Electromagnetism

Willoughby Seago

September 22, 2020

These are my notes for the *electromagnetism* course from the University of Edinburgh as part of the third year of the theoretical physics degree. When I took this course in the 2020/21 academic year it was taught by Dr Andreas Hermann[1] and Dr Jamie Cole[2]. These notes are based on the lectures delivered as part of this course, the notes provided as part of this course, and the book 'introduction to electrodynamics'[3]. The content within is correct to the best of my knowledge but if you find a mistake or just disagree with something or think it could be improved please let me know.

These notes were produced using LaTeX[4]. Graphs where plotted using Matplotlib[5], NumPy[6], and SciPy[7]. Diagrams were drawn with tikz[8].

This is version 1.0 of these notes, which is up to date as of 23/12/2020.

Willoughby Seago

s1824487@ed.ac.uk

[1] https://www.ph.ed.ac.uk/people/andreas-hermann
[2] https://www.ph.ed.ac.uk/people/jamie-cole
[3] Griffiths, D. J. *Introduction to Electrodynamics*, fourth edition (Cambridge University Press, Cambridge, 2017)
[4] https://www.latex-project.org/
[5] https://matplotlib.org/
[6] https://numpy.org/
[7] https://scipy.org/scipylib/
[8] https://www.ctan.org/pkg/pgf

Contents

List of Figures

Acronyms

BC boundary conditions.
EM electromagnetism.
emf electromotive force.
HWP half-wave plate.
LE Laplace's equation.
LIH linear isotropic homogeneous.
PDE partial differential equation.
PE Poisson's equation.
QWP quarter-wave plate.

1 Vector Calculus

1.1 Gradient

The **gradient** of a scalar field, $f\colon \mathbb{R}^n \to \mathbb{R}$, is a function, $\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$, defined in three dimensions, in Cartesian coordinates, (x, y, z), by

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial x}\boldsymbol{e_x} + \frac{\partial f}{\partial y}\boldsymbol{e_y} + \frac{\partial f}{\partial z}\boldsymbol{e_z} = \partial_i f \boldsymbol{e_i}$$

with the Einstein summation convention applied in the last term. The gradient of a function can be thought of as a vector operator, ∇, acting on a scalar field, f, to give a vector field, ∇f. This vector field points in the direction of maximum increase of f and its magnitude is a measure of how fast f increases in this direction. Because of this the gradient is perpendicular to the level surfaces as travelling along a level surface means that by definition f is constant.

One important gradient to remember is

$$\nabla r = \hat{\boldsymbol{r}}$$

where, $\hat{\boldsymbol{r}}$, represents a unit vector in the direction of $\boldsymbol{r}$. Another important fact is that

$$\int_A^B \nabla f \cdot \mathrm{d}\boldsymbol{l} = f(\boldsymbol{r}_B) - f(\boldsymbol{r}_A)$$

where $\boldsymbol{r}_A$ and $\boldsymbol{r}_B$ are the points A and B respectively. This is independent of the path taken from A to B as ∇f is always a conservative field, we will see this again later.

The gradient looks different in different coordinates. In cylindrical coordinates, (ρ, φ, z), it is

$$\nabla f = \frac{\partial t}{\partial \rho}\boldsymbol{e_\rho} + \frac{1}{\rho}\frac{\partial f}{\partial \varphi}\boldsymbol{e_\varphi} + \frac{\partial f}{\partial z}\boldsymbol{e_z},$$

and in spherical coordinates, (r, ϑ, φ),[9] it is

$$\nabla f = \frac{\partial}{\partial f}\boldsymbol{e_r} + \frac{1}{r}\frac{\partial f}{\partial \vartheta}\hat{\boldsymbol{\vartheta}} + \frac{1}{r\sin\vartheta}\frac{\partial f}{\partial \varphi}\boldsymbol{e_\varphi}.$$

Importantly if $f = f(r)$ (i.e. there is no dependence on ϑ or φ) then

$$\nabla f = \frac{\partial f}{\partial r}\boldsymbol{e_r}.$$

This is consistent with the chain rule which is

$$\nabla f(r) = \frac{\mathrm{d}f}{\mathrm{d}r}\nabla r = \frac{\mathrm{d}f}{\mathrm{d}r}\hat{\boldsymbol{r}} = \frac{\mathrm{d}f}{\mathrm{d}r}\boldsymbol{e_r}.$$

Here f is a function of one variable so partial and total derivatives are equivalent. An important example of this chain rule is

$$\nabla\frac{1}{r} = -\frac{1}{r^2}\hat{\boldsymbol{r}} = -\frac{1}{r^3}\boldsymbol{r}.$$

1.2 Divergence and Gauss' Theorem

1.2.1 Divergence

The **divergence** of a vector field, $\boldsymbol{K}\colon \mathbb{R}^n \to \mathbb{R}^n$, is a function, $\nabla \cdot \boldsymbol{K}\colon \mathbb{R}^n \to \mathbb{R}$, defined in three dimensions, in Cartesian coordinates, (x, y, z), by

$$\operatorname{div} f = \nabla \cdot f = \frac{\partial K_x}{\partial x} + \frac{\partial K_y}{\partial y} + \frac{\partial K_z}{\partial z} = \partial_i K_i.$$

The divergence of a function can be thought of as the scalar product of a vector operator, ∇, and a vector field, $\boldsymbol{K}$, to give a scalar field, $\nabla \cdot \boldsymbol{K}$. The divergence gives a measure of how fast the flux lines of the vector field $\boldsymbol{K}$ converge towards a sink ($\nabla \cdot \boldsymbol{K} < 0$) or diverge away from a source ($\nabla \cdot \boldsymbol{K}) > 0$.

[9] We use the physics convention of ϑ as the polar angle from the z-axis and φ as the azimuthal angle from the x-axis.

One important divergence to remember is

$$\nabla \cdot \boldsymbol{r} = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} = 3.$$

The divergence looks different in different coordinates. In cylindrical coordinates, (ρ, φ, z), it is

$$\nabla \cdot \boldsymbol{K} = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho K_\rho) + \frac{1}{\rho}\frac{\partial K_\varphi}{\partial \varphi} + \frac{\partial K_z}{\partial z},$$

and in spherical coordinates, (r, ϑ, φ), it is

$$\nabla \cdot \boldsymbol{K} = \frac{1}{r^2}\frac{\partial}{\partial r}\big(r^2 K_r\big) + \frac{1}{r\sin\vartheta}\frac{\partial}{\partial \vartheta}(K_\vartheta \sin\vartheta) + \frac{1}{r\sin\vartheta} p dv K_\varphi \varphi.$$

1.2.2 Gauss' Theorem

Gauss' theorem,[10] also known as the **divergence theorem**, is

$$\int_V \nabla \cdot \boldsymbol{K}\,\mathrm{d}V = \oint_A \boldsymbol{K} \cdot \mathrm{d}\boldsymbol{S}$$

where A is a closed surface enclosing the volume V, $\mathrm{d}V = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \mathrm{d}^3 r$ is a volume element, $\mathrm{d}\boldsymbol{S} = \hat{\boldsymbol{n}}\,\mathrm{d}S$ is a surface element pointing out of the volume, and $\boldsymbol{K}$ is a vector field. Gauss' theorem holds for any such closed surface, A, and the volume it defines and any vector field, $\boldsymbol{K}$. This theorem relates a surface integral over a closed surface to an integral of the volume enclosed. This is very useful if we don't care about the nature of the field inside the volume and we only want to know its integral over that volume as it much easier to perform a surface integral.

1.3 Curl and Stokes' Theorem

1.3.1 Curl

The **curl** of a vector field, $\boldsymbol{K}\colon \mathbb{R}^3 \to \mathbb{R}^3$, is a function $\nabla \times \boldsymbol{K}\colon \mathbb{R}^3 \to \mathbb{R}^3$, defined in Cartesian coordinates, (x, y, z), by

$$\begin{aligned}
\operatorname{curl}\boldsymbol{K} = \nabla \times \boldsymbol{K} &= \left(\frac{\partial K_z}{\partial y} - \frac{\partial K_y}{\partial z}\right)\boldsymbol{e_x} + \left(\frac{\partial K_x}{\partial z} - \frac{\partial K_z}{\partial x}\right)\boldsymbol{e_y} + \left(\frac{\partial K_y}{\partial x} - \frac{\partial K_x}{\partial y}\right) \\
&= \begin{vmatrix} \boldsymbol{e_x} & \boldsymbol{e_y} & \boldsymbol{e_z} \\ \partial_x & \partial_y & \partial_z \\ K_x & K_y & K_z \end{vmatrix} \\
&= \varepsilon_{ijk}\partial_i K_j \boldsymbol{e_k}
\end{aligned}$$

where ε_{ijk} is the Levi-Civita symbol. The curl looks different in different coordinates. In cylindrical coordinates, (ρ, φ, z), it is

$$\nabla \times \boldsymbol{K} = \left(\frac{1}{\rho}\frac{\partial K_z}{\partial \varphi} - \frac{\partial K_\varphi}{\partial z}\right)\boldsymbol{e_\rho} + \left(\frac{\partial K_\rho}{\partial z} - \frac{\partial K_z}{\partial \rho}\right)\boldsymbol{e_\varphi} + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}(\rho K_\varphi) - \frac{\partial K_\rho}{\partial \varphi}\right)\boldsymbol{e_z},$$

and in spherical coordinates, (r, ϑ, φ), it is

$$\begin{aligned}
\nabla \times \boldsymbol{K} = \frac{1}{r\sin\vartheta}&\left(\frac{\partial}{\partial \vartheta}(K_\varphi \sin\vartheta) - \frac{\partial K_\vartheta}{\partial \varphi}\right)\boldsymbol{e_r} \\
&+ \frac{1}{r}\left(\frac{1}{\sin\vartheta}\frac{\partial K_r}{\partial \varphi} - \frac{\partial}{\partial r}(r K_\varphi)\right)\boldsymbol{e_\vartheta} \\
&+ \frac{1}{r}\left(\frac{\partial}{\partial r}(r K_\vartheta) - \frac{\partial K_r}{\partial \vartheta}\right)\boldsymbol{e_\varphi}.
\end{aligned}$$

[10] Not to be confused with the similar Gauss' law, see section 2.5

1.3.2 Stokes' Theorem

Stokes' theorem is

$$\int_A (\nabla \times \boldsymbol{K}) \cdot \mathrm{d}\boldsymbol{S} = \oint_C \boldsymbol{K} \cdot \mathrm{d}\boldsymbol{l}$$

where C is a closed curve bounding the surface A, $\mathrm{d}\boldsymbol{S} = \hat{\boldsymbol{n}}\mathrm{d}S$ is a surface element, $\mathrm{d}\boldsymbol{l}$ is a line element that is related to $\mathrm{d}\boldsymbol{S}$ by the right hand rule, and $\boldsymbol{K}$ is a vector field. Stokes' theorem holds for all closed curves, C, and any surface that it defines, as well as any vector field, $\boldsymbol{K}$. This theorem relates a line integral over a closed curve to the integral of a surface defined by the curve. This is very useful if we don't care about the nature of the field inside the surface and we only want to know its integral over that area as it much easier to perform a line integral.

1.4 Examples of Non-Vanishing Divergence and Curl

1.4.1 Non-Vanishing Divergence

The divergence is non-vanishing if the vector field grows along the propagation direction. Consider the vector field drawn in figure 1.1. We can apply Gauss' theorem to the dashed box. Along the top and

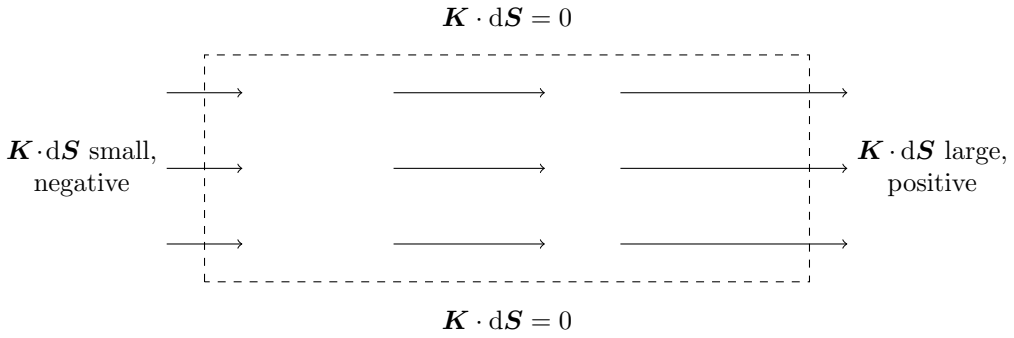


Figure 1.1: A vector field, $\boldsymbol{K}$, with non-vanishing divergence.

bottom the surface is parallel to the vector field so the surface normal is perpendicular to the vector field. This means that $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{S} = 0$. On the right hand side $\boldsymbol{K}$ is large and in the same direction as $\mathrm{d}\boldsymbol{S}$ so $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{S}$ is positive and large. On the left hand side $\boldsymbol{K}$ is small and in the opposite direction to $\mathrm{d}\boldsymbol{S}$ so $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{S}$ is negative and small in magnitude. The result is that after integrating over the whole surface the left hand side fails to cancel the right hand side and the result is positive. Therefore we have

$$\oint_S \boldsymbol{K} \cdot \mathrm{d}\boldsymbol{S} > 0.$$

This is true for all surfaces, S, that we can draw, specifically it is true if we take S to be arbitrarily small. Using Gauss' theorem this gives us

$$\int_V \nabla \cdot \boldsymbol{K}\, \mathrm{d}V > 0.$$

Since this holds for all surfaces, S, it must hold for all volumes, V, so we must have

$$\nabla \cdot \boldsymbol{K} > 0,$$

meaning the divergence is non-vanishing.

1.4.2 Non-Vanishing Curl

The curl is non-vanishing if the vector field grows perpendicular to the propagation direction. Consider the vector field drawn in figure 1.2. We can apply Stokes' theorem to the dashed box. Along the left and right hand side the curve is perpendicular to the vector field. This means that $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{l} = 0$. On the bottom $\boldsymbol{K}$ is large and in the same direction as $\mathrm{d}\boldsymbol{l}$ so $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{l}$ is positive and large. On the top $\boldsymbol{K}$ is small and in the opposite direction to $\mathrm{d}\boldsymbol{l}$ so $\boldsymbol{K} \cdot \mathrm{d}\boldsymbol{l}$ is negative and small in magnitude. The result is that after

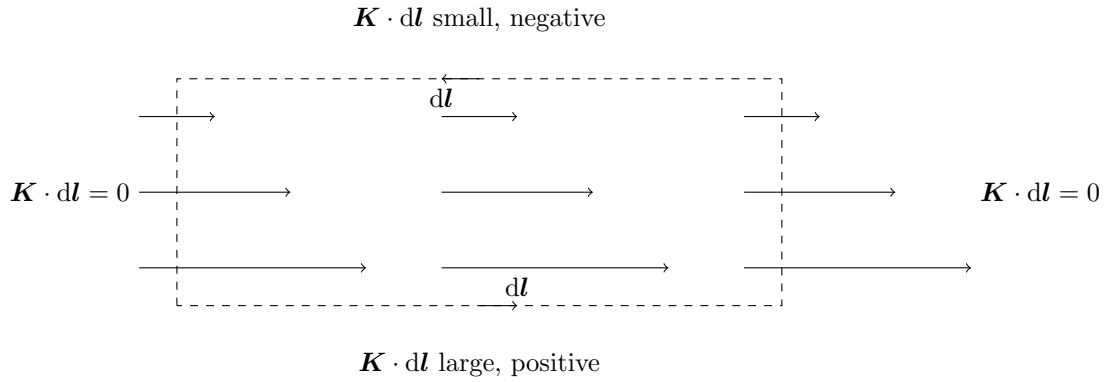


Figure 1.2: A vector field, $\boldsymbol{K}$, with non-vanishing curl.

integrating over the whole curve the top fails to cancel the bottom and the result is positive. Therefore we have

$$\oint_C \boldsymbol{K} \cdot \mathrm{d}\boldsymbol{l} > 0.$$

This is true for all curves, C, that we can draw, specifically it is true if we take C to be arbitrarily small. Using Stokes' theorem this gives us

$$\int_A (\nabla \times \boldsymbol{K}) \cdot \mathrm{d}\boldsymbol{S} > 0.$$

Since this holds for all curves, C, it must also hold for all surfaces, A, so we must have

$$\nabla \times \boldsymbol{K} > 0,$$

meaning the curl is non-vanishing.

