathcal{P}\int_{-\infty}^{+\infty} \frac{d\omega'}{\pi} \frac{\operatorname{Im}\alpha(\omega')}{\omega' - \omega}$$

and

$$\operatorname{Im}\alpha(\omega) = -\mathcal{P}\int_{-\infty}^{+\infty} \frac{d\omega'}{\pi} \frac{\operatorname{Re}\alpha(\omega')}{\omega' - \omega}$$

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 $\operatorname{Re}\alpha(\omega)$ for all frequencies in order to reconstruct $\operatorname{Im}\alpha(\omega)$ 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

$$\mathbf{J}(\omega) = \sigma(\omega)\mathbf{E}(\omega) \tag{7.46}$$

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 below for $\sigma_{DC} = 1$ and $\tau = 1$: The conductivity is complex simply

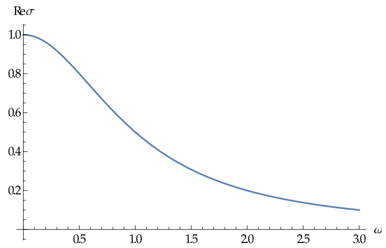


Figure 78: The real, dissipative part of the conductivity

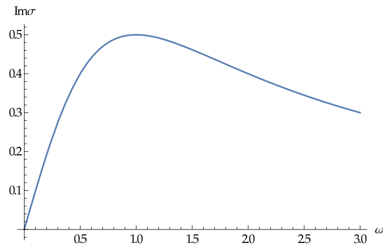


Figure 79: The imaginary, reactive part of the conductivity

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[5]. (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:

$$\begin{aligned} \nabla \cdot \mathbf{D} = \rho \quad &\text{and} \quad \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \\ \nabla \cdot \mathbf{B} = 0 \quad &\text{and} \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \end{aligned} \tag{7.48}$$

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?

[5]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 before, 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

$$\sigma(\omega) = \frac{\sigma_{\rm DC}}{1 - i\omega\tau}$$

Low Frequencies

At frequencies that are low compared to the scattering time, $\omega\tau \ll 1$, we have $\sigma(\omega) \approx \sigma_{\rm DC}$. This means that the real and imaginary parts of $\epsilon^{\rm eff}$ are

$$\epsilon^{\rm eff} = \epsilon_1^{\rm eff} + i\epsilon_2^{\rm eff} \approx \epsilon_1 + i\left(\epsilon_2 + \frac{\sigma_{\rm DC}}{\omega}\right) \tag{7.54}$$

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: $\epsilon^{\rm eff} \approx i\sigma_{\rm DC}/\omega$. The dispersion relation (7.53) then tells us that the wavenumber is

$$k = k_1 + ik_2 = \sqrt{i\mu\omega\sigma_{\rm DC}} = \sqrt{\frac{\mu\omega\sigma_{\rm DC}}{2}}\,(1+i)$$

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

$$\mathbf{E}(z,t) = \mathbf{E}(\omega)e^{-\delta z}\,e^{i(k_1 z - \omega t)}$$

where

$$\delta = \frac{1}{k_2} = \sqrt{\frac{2}{\mu\omega\sigma_{\rm DC}}}$$

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

$$\frac{|\mathbf{B}(\omega)|}{|\mathbf{E}(\omega)|} = \frac{k}{\omega} = \sqrt{\frac{\mu\sigma_{\rm DC}}{\omega}}$$

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 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 $\omega_p > \omega_0, \tau$. (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

$$\epsilon^{\rm eff}(\omega) \approx \epsilon_1 + i\frac{\sigma_{\rm DC}}{\omega}$$

Which can only vanish if we take the frequency to be purely imaginary,

$$\omega = -i\frac{\sigma_{\rm DC}}{\epsilon_1}$$

This is easy to interpret. Plugging it into the ansatz (7.49), we have

$$\mathbf{E}(\mathbf{x},t) = \mathbf{E}(\omega)\, e^{i\mathbf{k}\cdot\mathbf{x}}\, e^{-\sigma_{\rm DC}t/\epsilon_1}$$

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 $\sim \epsilon_1/\sigma_{\rm DC}$. 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 $\omega = \omega_p$:

$$\epsilon^{\rm eff}(\omega_p) \approx 0$$

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[6].