1.5 Laplacian

The **Laplacian** of a scalar field, $f\colon \mathbb{R}^n \to \mathbb{R}$, is a function, $\nabla^2 f\colon \mathbb{R}^n \to \mathbb{R}$, defined in three dimensions, in Cartesian coordinates, (x, y, z), by

$$\nabla^2 f = \nabla \cdot (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial x^2} = \partial_i \partial_i f.$$

The gradient of a scalar function can be thought of as a scalar operator, $\nabla^2 = \nabla \cdot \nabla$, acting on a scalar field, f, to give a scalar field, $\nabla^2 f$. The Laplacian is a measure of the curvature of a field, in the same way that the second derivative of $f\colon \mathbb{R} \to \mathbb{R}$ gives a measure of curvature (i.e. it is zero if and only if f describes a straight line).

Being a scalar operator the Laplacian can also be applied to a vector field, $\boldsymbol{K}\colon \mathbb{R}^n \to \mathbb{R}^n$ in which case the Laplacian is a function, $\nabla^2 \boldsymbol{K}\colon \mathbb{R}^n \to \mathbb{R}^n$, defined in three dimensions, in Cartesian coordinates, (x, y, z), by:

$$\nabla^2 \boldsymbol{K} = \nabla^2 K_x \boldsymbol{e_x} + \nabla^2 K_y \boldsymbol{e_y} + \nabla^2 K_z \boldsymbol{e_z} = \nabla^2 K_i \boldsymbol{e_i},$$

where the Laplacians after the first equals are as defined above acting on a scalar field. The Laplacian looks different in different coordinates. In cylindrical coordinates, (ρ, φ, z), it is

$$\nabla^2 f = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 f}{\partial \varphi^2} + \frac{\partial^2 f}{\partial z^2}.$$

In Spherical coordinates, (r, ϑ, φ), it is

$$\begin{aligned} \nabla^2 f &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \vartheta} \frac{\partial}{\partial \vartheta} \left(\sin \vartheta \frac{\partial f}{\partial \vartheta} \right) + \frac{1}{r^2 \sin^2 \vartheta} \frac{\partial^2 f}{\partial \varphi^2} \\ &= \frac{1}{r} \frac{\partial^2}{\partial r^2} (rf) + \frac{1}{r^2 \sin \vartheta} \frac{\partial}{\partial \vartheta} \left(\sin \vartheta \frac{\partial f}{\partial \vartheta} \right) + \frac{1}{r^2 \sin^2 \vartheta} \frac{\partial^2 f}{\partial \varphi^2} \end{aligned}$$

1.6 Useful Vector Identities

Let $\varphi, \psi \colon \mathbb{R}^3 \to \mathbb{R}$, and $\boldsymbol{A}, \boldsymbol{B} \colon \mathbb{R}^3 \to \mathbb{R}^3$ then

1. $\nabla(\varphi\psi) = \varphi\nabla\psi + (\nabla\varphi)\psi$
2. $\nabla \cdot (\varphi\boldsymbol{A}) = \varphi\nabla \cdot \boldsymbol{A} + (\nabla\varphi) \cdot \boldsymbol{A}$
3. $\nabla \times (\varphi\boldsymbol{A}) = \varphi(\nabla \times \boldsymbol{A}) + (\nabla\varphi) \times \boldsymbol{A}$
4. $\nabla(\boldsymbol{A} \cdot \boldsymbol{B}) = (\boldsymbol{A} \cdot \nabla)\boldsymbol{B} + (\boldsymbol{B} \cdot \nabla)\boldsymbol{A} + \boldsymbol{A} \times (\nabla \times \boldsymbol{B}) + \boldsymbol{B} \times (\nabla \times \boldsymbol{A})$
5. $\nabla \cdot (\boldsymbol{A} \times \boldsymbol{B}) = \boldsymbol{B} \cdot (\nabla \times \boldsymbol{A}) - \boldsymbol{A} \cdot (\nabla \times \boldsymbol{B})$
6. $\nabla \times (\boldsymbol{A} \times \boldsymbol{B}) = \boldsymbol{A}(\nabla \cdot \boldsymbol{B}) - \boldsymbol{B}(\nabla \cdot \boldsymbol{A}) + (\boldsymbol{B} \cdot \nabla)\boldsymbol{A} - (\boldsymbol{A} \cdot \nabla)\boldsymbol{B}$
7. $\nabla \times (\nabla f) = 0$
8. $\nabla \cdot (\nabla \times \boldsymbol{A}) = 0$
9. $\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2\boldsymbol{A}$

1.7 Taylor Expansions in 3 Dimensions

Let $f \colon \mathbb{R}^n \to \mathbb{R}$. A small change, $\mathrm{d}\boldsymbol{r}$, in $\boldsymbol{r}$ causes a change $\mathrm{d}f = \nabla f \cdot \mathrm{d}\boldsymbol{r}$ in f. This is the first term of a 3-dimensional Taylor expansion of f about the point $\boldsymbol{r'}$:

$$\begin{aligned} f(\boldsymbol{r}) &= \sum_{n=0}^{\infty} \frac{1}{n!}[(\boldsymbol{r} - \boldsymbol{r'}) \cdot \nabla]^n f(\boldsymbol{r})|_{\boldsymbol{r}=\boldsymbol{r'}} \\ &= f(\boldsymbol{r'}) + \sum_{i=1}^{3}(x_i - x_i')\frac{\partial f(\boldsymbol{r})}{\partial x_i}\bigg|_{\boldsymbol{r}=\boldsymbol{r'}} + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}(x_i - x_i')(x_j - x_j')\frac{\partial^2}{\partial f(\boldsymbol{r})\partial x_j}\bigg|_{x_i} \boldsymbol{r} = \boldsymbol{r'} + \cdots \end{aligned}$$

1.8 An Important Theorem

The following three statements concerning a vector field, $\boldsymbol{F} \colon V \subseteq \mathbb{R}^3 \to \mathbb{R}^3$, over some region in space, V, are equivalent:

1. $\nabla \times \boldsymbol{F} = 0$ – The vector field, $\boldsymbol{F}$, is irrotational.
2. $\boldsymbol{F} = \nabla\varphi$ for some $\varphi \colon \mathbb{R}^3 \to \mathbb{R}$ – The vector field, $\boldsymbol{F}$, is a gradient field of a scalar potential, φ.
3. The line integral
$$\int_A^B \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l}$$
is independent of the path from A to B for all $A, B \in V$. A consequence of this is that for any closed curve, C, in V we have
$$\oint_C \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l} = 0.$$

1.9 The Dirac Delta Distribution

The **Dirac Delta distribution** is a generalised function, δ, with two defining properties:

$$\delta(x - x_0) = \begin{cases} 0, & x \neq x_0, \\ \infty, & x = x_0, \end{cases}$$

and

$$\int_{-\infty}^{\infty} \delta(x - x_0)\ \mathrm{d}x = 1.$$

We can actually be more specific with this last property:

$$\int_a^b \delta(x - x_0)\ \mathrm{d}x = 1$$

if and only if $x_0 \in [a, b]$.

We can define an analogous distribution in 3 dimensions. We use the same symbol, δ, and it has the expected analogous properties:

$$\delta(\boldsymbol{r} - \boldsymbol{r_0}) = \begin{cases} 0, & \boldsymbol{r} \neq \boldsymbol{r_0}, \\ \infty, & \boldsymbol{r} = \boldsymbol{r_0}, \end{cases}$$

and

$$\int_{\mathbb{R}^3} \delta(\boldsymbol{r} - \boldsymbol{r_0}) \, \mathrm{d}V = 1$$

or more specifically for some volume $V \subseteq \mathbb{R}^3$

$$\int_V \delta(\boldsymbol{r} - \boldsymbol{r_0}) \, \mathrm{d}V = 1$$

if and only if $\boldsymbol{r_0} \in V$.

We can view the 3-dimensional delta distribution as a product of three 1-dimensional delta distributions:

$$\delta(\boldsymbol{r} - \boldsymbol{r_0}) = \delta(x - x_0)\delta(y - y_0)\delta(z - z_0)$$

where $\boldsymbol{r_0} = (x_0, y_0, z_0)$.

One way that we can view the delta distribution is as a limit of a sequence of functions, (f_ε), where

$$f_\varepsilon(x) = \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp\left(-\frac{1}{2}\left(\frac{x - x_0}{\varepsilon}\right)^2\right).$$

Here f_ε is a normal distribution centred at x_0 with a standard deviation (width) of ε. If we then take the limit as $\varepsilon \to 0$ we get

$$\delta(x - x_0) = \lim_{\varepsilon \to 0} f_\varepsilon(x) = \lim_{\varepsilon \to 0} \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp\left(-\frac{1}{2}\left(\frac{x - x_0}{\varepsilon}\right)^2\right).$$

Some useful properties of the delta distribution are:

- For $g\colon \mathbb{R} \to \mathbb{R}$ we have
$$\int_{-\infty}^{\infty} \delta(x - x_0) g(x) \, \mathrm{d}x = g(x_0).$$
- For $g\colon \mathbb{R} \to \mathbb{R}$ we have
$$\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x}\delta(x - x_0) g(x) \, \mathrm{d}x = -\left.\frac{\mathrm{d}g}{\mathrm{d}x}\right|_{x=x_0}.$$
- $\delta(x - x_0) = \delta(x_0 - x)$.

The properties with integrals actually hold so long as the point x_0 is between the limits of the integral, for example

$$\int_{x_0-\varepsilon}^{x_0+\varepsilon} \delta(x - x_0) g(x) \, \mathrm{d}x = g(x_0)$$

for $\varepsilon > 0$.

Part I

Electrostatics

2 Electrostatics Revision

2.1 Electric Charge

Charge is a discrete property of elementary particles. Charge is discretised into amounts of $e/3$ where $e = 1.602\,\mathrm{C}$ is the magnitude of the charge of an electron. As far as we know only quarks can have

this smallest possible amount of charge, for example a down quark has a charge of $q_{\mathrm{d}} = -e/3$. In most applications however we think of charge as being carried by electrons, which have a charge of $q_{\mathrm{e}^-} = -e$. The charge of the proton is, as far as we can tell, exactly the opposite of an electron, that is $q_{\mathrm{p}} = e$. This has been experimentally verified so we know that $q_{\mathrm{p}} + q_{\mathrm{e}^-} < 10^{-21}e$. Similarly the charge of an antimatter particle is exactly opposite that of the relevant matter particle. For example we have experimentally verified that $q_{\mathrm{p}} + q_{\bar{\mathrm{p}}} < 10^{-8}e$. This means that a vacuum has no charge.

We are interested in classical electromagnetism (EM) in which we deal with macroscopic charge distributions. Since e is so small we approximate charge as a continuous variable. We define a **charge density**, $\rho\colon \mathbb{R}^3 \to \mathbb{R}$, that takes a point in space, $\boldsymbol{r}$, and returns the charge, $\rho(\boldsymbol{r})\,\mathrm{d}V$, of the infinitesimal volume, $\mathrm{d}V$ around $\boldsymbol{r}$. This is sometimes referred to as a charge element. $\rho(\boldsymbol{r})$ has units of $\mathrm{C\,m^{-3}}$. Assuming that all charge is due to the presence of protons and neutrons we define the **number densities**, $n_{\mathrm{p}}, n_{\mathrm{e}^-}\colon \mathbb{R}^3 \to \mathbb{R}$, as functions that give the number of protons and electrons, $n_{\mathrm{p}}(\boldsymbol{r})\,\mathrm{d}V$ and $n_{\mathrm{e}^-}(\boldsymbol{r})\,\mathrm{d}V$, respectively in a small volume $\mathrm{d}V$ about the point $\boldsymbol{r}$. Using these we can write the charge density as

$$\rho(\boldsymbol{r}) = [n_{\mathrm{p}}(\boldsymbol{r}) - n_{\mathrm{e}^-}(\boldsymbol{r})]e.$$

The total charge enclosed in a volume V is given by

$$Q_V = \int_V \rho(\boldsymbol{r})\ \mathrm{d}V\,.$$

When working with lower dimension objects such as surfaces and lines we use lower dimensional analogues of the charge density ρ. A charged surface has a charge density of $\sigma\colon \mathbb{R}^2 \to \mathbb{R}$, which has units of $\mathrm{C\,m^{-2}}$, and gives us a charge element of $\sigma(x, y)\,\mathrm{d}S$. A charged curve has a charge density of $\lambda\colon \mathbb{R} \to \mathbb{R}$, which has units of $\mathrm{C\,m^{-1}}$, and gives us a charge element of $\lambda(x)\,\mathrm{d}l$. The total charge of an area A and curve L with these two charge densities are given by

$$Q_A = \int_A \sigma\ \mathrm{d}S\,, \qquad \text{and} \qquad Q_L = \int_L \lambda\ \mathrm{d}l$$

respectively.

2.2 Point Charges and the Delta Distribution

In electrostatics it is common to introduce a **point charge**, Q, at a point $\boldsymbol{r'}$. These are charge distributions that have a charge density of

$$\rho(\boldsymbol{r}) = Q\delta(\boldsymbol{r} - \boldsymbol{r'})$$

where δ is the Dirac delta distribution as defined in section 1.9. What this means is that the charge is zero everywhere apart from where the point charge is where the charge is Q.

One important property of the Dirac delta distribution is what happens when we have a sum of two delta distributions. If $g\colon \mathbb{R} \to \mathbb{R}$ is a sufficiently smooth function and $x_1, x_2 \in \mathbb{R}$ then we have

$$\begin{aligned}\int g(x)[\delta(x - x_1) + \delta(x - x_2)]\ \mathrm{d}x &= \int g(x)\delta(x - x_1)\ \mathrm{d}x + \int g(x)\delta(x - x_2)\ \mathrm{d}x \\ &= g(x_1) + g(x_2)\end{aligned}$$

where we have employed the sifting property of the delta distribution:

$$\int g(x)\delta(x - x')\ \mathrm{d}x = g(x').$$

This summing property of delta distributions generalises to any number of summands, $\delta(x - x_i)$. It also generalises to any number of dimensions. The reason that the sum of two delta distributions is important is it allows us to have an arbitrary number of point charges. Say we have N point charges, q_i, at positions $\boldsymbol{r_i}$, where $i = 1, \ldots, N$, then the charge density of this system is

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{N} q_i\delta(\boldsymbol{r} - \boldsymbol{r_i}).$$

We can see that this gives us the correct result for total charge. Assuming that all of these point charges lie in some volume, V, the total charge is

$$\begin{aligned} Q_V &= \int_V \rho(\boldsymbol{r})\ \mathrm{d}^3 r \\ &= \int_V \sum_{i=1}^{N} q_i \delta(\boldsymbol{r} - \boldsymbol{r_i})\ \mathrm{d}^3 r \\ &= \sum_{i=1}^{N} q_i \int_V \delta(\boldsymbol{r} - \boldsymbol{r_i})\ \mathrm{d}^3 r \\ &= \sum_{i=1}^{N} q_i \end{aligned}$$

So the total charge is just the sum of all the point charges. This is exactly what we would expect. The ability to add point charges and charge densities like this is called **superposition**.

2.3 Coulomb's Law

Let q_1 and q_2 be point charges at points $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ respectively. Empirically we know that the force exerted on q_1 due to q_2 is given by **Coulomb's Law**:

$$\boldsymbol{F_1} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{\boldsymbol{r}}_{12} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^3}\boldsymbol{r}_{12}$$

where $\boldsymbol{r}_{12} = \boldsymbol{r}_1 - \boldsymbol{r}_2$. Here $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{C\,N^{-1}\,m^{-2}}$ is the **permittivity of free space**, also called the **electric constant**. A handy number to remember is

$$\frac{1}{4\pi\varepsilon_0} = 8.988 \times 10^9\,\mathrm{N\,m^2\,C^{-1}} \approx 9 \times 10^9\,\mathrm{N\,m^2\,C^{-1}}$$

The $1/r^2$ dependence of this law has been verified up to 10^{-6}.

We can use the superposition principle to write a continuous charge density as a sum of point charges and then apply Coulomb's law to each. This gives us the total force on a charge, q, at the point $\boldsymbol{r}$ due to a charge density ρ:

$$\boldsymbol{F}(\boldsymbol{r}) = \frac{q}{4\pi\varepsilon_0}\int \frac{\boldsymbol{r} - \boldsymbol{r'}}{|\boldsymbol{r} - \boldsymbol{r'}|^3}\rho(\boldsymbol{r'})\ \mathrm{d}^3 r'\,.$$

2.4 Electric Field

If we have a positive test charge, q, then we can factorise the Coulomb force on this charge into $\boldsymbol{F} = q\boldsymbol{E}$. This defines the **electric field**, $\boldsymbol{E}\colon \mathbb{R}^3 \to \mathbb{R}^3$. $\boldsymbol{E}(\boldsymbol{r})$ has units of $\mathrm{N\,C^{-1}}$ and gives the force per unit charge experienced by a positive test charge, q, at the point $\boldsymbol{r}$.

For a point charge the electric field is radially away from the charge if the charge is positive and towards the charge if it is negative. More explicitly using Coulomb's law we see that the electric field for a point charge, q at $\boldsymbol{r'}$ is

$$\boldsymbol{E} = \frac{q}{4\pi\varepsilon}\frac{\boldsymbol{r} - \boldsymbol{r'}}{|\boldsymbol{r} - \boldsymbol{r'}|^3}.$$

2.5 Gauss' Law

Gauss' law for electric fields states that for a closed surface A if the electric field is $\boldsymbol{E}$ then the total charge, Q_{enc}, enclosed by A is given by

$$\oint_A \boldsymbol{E}\cdot \mathrm{d}\boldsymbol{S} = \frac{Q_{\mathrm{enc}}}{\varepsilon_0}.$$

Here $\mathrm{d}\boldsymbol{S}$ is a surface element that points out of the volume enclosed by A.

We can think of $\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}$ as the **electric flux density** through A at $\boldsymbol{r}$. We define the total **electric flux**, Φ_E, to be the integral of the electric flux density over the whole surface:

$$\Phi_E = \int_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}\,.$$

We see that for a closed surface, A, the electric flux is exactly given by

$$\Phi_E = \frac{Q_{\text{enc}}}{\varepsilon_0}.$$

We can show that Gauss law holds for a point charge, q, at $\boldsymbol{r}$. We start by applying the divergence theorem to get

$$\begin{aligned} I &= \int_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} \\ &= \int_V \nabla \cdot \boldsymbol{E}\ \mathrm{d}V \\ &= \int_V \nabla \cdot \left[\frac{q}{4\pi\varepsilon_0} \frac{\boldsymbol{r} - \boldsymbol{r'}}{|\boldsymbol{r} - \boldsymbol{r'}|^3} \right] \mathrm{d}^3 r \\ &= \frac{q}{4\pi\varepsilon_0} \int_V \nabla \cdot \left[\frac{\boldsymbol{r} - \boldsymbol{r'}}{|\boldsymbol{r} - \boldsymbol{r'}|^3} \right] \mathrm{d}^3 r \end{aligned}$$

In the first tutorial we showed that

$$\nabla \cdot \left(\frac{\boldsymbol{r}}{r^3} \right) = 4\pi\delta(\boldsymbol{r}).$$

We can use this here as $\nabla\cdot$ acts only on $\boldsymbol{r}$, not on $\boldsymbol{r'}$. This gives us

$$\begin{aligned} I &= \frac{q}{4\pi\varepsilon_0} \int_V 4\pi\delta(\boldsymbol{r} - \boldsymbol{r'})\ \mathrm{d}^3 r \\ &= \frac{q}{4\pi\varepsilon_0} 4\pi \\ &= \frac{q}{\varepsilon_0} \end{aligned}$$

where we assume that the point charge is in the volume over which we are integrating. If it isn't then $I = 0$ instead. Either way we get that I is given by the charge enclosed divided by ε_0. Thus we have shown that Gauss' law holds for a point charge.

We can then use the superposition property to write a continuous charge distribution, $\rho(\boldsymbol{r})$, as a sum of N point charges, q_i. If each charge contributes an electric field of $\boldsymbol{E_i}$ then the total electric field is

$$\boldsymbol{E} = \sum_{i=1}^{N} \boldsymbol{E_i}.$$

This comes from the fact that the total force on a test charge is

$$\boldsymbol{F} = \sum_{i=1}^{N} \boldsymbol{F_i}$$

where $\boldsymbol{F_i}$ is the force due to the point charge q_i. The total enclosed charge is then given by

$$\begin{aligned} \frac{Q_{\text{enc}}}{\varepsilon_0} &= \frac{1}{\varepsilon_0} \sum_{i=1}^{N} q_i \\ &= \sum_{i=1}^{N} \frac{q_i}{\varepsilon_0} \\ &= \sum_{i=1}^{N} \oint_A \boldsymbol{E_i} \cdot \mathrm{d}\boldsymbol{S} \end{aligned}$$

$$= \oint_A \sum_{i=1}^{N} \boldsymbol{E_i} \cdot \mathrm{d}\boldsymbol{S}$$
$$= \oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}\,.$$

So Gauss' law holds for any charge distribution.

2.6 Electrostatic Potential

It is trivial to show that

$$\frac{\hat{\boldsymbol{r}}}{r^2} = \frac{\boldsymbol{r}}{r^3} = -\nabla\left(\frac{1}{r}\right).$$

Hence for a point charge, q, at $\boldsymbol{r'}$ the electric field is given by

$$\boldsymbol{E}(\boldsymbol{r}) = -\nabla\left[\frac{q}{r\pi\varepsilon_0}\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right]$$

where we have again used the fact that ∇ acts on $\boldsymbol{r}$ and not $\boldsymbol{r'}$. We define the **electrostatic potential**, $V\colon \mathbb{R}^3 \to \mathbb{R}$, for a point charge to be

$$V(\boldsymbol{r}) = \frac{1}{r\pi\varepsilon_0}\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}$$

such that

$$\boldsymbol{E} = -\nabla V.$$

By the superposition property we can define the electrostatic potential for any charge density, ρ, to be

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\,\mathrm{d}^3r'\,.$$

So again we have $\boldsymbol{E} = -\nabla V$. The force is then given by

$$\boldsymbol{F} = q\boldsymbol{E} = -q\nabla V.$$

Also we have

$$\nabla \times \boldsymbol{E} = -\nabla \times (\nabla V) = \boldsymbol{0}$$

since the curl of a gradient is zero (see section 1.6). This is one of Maxwell's laws. We also have that

$$\int_A^B \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l}$$

is path independent (see section 1.8). Note that this only holds for static electric fields. For a general electric field we will see later that $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$.

3 Applications of Gauss' Law

3.1 Conductors and Insulators

Conductors and insulators are idealised materials. By this we mean that the properties we are about to list are exact in theory but are really only an approximation of reality.

3.1.1 Conductors

A **conductor** is a material in which charges can move freely. Charges can be separated (e.g. electrons removed from atoms) at an arbitrary rate, velocity, and magnitude (i.e. you won't run out of charges to separate and you can separate them instantaneously).

One important consequence of this is that an external field, $\boldsymbol{E}$, will lead to charges rearranging until an equilibrium is reached. This means an internal electric field, $\boldsymbol{E}_{\mathbf{int}}$, is induced and as an equilibrium is reached we must have $\boldsymbol{E} + \boldsymbol{E}_{\mathbf{int}} = \boldsymbol{0}$. This means that inside a conductor the charge density is given by $\rho = 0$ and the potential, V, is constant as $\nabla \cdot V = 0$ if V is constant, often it makes sense to set $V = 0$

as a potential is only defined relative to some other place. The result is that all free charge ends up on the surface so we have a surface charge density, σ, instead of a volume charge density.

Just outside of a conductor the electric field must be normal to the surface. Suppose that it wasn't. Then there would be a component that went along the surface meaning that there would be movement of charges along the surface meaning that the system wouldn't be at equilibrium so this can't happen. We will show later that $E = \sigma/\varepsilon_0$ just outside the surface.

3.1.2 Insulator

An **insulator** is a material in which there is no motion of charges. The charge density, ρ, can have any form and the potential, V, is generally non-uniform meaning that in general $\boldsymbol{E}(\boldsymbol{r}) \neq 0$.

3.2 Gauss' Law in Differential Form

We start from Gauss' law as we defined it previously, as well as the definition of the charge density over the volume V defined by the closed surface A:

$$\oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \frac{Q_{\text{enc}}}{\varepsilon_0} = \int_V \frac{\rho(\boldsymbol{r})}{\varepsilon_0}\, \mathrm{d}V\,.$$

We then apply the divergence theorem to get

$$\oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \int_V \nabla \cdot \boldsymbol{E}\, \mathrm{d}V = \int_V \frac{\rho(\boldsymbol{r})}{\varepsilon_0}\, \mathrm{d}V\,.$$

This holds for all volumes V and therefore the integrands of the two integrals must be identical, that is

$$\nabla \cdot \boldsymbol{E} = \frac{\rho(\boldsymbol{r})}{\varepsilon_0}.$$

This is Gauss' law in differential form. It forms the first of Maxwell's laws. Another of Maxwell's laws that we have used before is that for a static electric field $\nabla \times \boldsymbol{E} = \boldsymbol{0}$.

3.3 Using Gauss' Law

Gauss' law gives a very quick method of finding the electric field so long as we are in a situation where the symmetry of the charge distribution allows us to choose a surface, A, over which the integral becomes trivial.

There are typically three symmetries that we look for:

- Spherical symmetry – Choose a Gaussian surface of concentric spheres.
- Cylindrical symmetry – Choose a Gaussian surface of coaxial cylinders.
- Planar symmetry – Choose a 'pillbox' shaped Gaussian surface.

3.4 Spherical Symmetry

3.4.1 Insulating Sphere

An insulating sphere of radius a has a uniform charge distribution ρ. We argue that by the symmetry of the situation the electric field must be

$$\boldsymbol{E}(\boldsymbol{r}) = E(r)\boldsymbol{e}_{\boldsymbol{r}}.$$

There are two factors to this argument. First with the origin at the centre of the sphere, as is the only sensible choice, we are free to rotate the axis as much as we like without changing the physics, for this reason the strength of the field must be rotationally invariant so can't depend on the spherical coordinates ϑ or φ. The second part of the argument is that if there were a component of the electric field in the $\boldsymbol{e}_{\vartheta}$ or $\boldsymbol{e}_{\boldsymbol{\varphi}}$ direction then by the rotational symmetry this component must be the same all the way around the sphere. This means that the vector field has a closed loop. This in turn means that

the curl of the electric field is nonzero, which it cannot be in an electrostatics situation. Thus the field strength can only rely on r and the only direction the field can point is $\boldsymbol{e_r}$.

Now construct a Gaussian surface of a sphere of radius R centred on the origin. The field is always normal to this surface so

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(R)\boldsymbol{e_r} \cdot \mathrm{d}S\, \boldsymbol{e_r} = E(R)\, \mathrm{d}S\,.$$

Thus

$$\begin{aligned}\frac{Q_{\text{enc}}}{\varepsilon_0} &= \oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} \\ &= E(R) \oint_A \mathrm{d}S \\ &= E(R) 4\pi R^2.\end{aligned}$$

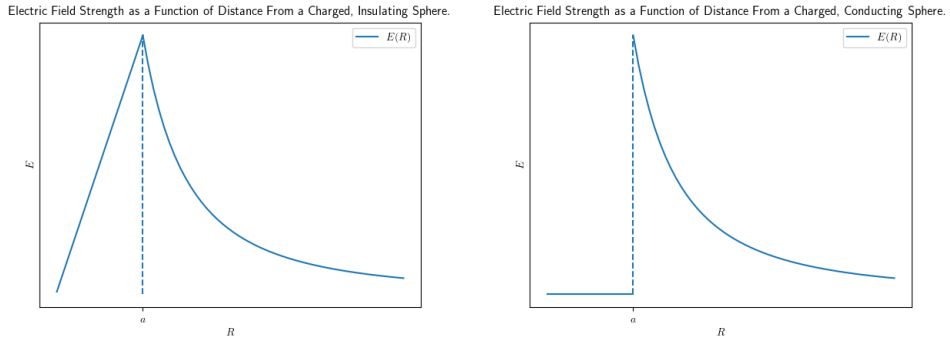
The value of Q_{enc} depends on R. If $R > a$ then

$$Q_{\text{enc}} = \rho V = \frac{4}{3}\pi a^3 \rho \implies E(R) = \frac{Q_{\text{enc}}}{\varepsilon_0} = \frac{4}{3}\pi a^3 \rho \frac{1}{\varepsilon_0 4\pi R^2} = \frac{\rho a^3}{3\varepsilon_0 R^2} = \frac{Q}{4\pi\varepsilon_0}\frac{1}{R^2},$$

where Q is the total charge of the sphere. This is the same as the electric field for a point charge, Q. If instead $R < a$ then

$$Q_{\text{enc}} = \rho V = \frac{4}{3}\pi R^3 \rho \implies E(R) = \frac{Q_{\text{enc}}}{\varepsilon_0} = \frac{4}{3}\pi R^3 \rho \frac{1}{\varepsilon_0 4\pi R^2} = \frac{\rho}{3\varepsilon_0} R.$$

This is plotted in figure 3.1a.



(a) The electric field strength as a function of distance from the centre of a charged insulating sphere.

(b) The electric field strength as a function of distance from the centre of a charged conducting sphere.

Figure 3.1: Electric field strength due to insulating and conducting spheres.

3.4.2 Conducting Sphere

Consider now the same set up as above but the sphere is made of a conductor. Outside of the sphere all of the same logic applies and so for $R > a$ we have

$$E(R) = \frac{Q}{4\pi\varepsilon_0}\frac{1}{R^2} = \frac{\rho a^3}{3\varepsilon_0 R^2}.$$

Inside the sphere the electric field is zero so we have $E(R) = 0$ for $R < a$. This is plotted in figure 3.1b.