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.

[6] 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,

$$\rho^{\rm ind}(\mathbf{r}) = \rho_0\left(e^{-q\phi(\mathbf{r})/k_BT} - 1\right) \approx -\frac{q\rho_0\phi(\mathbf{r})}{k_BT} + \ldots \tag{7.64}$$

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,

$$\epsilon(\mathbf{k}) = 1 + \frac{k_D^2}{k^2} \tag{7.65}$$

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

$$\rho^{\rm ind}(\mathbf{r}) = -\frac{Qe^{-r/\lambda_D}}{4\pi\lambda_D^2 r}$$

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 $\rho^{\rm ext}(\mathbf{r})$.

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 $\rho^{\text{ext}}(\mathbf{r})$ in the system. The story is the same as we saw before: the mobile charges move, resulting in an induced charge density $\rho^{\text{ind}}(\mathbf{r})$, 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 $\mu \to \mu - q\phi(\mathbf{r})$.

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 $\rho^{\text{ind}}(\mathbf{r}) = \rho(\mathbf{r}) - \rho_0$. Combining (7.69) and (7.67), we have

$$\rho^{\text{ind}}(\mathbf{r}) = g_s \int \frac{d^3k}{(2\pi)^3} \left[\frac{q}{e^{(E_k + q\phi(\mathbf{r}) - \mu)/k_BT} + 1} - \frac{q}{e^{(E_k-\mu)/k_BT} + 1} \right]$$

But we can rewrite this using the notation of (7.67) simply as

$$\rho^{\text{ind}}(\mathbf{r}) = \rho\left(\mu - q\phi(\mathbf{r})\right) - \rho(\mu) \approx -q\phi(\mathbf{r}) \, \frac{\partial \rho(\mu)}{\partial \mu}$$

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[7]

$$g(E) = g_s \frac{1}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}$$

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

$$\rho_0 = q\int_0^{E_F} dE\ g(E)$$

[7]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

$$\frac{\rho^{\text{ind}}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{f(|\mathbf{k}+\mathbf{k}'|) - f(k')}{E(|\mathbf{k}+\mathbf{k}'|) - E(k')} \tag{7.72}$$

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 $\rho^{\text{ind}}(\mathbf{r})$, 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

$$\frac{\rho^{\text{ind}}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{\partial f}{\partial E} = q^2 \int dE\ g(E) \frac{\partial f}{\partial E}$$

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

$$\frac{\rho^{\text{ind}}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g(E_F) \quad \Rightarrow \quad \epsilon(\mathbf{k}) = 1 + \frac{q^2 g(E_F)}{\epsilon_0 k^2}$$

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

$$\frac{\rho^{\text{ind}}(\mathbf{k})}{\phi(\mathbf{k})} = q^2 g_s \int \frac{d^3k'}{(2\pi)^3} \frac{f(|\mathbf{k}+\mathbf{k}'|) - f(k')}{E(|\mathbf{k}+\mathbf{k}'|) - E(k')}$$

Our task is to do this integral properly.

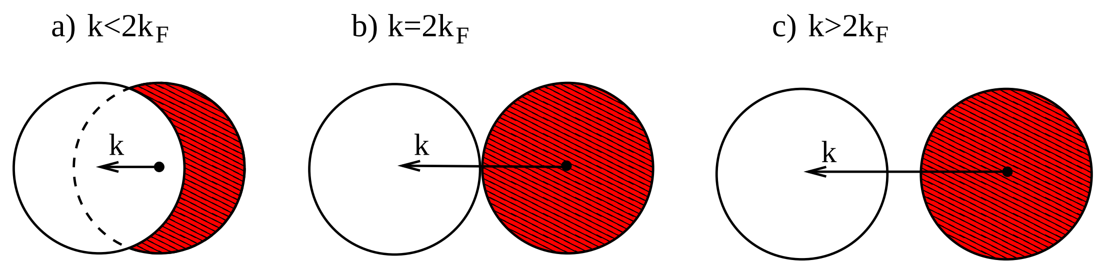


Figure 80: The two Fermi surfaces in momentum space. The integration region Σ is shown shaded in red for a) $k < 2k_F$, b) $k = 2k_F$ and c) $k > 2k_F$.