3.5 Cylindrical Symmetry

Take an infinitely long, thin wire with uniform charge density, λ. We argue that by the symmetry of the situation the electric field must be

$$\boldsymbol{E}(\boldsymbol{r}) = E(\rho)\boldsymbol{e_\rho}$$

where we are working in cylindrical coordinates, (ρ, φ, z). The argument for this is similar to the spherical case. If we place the origin on the wire, this is the only sensible choice for the location of the origin, then we are free to place it anywhere along the wire and we can define $\varphi = 0$ as any position around the wire. This means that the field cannot depend on z or φ as then the origin position we pick would effect the field strength which is non-physical. Similarly if there is a component in the $\boldsymbol{e_\varphi}$ direction then it must be the same all the way around the cylinder which means that there is a closed loop in the electric field meaning that $\nabla \times \boldsymbol{E} \neq \boldsymbol{0}$ which cannot be the case in an electrostatics situation. Finally the field can't have a component in the $\boldsymbol{e_z}$ direction as this would cause motion of charge along the wire which would result in the charge distribution not being uniform. Therefore we are left only with dependence on ρ and a component in the direction $\boldsymbol{e_\rho}$.

Now construct a Gaussian surface of a cylinder of radius ρ which shares an axis with the wire. The field is always normal to this surface over the curved part so

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(\rho)\boldsymbol{e_\rho} \cdot \mathrm{d}S\, \boldsymbol{e_\rho} = E(\rho)\, \mathrm{d}S\,.$$

Over the flat ends of the cylinder the field is parallel to the surface so for the top surface

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(\rho)\boldsymbol{e_\rho} \cdot \mathrm{d}S\, \boldsymbol{e_z} = 0,$$

the case of the bottom surface is the same but $\mathrm{d}S$ is negative. From Gauss' law we then have

$$\begin{aligned} \frac{Q_{\text{enc}}}{\varepsilon_0} &= \oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} \\ &= E(\rho) \int_{\text{CSA}} \mathrm{d}S \\ &= E(\rho) 2\pi\rho L \end{aligned}$$

where CSA is the curved surface area of the cylinder and L is the length of the cylinder. The charge enclosed is simply $Q_{\text{enc}} = \lambda L$ so rearranging the above equation gives us

$$E(\rho) = \frac{\lambda L}{\varepsilon} \frac{1}{2\pi\rho L} = \frac{\lambda}{2\pi\rho\varepsilon_0}.$$

This is plotted in figure 3.2.

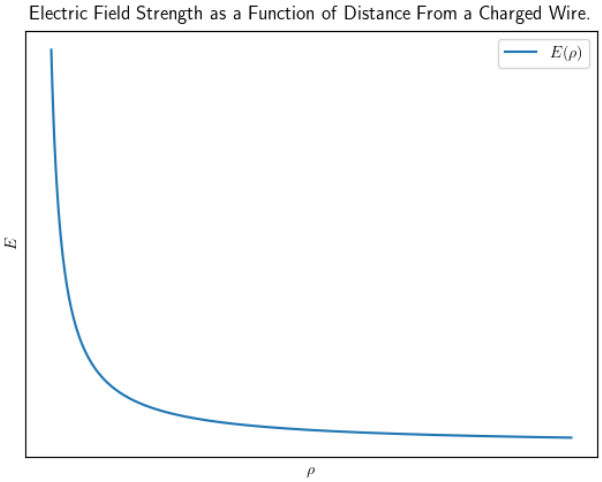


Figure 3.2: The electric field strength due to a charged wire.

3.6 Planar Symmetry

3.6.1 Insulating Plane

Take an infinite plane with uniform charge density σ. By symmetry we argue that

$$\boldsymbol{E}(\boldsymbol{r}) = E(z)\boldsymbol{e_z}$$

where we define $\boldsymbol{e_z}$ as normal to the plane and place the origin in the plane. The argument for this is, again, similar to the previous arguments. Since we are free to place the origin anywhere in the plane there can be no x or y dependence in the field. We are also free to rotate the axis around the z axis which means that any component in the $\boldsymbol{e_x}$ or $\boldsymbol{e_y}$ direction must form a complete loop in the electric field which would mean $\nabla \times \boldsymbol{e_E} \neq \boldsymbol{0}$, this is not possible in electrostatics so there must be no $\boldsymbol{e_x}$ or $\boldsymbol{e_y}$ components.

We choose a Gaussian surface that is a cylinder of radius R. We place it so that its flat faces are parallel to the plane and one is above the plane and the other below. As well as this we have both faces the same distance from the plane.

Along the curved surface the field is parallel to the surface so $\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = 0$. On the top face

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(z)\boldsymbol{e_z} \cdot \mathrm{d}S\, \boldsymbol{e_z} = E(z)\,\mathrm{d}S\,.$$

For the bottom case we have $z < 0$. Since we are free to define z-axis in either direction we must have mirror symmetry in the (x, y)-plane meaning that $E(-z) = -E(z)$. This means that for the bottom face of the cylinder we have

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(z)\boldsymbol{e_e} \cdot (-\,\mathrm{d}S\, \boldsymbol{e_e}) = -E(z)\,\mathrm{d}S\,.$$

However since $z < 0$ we have $E(z) = E(-|z|) = -E(|z|)$ so

$$\boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = E(|z|)\,\mathrm{d}S\,.$$

This means that the contribution from the two faces is equal to twice the contribution from the top face. Applying Gauss' law we have

$$\begin{aligned}\frac{Q_{\text{enc}}}{\varepsilon_0} &= \oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} \\ &= E(z) \int_{2\circ} \mathrm{d}S \\ &= 2\pi R^2 E(z)\end{aligned}$$

where $2\circ$ is the two circular faces. The charge enclosed is $Q_{\text{enc}} = \pi R^2 \sigma$ so rearranging the above equation gives us

$$E(z) = \operatorname{sgn}(z)\frac{Q_{\text{enc}}}{2\pi R^2 \varepsilon_0} = \operatorname{sgn}(z)\frac{\pi R^2 \sigma}{2\pi R^2 \varepsilon_0} = \operatorname{sgn}(z)\frac{\sigma}{2\varepsilon_0},$$

where

$$\operatorname{sgn}(z) = \begin{cases} 1, & z > 0, \\ 0, & z = 0, \\ -1, & z < 0. \end{cases}$$

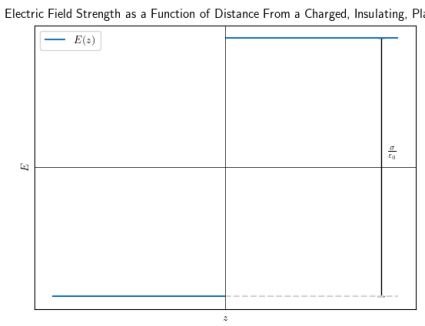
Notice that the only dependence on z is which side of the plane we are on as that affects the direction of the field. The field strength is constant. There is a discontinuous jump of σ/ε_0 as we move from one side of the plane to the other. This is plotted in figure 3.3a If instead we have a conducting, charged, plane then the situation is a little different. By a conducting plane what we mean is that the plane is the surface of a conductor that continues on below the plane forever. The same logic as before works above the plane but now we have that the electric field below the plane is zero as it is inside a conductor. Thus

$$\begin{aligned}\frac{Q_{\text{enc}}}{\varepsilon_0} &= \oint_A \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} \\ &= E(z) \int_{\circ} \mathrm{d}S\,, \qquad z > 0 \\ &= E(z)\pi R^2\end{aligned}$$

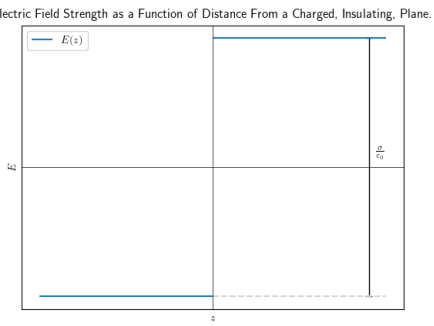
So the electric field strength is

$$E(z) = \begin{cases} \frac{\sigma}{\varepsilon_0}, & z > 0, \\ 0, & z \leq 0. \end{cases}$$

This is plotted in figure 3.3b. Note that in both cases there is a discontinuity of σ/ε_0 when passing from one side of the plane to the other.



(a) Electric field strength as a function of distance from an insulating charged plane.



(b) Electric field strength as a function of distance from an insulating charged plane.

Figure 3.3: Electric field strength due to insulating and conducting spheres.

4 Poisson's Equation

Poisson's equation (PE) is a partial differential equation (PDE) of the form

$$\nabla^2 \varphi = f$$

where $\varphi, f \colon \mathbb{R}^3 \to \mathbb{R}$. The specific case of $f(\boldsymbol{r}) = 0$ gives us

$$\nabla^2 \varphi = 0$$

which is Laplace's equation (LE). This crops up a lot in EM as the electrostatic potential and electric field are connected by

$$\boldsymbol{E} = -\nabla V$$

and the electric field and charge density are related by Gauss' law:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0}.$$

Combining these we get

$$\nabla \cdot (\nabla V) = \nabla^2 V = -\frac{\rho}{\varepsilon_0}.$$

This simplifies to LE if $\rho = 0$ everywhere in the region of interest.

The most obvious solution to Laplace's equation is a potential of $V(\boldsymbol{r}) = V_0$ for some constant V_0. However as with all PDE there will be a set of boundary conditions (BC) and typically $V(\boldsymbol{r}) = V_0$ won't satisfy these BC. Both PE and LE are among the most important PDE in physics and appear in many scenarios.

4.1 Properties of Poisson's Equation

If $V(\boldsymbol{r})$ is known then it is trivial to compute $\rho(\boldsymbol{r}) = -\varepsilon \nabla^2 V$ or $\boldsymbol{E}(\boldsymbol{r}) = -\nabla V$. The more likely scenario however is that we know $\rho(\boldsymbol{r})$ and we want to find the electric field. The first thing we should attempt should be to use Gauss' law, however this requires high levels of symmetry to be the most useful method. Lacking this symmetry the next thing that we can try to do is solve PE for the potential. There is an explicit solution, given by the original definition of the potential:

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|} \, \mathrm{d}^3 r' \, .$$

However this integral often has no analytic solution. There are methods for solving it numerically but this is not a numerical methods course so we won't discuss them.

There are two useful properties of PE that we use to find solutions:

1. The first useful property is linearity. If V_1 and ρ_1 satisfy PE and V_2 and ρ_2 satisfy PE then V_1+V_2 and $\rho_1+\rho_2$ satisfy PE. That is

$$\nabla^2(V_1+V_2) = -\frac{1}{\varepsilon_0}(\rho_1+\rho_2).$$

   This follows trivially from the linearity of ∇^2 as an operator which in turn follows from the linearity of partial derivatives as operators. In terms of EM this is just a restatement of the superposition principle. One use of this is if we have a shaped charge distribution, ρ, that can be created from simpler shaped charged distributions, ρ_i, then we can find the potential of each individual charge distribution and sum them together to get the potential of the entire charge distribution. For example a charged plane with a circular hole can be thought of as a plane without a hole and a charged disc the size of the hole, at the same position but with the opposite charge to the plane over he same area.

2. The second useful property is that the solution to PE is unique (possibly up to a constant term) for a given set of BC. There are two common ways that BC are given:
   - V is specified on the boundary – known as Dirichlet BC. The solution will be unique.
   - $\boldsymbol{E}$ is specified on the boundary – known as Neumann BC. The solution will be unique up to a constant term.

4.1.1 Proof of Uniqueness

Theorem 1: Uniqueness of the solution Poisson's Equation

Consider a region, $\mathcal{R}$, with boundary, $\mathcal{B}^a$. Let $\rho(\boldsymbol{r})$ be specified within $\mathcal{R}$. Let the BC be given by either
1. V is specified on $\mathcal{B}$.
2. $\boldsymbol{E} = -\nabla V$ is specified on $\mathcal{B}$.

Then any solution of Poisson's equation (PE),

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0},$$

which satisfies the boundary conditions is unique (up to a constant term in the case of the second set of boundary conditions.)

[a] It is possible that the boundary could be at infinity.

Proof. Suppose that V_1 and V_2 are two distinct solutions to $\nabla^2 V = -\rho/\varepsilon_0$. Define $\psi = V_1 - V_2$. Then

$$\begin{aligned}\nabla^2\psi &= \nabla^2(V_1 - V_2)\\ &= \nabla^2 V_1 - \nabla^2 V_2\\ &= \rho - \rho\\ &= 0.\end{aligned}$$

Thus ψ is a solution to LE. Multiplying by ψ we have

$$\psi\nabla^2\psi = \psi\cdot 0 = 0.$$

Consider the following:

$$\nabla\cdot(\psi\nabla\psi) - (\nabla\psi)\cdot(\nabla\psi)$$

Applying the product rule for the divergence of the product of a scalar field and vector field to the first term we get

$$(\nabla\cdot\psi)\cdot(\nabla\cdot\psi) - \psi(\nabla\cdot(\nabla\psi) - (\nabla\psi)\cdot(\nabla\psi).$$

The first and last terms cancel and we are left with

$$\psi(\nabla\cdot(\nabla\psi)) = \psi\nabla^2\psi = 0.$$

So the whole term is zero over the whole of $\mathcal{R}$. This means that integrating this term over $\mathcal{R}$ will also be zero:

$$\int_{\mathcal{R}} \nabla \cdot (\psi \nabla \psi) - (\nabla \psi) \cdot (\nabla \psi) \, \mathrm{d}V = 0.$$

Applying the divergence theorem to the first term this becomes

$$\int_{\mathcal{B}} \psi \nabla \psi \cdot \mathrm{d}\boldsymbol{S} - \int_{\mathcal{R}} (\nabla \psi) \cdot (\nabla \psi) \, \mathrm{d}V = 0.$$

For either set of boundary conditions the first term is zero. The second term is non-negative everywhere as the integrand is the norm of a vector. Therefore for the integral to be zero we must have that the integrand is zero. The norm of a vector is only zero when that vector is zero therefore

$$\nabla \psi = 0.$$

This means that $\psi = V_1 - V_2$ is constant. So V_1 and V_2 differ by at most a constant. If the boundary conditions were given in terms of V at $\mathcal{B}$ then we know that $V_1 = V_2$ at the boundaries so this constant is zero. □

Example 4.1. Consider a cavity, $\mathcal{R}$, in a conductor. We claim that if $\rho = 0$ in the cavity then $\boldsymbol{E} = \boldsymbol{0}$ inside the cavity.

The inner surface of the conductor is an equipotential since it is conducting. This means that $V = V_0$ for some constant V_0. This is our BC. Inside the cavity we have $\nabla^2 V = 0$ since there is no charge. Thus we have to solve LE subject to the condition that $V = V_0$ on the boundary. One solution to this is $V = V_0$ everywhere in $\mathcal{R}$. By the uniqueness theorem above we know that this is the only solution. Now we simply compute $\boldsymbol{E} = -\nabla V = -\nabla V_0 = \boldsymbol{0}$.

4.2 The Method of Images

The method of images is a method for solving PE by placing 'image charges' outside of the region, $\mathcal{R}$, such that they reproduce the required BC. These charges don't affect PE inside $\mathcal{R}$ as they aren't in the region so don't change ρ in the region. However the field that results from the superposition of these image charges as well as any pre-existing charges is the correct solution to PE.

Example 4.2. Consider a conducting plane with a point charge, Q, placed a distance a above the plane. What is the potential in the region, $\mathcal{R}$, above the plane?

The charge density above the plane is

$$\rho(\boldsymbol{r}) = Q\delta(z-a)\delta(x)\delta(y)$$

where we have placed the origin in the plane directly below the charge and have the z-axis normal to the plane. Our boundary condition is that $V(x, y, 0) = 0$ as at the plane $\boldsymbol{E} = 0$ so $V = V_0$ and we choose $V_0 = 0$ for simplicity.

Unfortunately $V = V_0$ is not a solution to PE here as we know that the potential from a point charge falls away as $1/r$ so at the origin (in the plane directly below the charge) we expect the potential to be $V = 1/a$ of what it is at the point charge. We need to solve

$$\nabla^2 V(\boldsymbol{r}) = -\frac{\rho(\boldsymbol{r})}{\varepsilon_0}.$$

To do this we can add a homogenous solution, V_{im} to PE. By homogenous here we mean that $\nabla^2 V_{\text{im}} = 0$, i.e. a solution to LE. This solution only needs to be homogeneous for $z \geq 0$ since this is the region of interest. We know that the potential due to the point charge is

$$V_Q(\boldsymbol{r}) = \frac{Q}{r\pi\varepsilon_0} \frac{1}{\sqrt{x^2 + y^2 + (z-a)^2}}.$$

We place an image charge, $-Q$, at $z = -a$ which gives us an image potential of

$$V_{\text{im}}(\boldsymbol{r}) = -\frac{Q}{r\varepsilon_0} \frac{1}{\sqrt{x^2 + y^2 + (z+a)^2}}.$$

At every point on the plane we have

$$V_Q + V_{\text{im}} = \frac{Q}{r\varepsilon_0}\frac{1}{\sqrt{x^2+y^2+(0-a)^2}} - \frac{Q}{r\varepsilon_0}\frac{1}{\sqrt{x^2+y^2+(0+a)^2}} = 0$$

so the boundary conditions are satisfied. Thus $V = V_Q + V_{\text{im}}$ is the solution and gives the potential everywhere.

In reality this potential isn't caused by two point charges. Rather the real point charge causes the charge density of the plane to become non-uniform in a way that the final potential is as given above.

---

5 Electric Dipoles and Multipoles

The motivation behind this section is to study a general, non-trivial, charge distribution, $\rho(\boldsymbol{r})$, and in particular find an approximation of the potential,

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\,\mathrm{d}^3r'\,,$$

which applies to points, $\boldsymbol{r}$, far from where $\rho(\boldsymbol{r'}) \neq 0$.

5.1 Electric Dipoles

An **electric dipole** is formed from two charges, $\pm q$, fixed distance a apart. The vector from q to $-q$ is defined to be $\boldsymbol{a}$. The **electric dipole moment** is then defined to be $\boldsymbol{p} = q\boldsymbol{a}$. An electric dipole is shown in figure 5.1. Dipoles like this are important because many molecules, such as H_2O and CO have

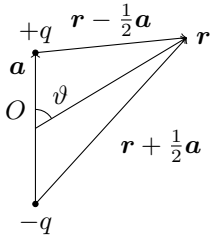


Figure 5.1: An electric dipole.

permanent dipoles and all molecules/atoms acquire an induced dipole in an external field.

5.1.1 Field of an Electric Dipole

The electric potential of an electric dipole is simply the superposition of the potential due to the two point charges:

$$V(\boldsymbol{r}) = \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r_+} - \frac{1}{r_-}\right)$$

where

$$\boldsymbol{r}_\pm = \boldsymbol{r} \mp \frac{1}{2}\boldsymbol{a}.$$

That is $\boldsymbol{r}_\pm$ is the vector from $\pm q$ to $\boldsymbol{r}$. The next question that we ask is what is the field like for $r \gg a$? This is known as the far field approximation. We can Taylor expand the $1/r_\pm$ terms in the potential. To do this we first note that

$$\begin{aligned} r_\pm^2 &= \boldsymbol{r}_\pm \cdot \boldsymbol{r}_\pm \\ &= \left(\boldsymbol{r} \mp \frac{1}{2}\boldsymbol{a}\right)\cdot\left(\boldsymbol{r} \mp \frac{1}{2}\boldsymbol{a}\right) \\ &= r^2 \mp \boldsymbol{a}\cdot\boldsymbol{r} + \frac{1}{4}a^2 \\ &= r^2 \mp ar\cos\vartheta + \frac{1}{4}a^2 \end{aligned}$$

$$= r^2\left(1 \mp \frac{a}{r}\cos\vartheta + \frac{a^2}{4r^2}\right).$$

Hence

$$\frac{1}{r_\pm} = \left[r^2\left(1 \mp \frac{a}{r}\cos\vartheta + \frac{a^2}{4r^2}\right)\right]^{-1/2} = \frac{1}{r}\left[\left(1 \mp \frac{a}{r}\cos\vartheta + \frac{a^2}{4r^2}\right)\right]^{-1/2}.$$

For $r > a$ this is of the form $(1+\varepsilon)^p$ with $|\varepsilon| < 1$ needed to use

$$(1+\varepsilon)^p = 1 + p\varepsilon + \mathcal{O}\big(\varepsilon^2\big).$$

Doing this gives

$$\begin{aligned}\frac{1}{r_\pm} &= \frac{1}{r}\left[1 - \frac{1}{2}\left(\mp\frac{a}{r}\cos\vartheta + \frac{a^2}{4r^2}\right) + \mathcal{O}\left(\frac{a^2}{r^2}\right)\right] \\ &= \frac{1}{r} \pm \frac{a}{2r^2}\cos\vartheta + \mathcal{O}\left(\frac{a^2}{r^3}\right).\end{aligned}$$

Substituting this into the potential gives

$$\begin{aligned}V(\boldsymbol{r}) &= \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r_+} - \frac{1}{r_-}\right) \\ &\approx \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{r} + \frac{a}{2r^2}\cos\vartheta - \frac{1}{r} + \frac{a}{2r^2}\cos\vartheta\right) \\ &= \frac{qa\cos\vartheta}{r^2\pi\varepsilon_0} \\ &= \frac{\boldsymbol{p}\cdot\hat{\boldsymbol{r}}}{4\pi\varepsilon_0 r^2}.\end{aligned}$$

Notice that this drops off as $1/r^2$ whereas the potential from a point charge drops off slower as $1/r$. You can think of this as the fact that there are positive and negative charges close together causing parts of the potential to cancel. What we have derived here is the far field limit, in that it is only valid for $r \gg a$. It is also known as the 'ideal dipole' potential which is what we would get if we too a limit as $a \to 0$, and $q \to \infty$ in such a way that $\boldsymbol{p}$ remains constant.

5.1.2 Dipole Interaction With an External Electric Field

The electrostatic energy of a point charge, q, in a potential, V, is $U = qV$. The energy of a dipole is then the superposition of the two point charges:

$$U_{\text{dip}} = qV_{\text{ext}}(\boldsymbol{a}/2) - qV_{\text{ext}}(-\boldsymbol{a}/2).$$

Here V_{ext} is the external potential. Taking a Taylor series gives

$$\begin{aligned}U_{\text{dip}} &\approx q\left[V_{\text{ext}}(0) + \frac{1}{2}\boldsymbol{a}\cdot\nabla V_{\text{ext}}\right] - q\left[V_{\text{ext}}(0) - \frac{1}{2}\boldsymbol{a}\cdot\nabla V_{\text{ext}}\right] \\ &= q\boldsymbol{a}\cdot\nabla V_{\text{ext}} \\ &= -q\boldsymbol{a}\cdot\boldsymbol{E}_{\mathbf{ext}} \\ &= -p\cdot\boldsymbol{E}_{\mathbf{ext}}\end{aligned}$$

We see that the dipole energy is minimised if $\boldsymbol{p}$ is parallel to $\boldsymbol{E}_{\mathbf{ext}}$ and the energy is maximised when they are antiparallel. The change in energy to change between parallel and antiparallel is

$$\Delta U = 2p|E_{\text{ext}}|.$$

The force experienced by the dipole is

$$\boldsymbol{F} = -\nabla U_{\text{dip}} = \nabla(\boldsymbol{p}\cdot\boldsymbol{E}_{\mathbf{ext}}).$$

If $\boldsymbol{E}_{\mathbf{ext}}$ is a uniform field (doesn't depend on $\boldsymbol{r}$) then the force is zero. This is because the two charges experience equal and opposite forces. However there is still a torque because the two charges aren't in the same location. This torque acts to align the dipole with $\boldsymbol{E}_{\mathbf{ext}}$ and is given by:

$$\boldsymbol{\tau} = \frac{1}{2}\boldsymbol{a}\times q\boldsymbol{E}_{\mathbf{ext}} - \frac{1}{2}\boldsymbol{a}\times(-q\boldsymbol{E}_{\mathbf{ext}}) = \boldsymbol{p}\times\boldsymbol{E}_{\mathbf{ext}}.$$

The work done by the torque to rotate the dipole from aligned with the field to an angle ϑ from the field is

$$W = \int_0^{\vartheta} \tau \, \mathrm{d}\vartheta = \int_0^{\vartheta} qE_{\text{ext}} \, \mathrm{d}\vartheta = pE_{\text{ext}}(1 - \cos\vartheta).$$

In a non-uniform field the force is generally more complex. It acts to move the dipole along the gradient of the field.

5.2 Multipole Expansion

Given a charge distribution ρ we say that ρ is bounded inside a region, $\mathcal{R}$, if, for $\boldsymbol{r} \notin \mathcal{R}$, $\rho(\boldsymbol{r}) = 0$. Let ρ be a charge distribution that is bounded inside the region $\mathcal{R}$. Then from the definition of the potential we know that

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0} \int_{\mathcal{R}} \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|} \, \mathrm{d}^3 r' \, .$$

This holds for all $\boldsymbol{r}$. However if $\boldsymbol{r} \notin \mathcal{R}$, that is $r \gg r'$, then we can make use of a Taylor expansion. Following the same steps as we did for a dipole we see that

$$\frac{1}{|\boldsymbol{r} - \boldsymbol{r'}|} = \frac{1}{r}\left[1 - \frac{2\boldsymbol{r}\cdot\boldsymbol{r'}}{r^2} + \frac{r'^2}{r^2}\right]^{-1/2}.$$

We now Taylor expand this but keep higher order terms:

$$\begin{aligned}
\left[1 - \frac{2\boldsymbol{r}\cdot\boldsymbol{r'}}{r^2} + \frac{r'^2}{r^2}\right]^{-1/2} &= 1 - \frac{1}{2}\left(-\frac{2\boldsymbol{r}\cdot\boldsymbol{r'}}{r^2} + \frac{r'^2}{r^2}\right) + \frac{3}{8}\left(-\frac{2\boldsymbol{r}\cdot\boldsymbol{r'}}{r^2} + \frac{r'^2}{r^2}\right)^2 + \mathcal{O}\left(\frac{1}{r^3}\right) \\
&= 1 + \frac{\boldsymbol{r}\cdot\boldsymbol{r'}}{r^2} - \frac{1}{2}\frac{r'^2}{r^2} + \frac{3}{2}\frac{(\boldsymbol{r}\cdot\boldsymbol{r'})^2}{r^4} - \frac{3}{2}\frac{r'^2(\boldsymbol{r}\cdot\boldsymbol{r'})}{r^4} - \frac{3}{8}\frac{r'^4}{r^4} + \mathcal{O}\left(\frac{1}{r^3}\right) \\
&= 1 + \frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}}{r} - \frac{1}{2}\frac{r'^2}{r^2} + \frac{3}{2}\frac{(\hat{\boldsymbol{r}}\cdot\boldsymbol{r'})^2}{r^2} - \frac{3}{2}\frac{r'^2(\hat{\boldsymbol{r}}\cdot\boldsymbol{r})}{r^3} - \frac{3}{8}\frac{r'^4}{r^4} + \mathcal{O}\left(\frac{1}{r^3}\right)
\end{aligned}$$

keeping only terms of order $1/r^2$ or lower

$$\begin{aligned}
&\approx 1 + \frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}}{r} - \frac{1}{2}\frac{r'^2}{r^2} + \frac{3}{4}\frac{(\hat{\boldsymbol{r}}\cdot\boldsymbol{r'})^2}{r^2} \\
&= 1 + \frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}}{r} + \frac{3(\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}) - r'^2}{2r^2}
\end{aligned}$$

Using this in the definition of the potential gives

$$\begin{aligned}
V(\boldsymbol{r}) &\approx \frac{1}{4\pi\varepsilon_0}\int_{\mathcal{R}} \mathrm{d}^3 r' \, \rho(\boldsymbol{r'})\frac{1}{r}\left[1 + \frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}}{r} + \frac{3(\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}) - r'^2}{2r^2}\right] \\
&= \frac{1}{4\pi\varepsilon_0}\int_{\mathcal{R}} \mathrm{d}^3 r' \, \rho(\boldsymbol{r'})\left[\frac{1}{r} + \frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}}{r^2} + \frac{3(\hat{\boldsymbol{r}}\cdot\boldsymbol{r'}) - r'^2}{2r^3}\right].
\end{aligned}$$

Defining some new terms this becomes

$$V(\boldsymbol{r}) \approx \frac{1}{4\pi\varepsilon_0}\frac{Q}{r} + \frac{1}{r\pi\varepsilon_0}\frac{\hat{\boldsymbol{r}}\cdot\boldsymbol{P}}{r^2} + \frac{1}{4\pi\varepsilon_0}\frac{1}{r^3}\frac{1}{2}\sum_{i,j}\mathcal{Q}_{ij}\hat{r}_i\hat{r}_j.$$

Where

$$Q = \int_{\mathcal{R}} \mathrm{d}^3 r' \, \rho(\boldsymbol{r'})$$

is the total charge,

$$\boldsymbol{P} = \int_{\mathcal{R}} \mathrm{d}^3 r' \, \boldsymbol{r'}\rho(\boldsymbol{r'})$$

is the net dipole moment, a vector with Cartesian components

$$P_i = \int_{\mathcal{R}} \mathrm{d}^3 r' \, r'_i\rho(\boldsymbol{r'}),$$

and $\mathcal{Q}$ is the quadrupole tensor which has Cartesian components

$$\mathcal{Q}_{ij} = \int_{\mathcal{R}} \mathrm{d}^3 r' \, (3r'_i r'_j - r'^2 \delta_{ij}) \rho(\boldsymbol{r}').$$

This representation of V is the **multipole expansion** of V in Cartesian coordinates. The first term is the monopole term. It is dominant when $Q \neq 0$ and reasonably approximates the far field charge distribution as a point charge at the origin. When the total charge, Q, is zero then the second term, the dipole term, dominates. If this term also vanishes then the third term, the quadrupole term, dominates. Note that another way of writing this terms is

$$\frac{1}{4\pi\varepsilon_0} \frac{1}{r^3} \frac{1}{2} \hat{\boldsymbol{r}}^{\mathsf{T}} \mathcal{Q} \hat{\boldsymbol{r}} = \frac{1}{4\pi\varepsilon_0} \frac{1}{r^5} \frac{1}{2} \boldsymbol{r}^{\mathsf{T}} \mathcal{Q} \boldsymbol{r}.$$

It is possible to take even more terms of the Taylor series and end up with more terms, however we will stop at three. Note that each term of the expansion is a solution to Laplace's equation (LE) and therefore we can use the method of images with image dipoles and quadrupoles as well as image charges.

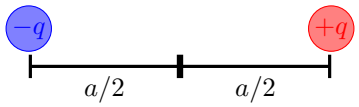


Figure 5.2: The dipole setup used in example 5.1

Example 5.1. A dipole formed of two charges, $\pm q$, aligned along the x-axis so that q is at $(a/2, 0, 0)$ and $-q$ is at $(-a/2, 0, 0)$ has a charge density given by

$$\rho(\boldsymbol{r}) = q[\delta(x - a/2) - \delta(x + a/2)]\delta(y)\delta(z).$$

Clearly $Q = 0$. The x component of the dipole moment is

$$\begin{aligned} P_x &= \int x\rho(\boldsymbol{r}) \, \mathrm{d}V \\ &= q \int x[\delta(x - a/2) - \delta(x + a/2)] \, \mathrm{d}x \int \delta(y) \, \mathrm{d}y \int \delta(z) \, \mathrm{d}z \\ &= q\frac{a}{2} - q\left(\frac{a}{2}\right) \\ &= qa. \end{aligned}$$

The y component is

$$\begin{aligned} P_y &= \int y\rho(\boldsymbol{r}) \, \mathrm{d}V \\ &= \int [\delta(x - a/2) - \delta(x + a/2)] \, \mathrm{d}x \int y\delta(y) \, \mathrm{d}y \int \delta(z) \, \mathrm{d}z \\ &= 0 \end{aligned}$$

since the middle integral is zero. Similarly we can show that $P_z = 0$. This means that $\boldsymbol{P} = q\boldsymbol{a}$, so we are justified in calling $\boldsymbol{P}$ the dipole moment.