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 $E = \hbar^2k^2/2m$. 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

$$\begin{aligned}\frac{\rho^{\rm ind}(\mathbf{k})}{\phi(\mathbf{k})} &= q^2 g_s \int_{k\leq k_F} \frac{d^3k'}{(2\pi)^3}\left[\frac{1}{-k^2+2\mathbf{k}\cdot\mathbf{k}'} + \frac{1}{-k^2-2\mathbf{k}\cdot\mathbf{k}'}\right] \\ &= -q^2 g_s \frac{2m}{\hbar^2}\int_{k'\leq k_F} \frac{d^3k'}{(2\pi)^3}\,\frac{2}{k^2-2\mathbf{k}'\cdot\mathbf{k}}\end{aligned}$$

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 $k \leq 2k_F$, 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 $k_{TF}^2 = q^2 g(E_F)/\epsilon_0 = g_s q^2 m k_F/2\pi^2\hbar^2\epsilon_0$. 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

$$\phi^{\text{ext}}(\mathbf{k}) = \frac{Q}{\epsilon_0 k^2}$$

and, after screening, the true potential is $\phi(\mathbf{k}) = \phi^{\text{ext}}(\mathbf{k})/\epsilon(\mathbf{k})$. 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 $\rho^{\text{ind}}(\mathbf{r})$. From the definition of the dielectric function (7.63), the induced charge density in the presence of a point charge $\phi^{\text{ext}}(\mathbf{k}) = Q/\epsilon_0 k^2$ is given by,

$$\rho^{\text{ind}}(\mathbf{k}) = -Q \frac{\epsilon(\mathbf{k}) - 1}{\epsilon(\mathbf{k})}$$

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

$$\rho^{\text{ind}}(\mathbf{r}) = -Q \int \frac{d^3k}{(2\pi)^3}\, e^{i\mathbf{k}\cdot\mathbf{r}} \left(\frac{\epsilon(\mathbf{k}) - 1}{\epsilon(\mathbf{k})} \right) = -\frac{Q}{2\pi^2} \frac{1}{r} \int_0^\infty dk \, \left(\frac{k\epsilon(\mathbf{k}) - k}{\epsilon(\mathbf{k})} \right) \sin kr$$

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 $k = 2k_F$,

$$\begin{aligned}\rho^{\rm ind}(\mathbf{r}) &\approx \frac{Qk_{TF}^2}{8\pi^2k_F^2}\frac{1}{r^3}\int_I dk\ \frac{\sin kr}{k-2k_F}\\ &= \frac{Qk_{TF}^2}{8\pi^2k_F^2}\frac{1}{r^3}\int_I dk\ \left[\frac{\cos(2k_Fr)\sin((q-2k_F)r)}{k-2k_F}+\frac{\sin(2k_Fr)\cos((k-2k_F)r)}{k-2k_F}\right]\end{aligned}$$

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,

$$\rho^{\rm ind}(\mathbf{r}) \approx \frac{Qk_{TF}^2}{8\pi k_F^2}\frac{\cos(2k_Fr)}{r^3} \tag{7.74}$$

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[8]. You can make out a bright

[8]The figure is taken from *Direct Observation of Friedel Oscillations around Incorporated* $\mathrm{Si}_{\mathrm{Ga}}$ *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.

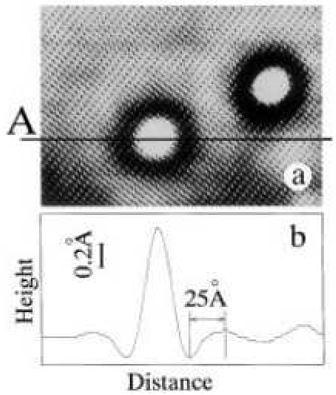


Figure 81: Friedel oscillations in GaAs doped with Silicon.