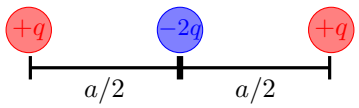


Figure 5.3: The quadrupole setup used in example 5.2

Example 5.2. Three charges are placed in a line along the x-axis. Two of the charges have charge q and are at $\pm a/2$. The third charge is at the origin and has charge $-2q$. The charge density is

$$\rho(\boldsymbol{r}) = q[\delta(x - a/2) + \delta(x + a/2) - 2\delta(x)]\delta(y)\delta(z).$$

Again, $Q = 0$. The x component of the dipole moment is

$$\begin{aligned} P_x &= \int x\rho(\boldsymbol{r})\,\mathrm{d}V \\ &= q\int x[\delta(x-a/2)+\delta(x+a/2)-2\delta(x)]\int\delta(y)\,\mathrm{d}y\int\delta(z)\,\mathrm{d}z \\ &= q\left[\frac{a}{2}-\frac{a}{2}+0\right] \\ &= 0 \end{aligned}$$

In a similar way to the previous example $P_y = P_z = 0$. Thus the dipole moment vanishes.

Next we calculate the quadrupole moment. In general

$$\mathcal{Q}_{ij} = \int (3x_i x_j - r^2\delta_{ij})\rho(\boldsymbol{r})\,\mathrm{d}V\,.$$

The $\mathcal{Q}_{xx}$ component is

$$\begin{aligned} \mathcal{Q}_{xx} &= \int (3x^2-r^2)q[\delta(x-a/2)+\delta(x+a/2)-2\delta(x)]\delta(y)\delta(z)\,\mathrm{d}V \\ &= \int (3x^2-(x^2+y^2+z^2))q[\delta(x-a/2)+\delta(x+a/2)-2\delta(x)]\delta(y)\delta(z)\,\mathrm{d}V \\ &= 3q\left(\frac{a}{2}\right)^2 - q\left(\frac{a}{2}\right)^2 + 3q\left(-\frac{a}{2}\right)^2 - q\left(-\frac{a}{2}\right)^2 \\ &= qa^2 \end{aligned}$$

Next we will calculate $\mathcal{Q}_{xy}$:

$$\begin{aligned} \mathcal{Q}_{xy} &= \int 3xyq[\delta(x-a/2)+\delta(x+a/2)-2\delta(x)]\delta(y)\delta(z)\,\mathrm{d}V \\ &= \int 3xq[\delta(x-a/2)+\delta(x+a/2)-2\delta(x)]\,\mathrm{d}x\int y\delta(y)\,\mathrm{d}y\int\delta(z)\,\mathrm{d}z \\ &= 0 \end{aligned}$$

If we calculated every component we would find that

$$\mathcal{Q} = qa^2\begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}.$$

We can then approximate the potential as

$$\begin{aligned} V(\boldsymbol{r}) &= \frac{1}{8\pi\varepsilon_0}\frac{1}{r^5}\boldsymbol{r}^{\mathsf{T}}\mathcal{Q}\boldsymbol{r} \\ &= \frac{qa^2}{8\pi\varepsilon_0}\frac{1}{r^5}\begin{pmatrix} x & y & z \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ &= \frac{qa^2}{8\pi\varepsilon_0}\frac{1}{r^5}\begin{pmatrix} x & y & z \end{pmatrix}\begin{pmatrix} x \\ -\frac{1}{2}y \\ -\frac{1}{2}z \end{pmatrix} \\ &= \frac{qa^2}{8\pi\varepsilon_0}\frac{1}{r^5}\left[x^2 - \frac{1}{2}(y^2+z^2)\right]. \end{aligned}$$

Example 5.3. Four charges $\pm q$ are placed on the corners of a square in the (x, y)-plane with side length a with the origin at the centre of the square. They are arranged such that diagonally opposite charges have the same sign and charges connected by an edge have opposite signs. The charge density is

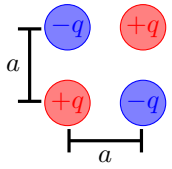


Figure 5.4: The quadrupole setup used in example 5.3

$$\rho(\boldsymbol{r}) = q[\delta(x-a/2)\delta(y-a/2) - \delta(x+a/2)\delta(y-a/2) \\ - \delta(x-a/2)\delta(y+a/2) + \delta(x+a/2)\delta(y+a/2)]\delta(z).$$

Again it can be shown that $\boldsymbol{P} = \boldsymbol{0}$ and clearly $Q = 0$. If we calculate the quadrupole components then we find that

$$\mathcal{Q} = qa^2 \begin{pmatrix} 0 & 3 & 0 \\ 3 & 0 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

The potential can then be approximated as

$$V(\boldsymbol{r}) = \frac{qa^2}{8\pi\varepsilon_0}\frac{1}{r^5}[6xy - 2z^2].$$

---

6 Electrostatic Energy and Capacitors

6.1 Electrostatic Energy of a General Charge Distribution

If we start with an assembly of $n-1$ point charges, q_i, at position $\boldsymbol{r_i}$ then the potential at $\boldsymbol{r}$ is given by the superposition of the potentials of each particle:

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{j=1}^{n-1}\frac{q_i}{|\boldsymbol{r} - \boldsymbol{r_j}|}.$$

If we then bring another charge, q_n, from infinity to $\boldsymbol{r_n}$ then the work required to do so is

$$W_n = q_n V(\boldsymbol{r_n}) = \frac{q_n}{\frac{q_n}{4\pi\varepsilon_0}}\sum_{j=1}^{n-1}\frac{q_j}{|\boldsymbol{r_n} - \boldsymbol{r_j}|}.$$

We can write out this sum for the first few values of n to see how it progresses:

$$\begin{aligned} W_1 &= 0 \\ W_2 &= \frac{q_2 q_1}{4\pi\varepsilon_0}\frac{1}{|\boldsymbol{r_2} - \boldsymbol{r_1}|} \\ W_3 &= \frac{q_3 q_1}{4\pi\varepsilon_0}\frac{1}{|\boldsymbol{r_3} - \boldsymbol{r_1}|} + \frac{q_3 q_2}{4\pi\varepsilon_0}\frac{1}{|\boldsymbol{r_3} - \boldsymbol{r_2}|} \end{aligned}$$

In general the total work required to assemble n charges, which is the electrostatic energy, U_E, is given by

$$\begin{aligned} U_E &= \sum_{i=1}^{n} W_i \\ &= \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{q_i q_j}{|\boldsymbol{r_i} - \boldsymbol{r_j}|} \\ &= \frac{1}{8\pi\varepsilon_0}\sum_{\substack{i,j \\ i\neq j}}^{n}\frac{q_i q_j}{|\boldsymbol{r_i} - \boldsymbol{r_j}|}. \end{aligned}$$

In the last step we collapsed two sums into one by noting that a sum over i and j with $i > j$ is equivalent to a sum over i and j with $i \neq j$ except that the second allows for $(i, j) = (2, 1)$ and $(i, j) = (1, 2)$. Due to symmetry the value of both of these terms is the same so we end up double counting the allowed states where $i > j$ if we include all $j \neq i$. For this reason we have to divide by 2 which explains the factor of 8 in the denominator.

In the limit of a continuous charge density, ρ, the electrostatic energy is

$$U_E = \frac{1}{2} \int \rho(\boldsymbol{r}) V(\boldsymbol{r}) \, \mathrm{d}^3 r \, .$$

Using Maxwell's first law this becomes

$$U_E = \frac{\varepsilon_0}{2} \int V(\boldsymbol{r}) (\nabla \cdot \boldsymbol{E}) \, \mathrm{d}^3 r \, .$$

We then use the product rule,

$$\nabla \cdot (V \boldsymbol{E}) = V \nabla \cdot \boldsymbol{E} + (\nabla V) \cdot \boldsymbol{E} = V \nabla \cdot \boldsymbol{E} - \boldsymbol{E} \cdot \boldsymbol{E} = V \nabla \cdot \boldsymbol{E} - E^2,$$

to change the integrand to

$$U_E = \frac{\varepsilon_0}{2} \int \nabla \cdot [V(\boldsymbol{r}) \boldsymbol{E}(\boldsymbol{r})] + E^2 \, \mathrm{d}^3 r \, .$$

Splitting the integral at the addition and applying the divergence theorem to the first integral we get

$$U_E = \frac{\varepsilon_0}{2} \oint_S V(\boldsymbol{r}) \boldsymbol{E}(\boldsymbol{r}) \cdot \mathrm{d}\boldsymbol{S} + \frac{\varepsilon_0}{2} \int E^2 \, \mathrm{d}^3 r \, .$$

So far we have made no assumptions about the volume over which our integration occurs. This means we are free to pick any boundary, S. We choose S to be at infinity. Assuming that our charge distribution is bound we know that outside of the region containing it V drops of as at least $1/r$ and E drops off as at least $1/r^2$. Thus VE drops of as at least $1/r^3$. The area of the surface however only grows as $1/r^2$ so the first integral must be zero. Thus the internal energy is

$$U_E = \frac{\varepsilon_0}{2} \int |\boldsymbol{E}(\boldsymbol{r})|^2 \, \mathrm{d}^3 r = \int u_E \, \mathrm{d}^3 r \, ,$$

where

$$u_E = \frac{\varepsilon_0}{2} |\boldsymbol{E}(\boldsymbol{r})|^2$$

is the energy density.

Note that this derivation started from a continuous charge distribution. This does not apply to point charges. If we were to apply it to point charges there are self interaction terms that we would have to exclude. We can see this form the point charge equations, if we try to include the interaction of a point charge with itself we get $|\boldsymbol{r_i} - \boldsymbol{r_i}| = 0$ so we end up dividing by zero. We are now assuming that the integral is over all space, if we wish to only consider a small volume we either require that $E = 0$ outside of that area or we have to consider the first integral that we reasoned to be zero for an infinite volume. As a final warning the electrostatic energy is quadratic in field strength so the superposition principle *does not* apply to energy densities.

6.2 Capacitors

A capacitor comprises of two neighbouring conducting bodies with equal and opposite charges, $\pm Q$. The capacitance is defined to be

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

where V is the potential difference between the two bodies. This is well defined as the surfaces of the conductors are equipotentials so it doesn't matter where we select on each conductor to measure the potential difference.

6.2.1 Parallel Plate Capacitors

The simplest capacitor is a parallel plate capacitor where we model the two conductors as thin, infinite, parallel, planes a distance d apart. We can easily calculate the field from the superposition of the fields from two charged planes. The set up and electric field from each plate is shown in figure 6.1. The magnitude of the field from each plate is the same and equal to $\sigma/2\varepsilon_0$, it is also the same everywhere. This is important because it means that between the plates where the fields align the net field is σ/ε_0. However outside of the plates the fields are anti-aligned and cancel so the net field is 0. The potential

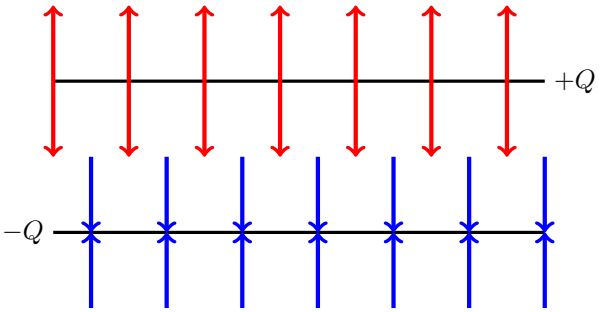


Figure 6.1: Parallel plate capacitor electric field.

difference between the fields is given by

$$V = -\int_0^d E_z \,\mathrm{d}z = \frac{\sigma}{\varepsilon_0} d = \frac{Qd}{A\varepsilon_0}$$

where we have chosen a coordinate system such that $\boldsymbol{e_z}$ is normal to the plates and points from the negative to the positive plate. The origin is also chosen to be in the negative plate. The area of each plate is A. The capacitance is then given by

$$C = \frac{Q}{V} = \frac{A\varepsilon_0}{d}.$$

Notice that the capacitance depends only on the geometry of the plates (their area and the distance between) it is independent of the charge and potential difference. It turns out that in general capacitance is a purely geometric property.

We can use the charge distribution to calculate the electrostatic energy of the charged capacitor.

$$\begin{aligned} U_E &= \frac{1}{2} Q(V_1 - V_2) \\ &= \frac{1}{2} QV \\ &= \frac{1}{2} \frac{Q^2}{E} \end{aligned}$$

We can also use the fact that the electric field vanishes outside of the plates to calculate the electrostatic energy of the charged capacitor.

$$\begin{aligned} U_E &= \frac{\varepsilon_0}{2} \int E^2 \,\mathrm{d}^3 r \\ &= \frac{\varepsilon_0}{2} \left(\frac{\sigma}{\varepsilon_0}\right)^2 \int \mathrm{d}^3 r \\ &= \frac{\sigma^2}{2\varepsilon_0} dA \\ &= \frac{dQ^2}{2\varepsilon_0 A} \\ &= \frac{1}{2} \frac{Q^2}{E}. \end{aligned}$$

Here we have used that the empty integral is just the volume which in this case since $E = 0$ outside of the capacitor is just the volume between the plates, dA.

6.2.2 Edge Effects

This section is non-examinable

So far we have assume that the plates are infinite and therefore the fields are uniform between the plates. However in reality the plates cannot be infinite. For a finite sized capacitor the field bulges out of the capacitor and isn't uniform between the plates.

Suppose that instead of infinite planes the capacitor is made from two discs of radius R. We can perform an integral over the charge distribution to obtain the potential a height z above the disc:

$$V(z) = \frac{1}{4\pi\varepsilon_0}\int_S \frac{\sigma}{(\rho^2+z^2)^{1/2}}\,\mathrm{d}S\,.$$

In plane polar coordinates $\mathrm{d}S = \rho\,\mathrm{d}\rho\,\mathrm{d}\varphi$ so

$$\begin{aligned} V(z) &= \sigma 4\pi\varepsilon_0 \int_0^{2\pi}\mathrm{d}\varphi\int_0^R \mathrm{d}\rho\,\frac{\rho}{(\rho^2+z^2)^{1/2}} \\ &= \frac{\sigma}{4\pi\varepsilon_0}2\pi\left[(\rho+z^2)^{1/2}\right]_{\rho=0}^{\rho=R} \\ &= \frac{\sigma}{2\varepsilon_0}\left[(R^2+z^2)^{1/2}-z\right]. \end{aligned}$$

This means that the z component of the electric field is

$$E_z = -\frac{\partial V}{\partial z} = -\frac{\sigma}{2\varepsilon_0}\left[\frac{z}{(R^2+z^2)^{1/2}}-1\right].$$

This depends on z so the field is not uniform between the plates. There are two interesting limiting cases. First $R \ll z$:

$$\boldsymbol{E}(R) = \frac{\sigma}{2\varepsilon_0}\left[1-\left(1+\frac{R^2}{z^2}\right)^{-1/2}\right]\boldsymbol{e_z}$$

Taylor expanding the binomial term gives

$$\begin{aligned} &\approx \frac{\sigma}{2\varepsilon_0}\left[1-\left(1-\frac{1}{2}\frac{R^2}{z^2}\right)\right]\boldsymbol{e_z} \\ &= \frac{\sigma R^2}{4\varepsilon_0 z^2}\boldsymbol{e_z} \\ &= \frac{Q}{4\pi\varepsilon_0 z^2}\boldsymbol{e_z} \end{aligned}$$

where $Q = \sigma\pi R^2$ is the total charge. We see that in this limit we can essentially view the capacitors as a point charge when we are far enough away that we can approximate it as having no volume.

The other interesting case is when $R \gg z$. In this case the term $z/(R^2+z^2)^{-1/2} \approx 0$ so

$$\boldsymbol{E}(R) = \frac{\sigma}{2\varepsilon_0}\boldsymbol{e_z}.$$

So in the case that the discs are very large they approximate infinite planes. It is only when $R \approx z$ (i.e. $A \approx d^2$) that we need to consider the edge effects of the capacitor. Fortunately in practice this is rarely necessary.

Part II
Magnetostatics

7 The Magnetic Field

7.1 The Magnetic Force

The **magnetic field**, $\boldsymbol{B}$, is defined by the force, $\boldsymbol{F}$, experienced by a charge, q, with velocity $\boldsymbol{v}$, in an external electric field $\boldsymbol{E}$:

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}).$$

This is called the **Lorentz force law**. The units of $\boldsymbol{B}$ are teslas, T, defined by $1\,\mathrm{T} = 1\,\mathrm{N\,A^{-1}\,m^{-1}}$. This is quite a large unit and it is common to use an alternative unit called a gauss defined by $10^{-4}\,\mathrm{T} = 1\,\mathrm{G}$ sometimes also denoted 1 Gs.

The important thing about this formula is that the force due to the magnetic field, $\boldsymbol{v}\times\boldsymbol{B}$, is perpendicular to the velocity. This means that magnetic fields do no work as $\boldsymbol{v} \cdot (\boldsymbol{v} \times \boldsymbol{B}) = \boldsymbol{0}$.

7.2 Current

A **current** is a moving density of charge. For a **steady current** at any point, $\boldsymbol{r}$, a constant (time independent) density of charge moves past $\boldsymbol{r}$ in a given time. Steady currents are important as if all currents are steady we will see that this leads to a constant magnetic field in the same way that stationary charges lead to a constant electric field. A current formed by $n(\boldsymbol{r})$ charges, q, moving past the point $\boldsymbol{r}$ with average velocity $\boldsymbol{v}(\boldsymbol{r})$ can be described by a **current density**

$$\boldsymbol{J}(\boldsymbol{r}) = n(\boldsymbol{r})q\boldsymbol{v}(\boldsymbol{r}) = \rho(\boldsymbol{r})\boldsymbol{v}(\boldsymbol{r}).$$

Here we have used $n(\boldsymbol{r})q = \rho(\boldsymbol{r})$. The current density is to the magnetic field as the charge density is to the electric field. The units of $\boldsymbol{J}$ are $\mathrm{A\,m^{-2}}$. We can similarly define a surface current density, usually denoted $\boldsymbol{K}$ or $\boldsymbol{j}$. This will have units of $\mathrm{A\,m^{-1}}$. We can also define a line current density, usually denoted $\boldsymbol{I}$, which has units of A.

The total current, I, is the current density through a surface, A:

$$I = \int_A \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S}\,,$$

where $\mathrm{d}\boldsymbol{S}$ is a surface element normal to A.

Example 7.1. An insulating disc of radius R with uniform charge density, σ, is rotated about its axis with an angular velocity, $\boldsymbol{\omega} = \omega\boldsymbol{e_z}$. As a result the disc has a surface current density:

$$\boldsymbol{K} = \sigma\boldsymbol{v} = \sigma\boldsymbol{\omega}\boldsymbol{r} = \sigma\omega\boldsymbol{e_z} \times r\boldsymbol{e_\rho} = \sigma\omega r\boldsymbol{e_\varphi}.$$

Notice that this is a steady current even though the disc is moving faster towards the edge, hence scaling with r. The important thing for a steady current is that the current is constant at any one point in space, not that it has the same value at all points of space.

7.3 Conductivity

The **conductivity**, σ, is an intrinsic bulk property of a material. It relates the current density to the electric field:

$$\boldsymbol{J} = \sigma\boldsymbol{E}.$$

This is **Ohm's law**. This assumes that $\boldsymbol{J}$ and $\boldsymbol{E}$ are parallel. This is only the case if the material is isotropic, that is all directions within the material are equivalent. If this isn't the case then we need to use the conductivity tensor, σ_{ij}, instead.

A typical value of σ for a metal is $10^9\,\Omega^{-1}\,\mathrm{m}^{-1}$. We define the **resistivity** as

$$\rho = \frac{1}{\sigma}.$$

A typical value of ρ for an insulator is $10^{16}\,\Omega\,\mathrm{m}$. In a superconductor $\rho = 0$.

Suppose a wire of cross sectional area A has a homogenous current density and electric field. The current in the wire is

$$I = \int_A \boldsymbol{J}\cdot\mathrm{d}\boldsymbol{S} = \int_A J\,\mathrm{d}S = \int \sigma E\,\mathrm{d}S = E\sigma A.$$

The potential difference between two points on the wire a distance d apart is

$$\Delta V = Ed.$$

The current can then be written as

$$I = \frac{A\sigma}{d}\Delta V.$$

Rearranging we get

$$\Delta V = IR, \qquad \text{where} \qquad R = \frac{d\rho}{A}.$$

R here is what we define as the **resistance**. It has units of ohms, Ω. This equation is the more familiar form of Ohm's law.

7.4 Current Elements

A current element, $\mathrm{d}\boldsymbol{\mathcal{I}}$, is a vector defined by

$$\begin{aligned}\mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r}) &= \boldsymbol{J}(\boldsymbol{r})\,\mathrm{d}V\,,\\ \mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r}) &= \boldsymbol{K}(\boldsymbol{r})\,\mathrm{d}S\,,\\ \mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r}) &= \boldsymbol{I}(\boldsymbol{r})\,\mathrm{d}l\,.\end{aligned}$$

The units of a current element are $\mathrm{A\,m}$. Be careful, $\boldsymbol{K}\,\mathrm{d}S \neq K\,\mathrm{d}S$, the first points in the direction of the surface current density, which is along the surface, and the second is normal to the surface. On the other hand $\boldsymbol{I}\,\mathrm{d}l = I\,\mathrm{d}\boldsymbol{l}$ since a line current element always points along the line and so does a line element.

A current is a moving charge element. For a bulk current density we have

$$\mathrm{d}\boldsymbol{\mathcal{I}}\,(\boldsymbol{r}) = \boldsymbol{J}(\boldsymbol{r})\,\mathrm{d}V = \rho(\boldsymbol{r})\boldsymbol{v}(\boldsymbol{r})\,\mathrm{d}V = \boldsymbol{v}(\boldsymbol{r})\,\mathrm{d}q\,.$$

This means that we can work out the force, $\mathrm{d}\boldsymbol{F}$, on a current element, $\mathrm{d}\boldsymbol{\mathcal{I}}$, in a magnetic field, $\boldsymbol{B}$:

$$\mathrm{d}\boldsymbol{F} = \mathrm{d}q\,\boldsymbol{v}\times\boldsymbol{B} = \mathrm{d}\boldsymbol{\mathcal{I}}\times\boldsymbol{B}.$$

If the current comes from a bulk current density, $\mathrm{d}\boldsymbol{\mathcal{I}} = \boldsymbol{J}\,\mathrm{d}V$ then

$$\mathrm{d}\boldsymbol{F} = \boldsymbol{J}\times\boldsymbol{B}\,\mathrm{d}V\,.$$

7.5 Biot Savart Law

The Biot Savart law is an empirical law relating current elements to the magnetic field. It states that if a current element $\mathrm{d}\boldsymbol{\mathcal{I}}\,(\boldsymbol{r'})$ is at position $\boldsymbol{r'}$ then the resulting magnetic field at $\boldsymbol{r}$ is given by

$$\mathrm{d}\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r'})\times(\boldsymbol{r}-\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}.$$

The Biot Savart law is to magnetic fields as Coulomb's law is to electric fields. It states that charge elements create magnetic fields which shows that a steady current causes a constant magnetic field. Here μ_0 is a constant known as the permeability of free space, or the electric constant. It has a value of

$$\mu_0 = 4\pi\cdot 10^{-7}\,\mathrm{H\,m}^{-1} = 4\pi\cdot 10^{-7}\,\mathrm{N\,A}^{-2}.$$

Here H is a henry defined as $1\,\mathrm{H} = 1\,\mathrm{N\,m\,A^{-2}}$.

Since the Biot Savart law is linear in $\mathrm{d}\boldsymbol{\mathcal{I}}$ the superposition of the magnetic fields of many current elements to get the total magnetic fields holds so

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

For a bulk current density this becomes

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

7.6 Magnetic Force Between Currents

The force on the current element $\mathrm{d}\boldsymbol{\mathcal{I}_1}$ at position $\boldsymbol{r_1}$ due to a current element, $\mathrm{d}\boldsymbol{\mathcal{I}_2}$, at $\boldsymbol{r_2}$ is

$$\mathrm{d}\boldsymbol{F_1} = \mathrm{d}\boldsymbol{\mathcal{I}_1}\times\mathrm{d}\boldsymbol{B_2} = \frac{\mu_0}{4\pi r_{12}^2}\,\mathrm{d}\boldsymbol{\mathcal{I}_1}\times(\mathrm{d}\boldsymbol{\mathcal{I}_2}\times\hat{\boldsymbol{r}}_{\boldsymbol{12}}).$$

Here $\mathrm{d}\boldsymbol{B_2}$ is the magnetic field due to the second current element and $\boldsymbol{r_{12}} = \boldsymbol{r_1} - \boldsymbol{r_2}$. If the each current element, $\mathrm{d}\boldsymbol{\mathcal{I}_i}$ is due to a bulk current density $\boldsymbol{J_i}$ then this becomes

$$\mathrm{d}\boldsymbol{F_1} = \frac{\mu_0}{4\pi r_{12}^2}\boldsymbol{J_1}\times(\boldsymbol{J_2}\times\hat{\boldsymbol{r}}_{\boldsymbol{12}})\,\mathrm{d}^3r_1\,\mathrm{d}^3r_2\,.$$

Example 7.2. A long straight wire is aligned along the z-axis. It carries a current, I, in the positive z direction. What is the magnetic field strength at a point a distance r from the wire?

A current element is given by

$$\mathrm{d}\boldsymbol{\mathcal{I}} = I\,\mathrm{d}z\,\boldsymbol{e_z} = I\,\mathrm{d}\boldsymbol{r'}$$

where $\boldsymbol{r'}$ is the position of the current element along the wire. Choosing the origin to be in the same plane as the point at which we are evaluating the field we have

$$\boldsymbol{r} = \rho\boldsymbol{e_\rho}$$

for the position at which we want to know the field strength (note that ρ here is cylindrical coordinates, not charge density or resistivity). Applying the Biot Savart law we get

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{r}) &= \frac{\mu_0}{4\pi}\int\frac{\mathrm{d}\boldsymbol{\mathcal{I}}\times(\boldsymbol{r}-\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\\
&= \frac{\mu_0}{4\pi}\int\frac{I\boldsymbol{e_z}\times(\rho\boldsymbol{e_\rho}-r'\boldsymbol{e_z})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\,\mathrm{d}z\\
&= \frac{\mu_0}{4\pi}I\left[\int\frac{\boldsymbol{e_z}\times\rho\boldsymbol{e_\rho}}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\,\mathrm{d}z - \int\frac{\boldsymbol{e_z}\times r'\boldsymbol{e_z}}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\,\mathrm{d}z\right]\\
&= \frac{\mu_0 I}{4\pi}\rho\boldsymbol{e_\varphi}\int(\rho^2+z^2)^{-3/2}\,\mathrm{d}z\\
&= \frac{\mu_0 I}{2\pi\rho}\boldsymbol{e_\varphi}.
\end{aligned}$$

We looked up the integral in the last step over all $z\in\mathbb{R}$ and found it to be $2/\rho^2$.

---

8 Divergence and Curl of the Magnetic Field

8.1 Divergence of the Magnetic Field

The Biot Savart law for a bulk current density, $\boldsymbol{J}(\boldsymbol{r'})$, is

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

$$= -\frac{\mu_0}{4\pi} \int \boldsymbol{J}(\boldsymbol{r'}) \times \nabla_{\boldsymbol{r}} \left(\frac{1}{|\boldsymbol{r} - \boldsymbol{r'}|} \right) \mathrm{d}^3 r'$$

Here we use a notation where differential operators with a subscript $\boldsymbol{r}$ act only on the components of $\boldsymbol{r}$. We can take the divergence of this fairly easily if we recognise that $\boldsymbol{J}(\boldsymbol{r'})$ is constant with respect to $\boldsymbol{r}$ so $\nabla_{\boldsymbol{r}} \cdot \boldsymbol{J}(\boldsymbol{r'}) = 0$.

$$\begin{aligned}
\nabla_{\boldsymbol{r}} \cdot \boldsymbol{B}(\boldsymbol{r}) &= -\frac{\mu_0}{4\pi} \int \nabla_{\boldsymbol{r}} \cdot \left[\boldsymbol{J}(\boldsymbol{r'}) \times \nabla_{\boldsymbol{r}} \left(\frac{1}{|\boldsymbol{r} - \boldsymbol{r'}|} \right) \right] \mathrm{d}^3 r' \\
&= -\frac{\mu_0}{4\pi} \int \boldsymbol{J}(\boldsymbol{r'}) \cdot \left[\nabla_{\boldsymbol{r}} \times \nabla_{\boldsymbol{r}} \left(\frac{1}{|\boldsymbol{r} - \boldsymbol{r'}|} \right) \right] \mathrm{d}^3 r' \\
&= 0
\end{aligned}$$

Here we have used the identity that grad curl is zero. Hence

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

This is **Maxwell's second law**. It states that there are no sinks or sources of the magnetic field, there are no magnetic monopoles, there are no point sources of the magnetic field, there are no 'magnetic charges'. The magnetic field always forms closed loops.

Applying the divergence theorem we have

$$\int_V \nabla \cdot \boldsymbol{B}(\boldsymbol{r}) \, \mathrm{d}^3 r = \oint_A \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} \, .$$

Identifying the left hand side as an integral of zero we have

$$\oint_a \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0.$$

This is **Gauss' law for magnetic fields**.

8.2 Magnetic Dipoles

Maxwell's second law rules out the existence of magnetic monopoles. Therefore a magnetic dipole is the most elementary magnetic pole. It turns out that a magnetic dipole can be formed from a current loop.

Suppose we have a circular loop or radius a carrying current I. If we align the axis of the loop with the z axis so that the current is going in the $\boldsymbol{e_\varphi}$ direction then the magnetic field on the axis can be found from the Biot Savart law,

$$\mathrm{d}\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \frac{\mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r'}) \, (\boldsymbol{r} - \boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|^3}.$$

Here $\boldsymbol{r} = z\boldsymbol{e_z}$, $\boldsymbol{r'} = a\boldsymbol{e_\rho}$, and $\mathrm{d}\boldsymbol{\mathcal{I}} = I \, \mathrm{d}\ell \, \boldsymbol{e_\varphi}$. Hence

$$\begin{aligned}
\mathrm{d}\boldsymbol{\mathcal{I}} \times (\boldsymbol{r} - \boldsymbol{r'}) &= I \, \mathrm{d}\ell \, \boldsymbol{e_\varphi} \times (z\boldsymbol{e_z} - a\boldsymbol{e_\rho}) \\
&= I \, \mathrm{d}\ell \, (z\boldsymbol{e_\varphi} \times \boldsymbol{e_z} - a\boldsymbol{e_\varphi} \times \boldsymbol{e_\rho}) \\
&= I \, \mathrm{d}\ell \, (z\boldsymbol{e_\rho} + a\boldsymbol{e_z}).
\end{aligned}$$

Radial from diametrically opposed sides of the loop will cancel leaving a net field only in the $\boldsymbol{e_z}$ direction:

$$\mathrm{d}B_z = \frac{\mu_0}{4\pi} \frac{Ia \, \mathrm{d}\ell}{(a^2 + z^2)^{3/2}}$$

Integrating this we get

$$\begin{aligned}
B_z &= \frac{\mu_0}{4\pi} \frac{Ia}{(a^2 + z^2)^{3/2}} \oint \mathrm{d}\ell \\
&= \frac{\mu_0}{4\pi} \frac{Ia}{(a^2 + z^2)^{3/2}} 2\pi a \\
&= \frac{\mu_0 I a^2}{2(a^2 + z^2)^{3/2}}.
\end{aligned}$$

At $z = 0$ (i.e. the centre of the loop)

$$B_z = \frac{\mu_0 I a^2}{2a^3} = \frac{\mu_0 I}{2a}.$$

In the far field (i.e. $z \gg a$)

$$B_z \approx \frac{\mu_0 I a^2}{2z^3}.$$

It can be shown that off-axis field is

$$\boldsymbol{B_{\mathbf{dip}}}(\boldsymbol{r}) = \frac{\mu_0}{4\pi r^3}[3(\boldsymbol{m}\cdot\hat{\boldsymbol{r}})\hat{\boldsymbol{r}} - \boldsymbol{m}]$$

in the far field. Here $\boldsymbol{m}$ is the dipole moment defined as

$$\boldsymbol{m} = IA\boldsymbol{e_z}$$

where A is the area of a current loop, in the (x, y)-plane, and I is the current it carries. In this case $A = \pi a^2$ so

$$\boldsymbol{m} = I\pi a^2 \boldsymbol{e_z}.$$

More generally we can define the dipole moment as

$$\boldsymbol{m} = I\boldsymbol{A} = I\int_A \mathrm{d}\boldsymbol{S}$$

where $\boldsymbol{A}$ is the vector area of the loop defined by an integral over the entire loop, which may not be planar, and has surface elements $\mathrm{d}\boldsymbol{S}$.

In an external magnetic field, $\boldsymbol{B_{\mathbf{ext}}}$, a magnetic dipole has energy

$$U = -\boldsymbol{m}\cdot\boldsymbol{B_{\mathbf{ext}}}.$$

There is also a torque

$$\boldsymbol{\tau} = \boldsymbol{m}\times\boldsymbol{B_{\mathbf{ext}}}.$$

8.3 Curl of the Magnetic Field

Starting again from the Biot Savart law for a bulk current density and this time taking the curl we have

$$\nabla_{\boldsymbol{r}}\times\boldsymbol{B}(\boldsymbol{r}) = -\frac{\mu_0}{4\pi}\int \nabla_{\boldsymbol{r}}\times\left[\boldsymbol{J}(\boldsymbol{r'})\times\nabla_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\right]\mathrm{d}^3r'$$

Looking just at the integrand we have

$$\begin{aligned}\nabla_{\boldsymbol{r}}\times\left[\boldsymbol{J}(\boldsymbol{r'})\times\nabla_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\right] &= \boldsymbol{J}(\boldsymbol{r'})\nabla^2_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right) - (\boldsymbol{J}(\boldsymbol{r'})\cdot\nabla_{\boldsymbol{r}})\nabla_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\\ &= -4\pi\delta(\boldsymbol{r}-\boldsymbol{r'})\boldsymbol{J}(\boldsymbol{r'}) - (\boldsymbol{J}(\boldsymbol{r'})\cdot\nabla_{\boldsymbol{r}})\nabla_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right).\end{aligned}$$

Hence the curl of the magnetic field is

$$\nabla_{\boldsymbol{r}}\times\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int 4\pi\delta(\boldsymbol{r}-\boldsymbol{r'})\boldsymbol{J}(\boldsymbol{r'})\,\mathrm{d}^3r' + \frac{\mu_0}{4\pi}\int(\boldsymbol{J}(\boldsymbol{r'})\cdot\nabla_{\boldsymbol{r}})\nabla_{\boldsymbol{r}}\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\mathrm{d}^3r'$$

The second term can be shown to be zero by writing it as a boundary integral at infinity which must vanish to have finite energy

$$\begin{aligned}&= \frac{\mu_0}{4\pi}\int 4\pi\delta(\boldsymbol{r}-\boldsymbol{r'})\boldsymbol{J}(\boldsymbol{r'})\,\mathrm{d}^3r'\\ &= \mu_0\int\boldsymbol{J}(\boldsymbol{r'})\delta(\boldsymbol{r}-\boldsymbol{r'})\,\mathrm{d}^3r'\\ &= \mu_0\boldsymbol{J}(\boldsymbol{r}).\end{aligned}$$

This is **Maxwell's fourth law**. Applying Stokes' theorem we have

$$\int_S \nabla\times\boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S} = \oint_C \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{l}\,.$$

Looking at the left hand side we have

$$\int_S \nabla \times \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \mu_0 \int_S \nabla \times \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} = \mu_0 I.$$

Hence

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 I.$$

This is **Ampere's law**.

Example 8.1. Consider an infinite wire of radius a aligned along the $\boldsymbol{e_z}$ direction carrying current density $\boldsymbol{J} \propto \boldsymbol{e_z}$. We must have that $B_z = 0$ as $\boldsymbol{J}$ and $\boldsymbol{B}$ must be perpendicular. Similarly we must have that $B_\rho = 0$ as if there where radial components then $\nabla \cdot \boldsymbol{B}$ would be non-zero violating the second of Maxwell's equations. Therefore $\boldsymbol{B} \propto \boldsymbol{e_\varphi}$. Also B_φ cannot depend on z as we are free to place the origin anywhere along the wire and it cannot depend on φ as we are free to define any angle around the wire as $\varphi = 0$. Thus we are left with

$$\boldsymbol{B} = B_\varphi(\rho)\boldsymbol{e_\varphi}.$$

For $\rho > a$ we have

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}l = B_\varphi 2\pi\rho = \mu_0 I_{\text{enc}} = \mu_0 J \pi a^2.$$

Hence

$$B_\varphi = \frac{\mu_0 J}{2}\frac{a^2}{\rho}.$$

For $\rho < a$ we still have

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = B_\varphi 2\pi\rho$$

but now

$$\mu_0 I_{\text{enc}} = \mu_0 J \pi \rho^2$$

so

$$B_\varphi = \frac{\mu_0 J}{2}\rho.$$

This is shown in figure 8.1.

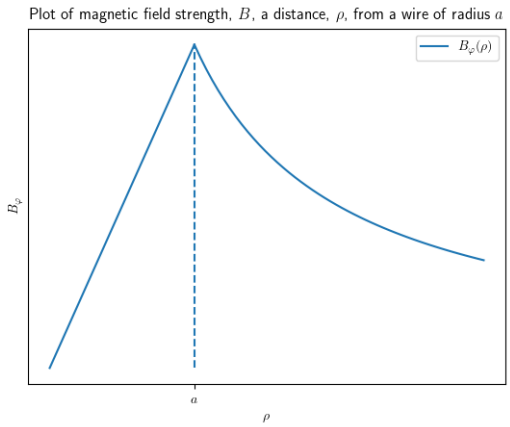


Figure 8.1: The magnetic field strength a distance ρ from a wire.

---

9 Ampère's Law and Vector Potentials

9.1 Applications of Ampère's Law

Ampère's Law states

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_A \boldsymbol{J} \cdot \mathrm{d}S = \mu_0 I_{\text{enc}}.$$

Here $\boldsymbol{B}$ is the magnetic field due to a current density $\boldsymbol{J}$. C is some closed curve, called an Ampèrian loop, bounding a surface A which has surface normals $\mathrm{d}\boldsymbol{S}$. I_{enc} is the current that passes through the surface. The integration is performed along the curve in such a way that the surface normals and direction of integration agree with the right hand grip rule. Symmetry permitting Ampère's law is usually the fastest way to calculate the magnetic field for a given current density. We look for cases where the Ampèrian loop is either parallel or perpendicular to the magnetic field. The cases where this is possible and the Ampèrian loops to use are

- Infinite straight line – coaxial circles (see example 8.1)
- Infinite plane – rectangular loop
- Infinite solenoid – rectangular loop
- Toroid – circle

9.1.1 Infinite Slab of Current

An infinite slab of thickness d has a current density, $\boldsymbol{J}$. Define the x direction to be the direction of this current density and the y direction to be in the slab but perpendicular to the current density and the z direction to be normal to the slab.

The magnetic field cannot have an x component as the current density must be perpendicular to the magnetic field. The magnetic field cannot have a z component as if it did we could place a Gaussian surface with a section of the slab in it and the integral over it would be non-zero meaning that we would have non-vanishing divergence. This is forbidden by Maxwell's second law. This leaves us with a magnetic field that can only be in the y direction.

If we choose to put the origin a distance $d/2$ into the slab then since we can pick any point at this depth in the slab to place the origin we must have that B_y is independent of x and y and so $B_y = B_y(z)$.

We choose an Ampèrian loop that is rectangular with two sides in the z direction and two in the y direction. The field is perpendicular to the sides in the z direction. The field is aligned with the sides in the y direction. Say that the length of these sides is b, then

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = 2b|B_y| = \mu_0 I_{\text{enc}} = \mu_0 bdJ$$

$$\implies |B_y| = \frac{1}{2}\mu_0 Jd.$$

We know that $\mathrm{d}\boldsymbol{\mathcal{I}} \propto \boldsymbol{e_x}$ so for $\boldsymbol{r} \propto \boldsymbol{e_z}$ we have, by the Biot Savart law, that

$$\boldsymbol{B} \propto \boldsymbol{J} \times \boldsymbol{r} \propto \boldsymbol{e_x} \times \boldsymbol{e_z} = -\boldsymbol{e_y}$$

This means that, outside of the slab, the magnetic field is

$$\boldsymbol{B} = \begin{cases} -\frac{1}{2}\mu_0 Jd\boldsymbol{e_y}, & z > \frac{1}{2}d, \\ +\frac{1}{2}\mu_0 Jd\boldsymbol{e_y}, & z < -\frac{1}{2}d. \end{cases}$$

If we want to know the magnetic field inside the slab then the set up for the Ampérian loop is similar but now one of the sides in the y direction should be inside the slab. We then use the field outside the loop that we already know to integrate along this loop. We find that

$$B_{y,\text{in}} = -\mu_0 Jz.$$

9.2 Toroid

A wire is curled up into a solenoid with n coils per unit length and radius R. The solenoid is the curled up into a toroid of radius $a > R$. Figure 9.4 shows a toroid (normal lines) with three possible Ampérian loops (dashed lines). The easiest loop to consider is the one that is entirely inside the hole of the toroid. Clearly there is no current enclosed in this so the magnetic field in the hole must be zero.

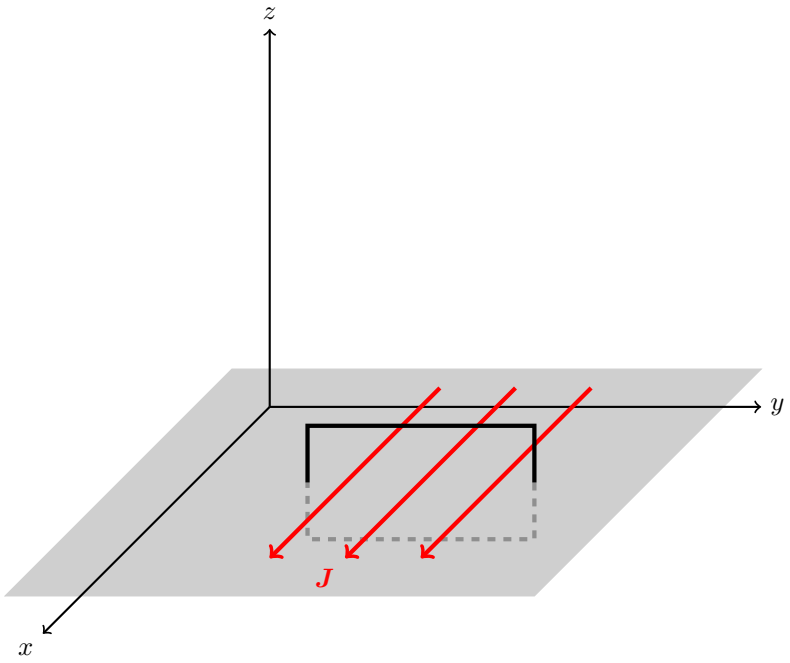


Figure 9.1: Infinite slab of current

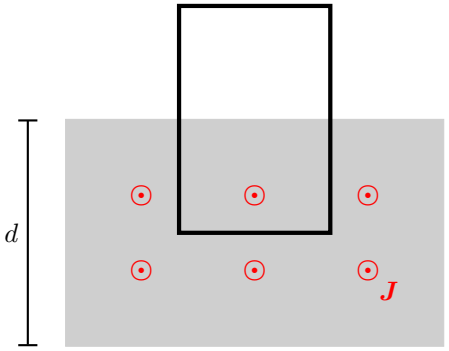


Figure 9.2: Finding the magnetic field inside a slab of current

Next we consider the outer most loop. At first it may seem like current does pass trough this but we must remember that the direction is important and therefore the *net* current is zero as there is just as much current in both directions since the current is forming loops in the coil. Therefore outside of the toroid the magnetic field must be zero. The only Ampérian loop with a net current enclosed is the one that is in the toroid.

All of the current element that pass through it are in the (ρ, z)-plane, this means that $\boldsymbol{B}$ can only be in the $\boldsymbol{e}_{\boldsymbol{\varphi}}$ direction. If the loop has radius ρ then we find that

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = 2\pi\rho B_\varphi = \mu_0 I_{\text{enc}} = \mu_0 n 2\pi a I$$

where I is the current carried by the wire. Rearranging this we get

$$B_\varphi = \frac{\mu_0 n I a}{\rho}.$$

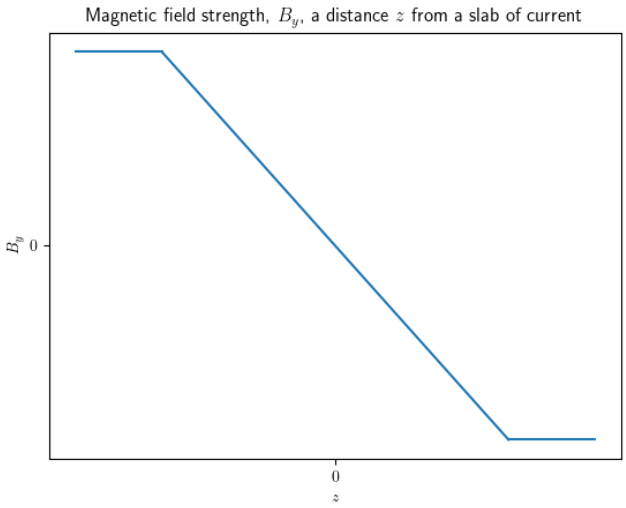


Figure 9.3: The magnetic field strength, B_y, a distance, z, from a slab carrying current density, $\boldsymbol{J}$.

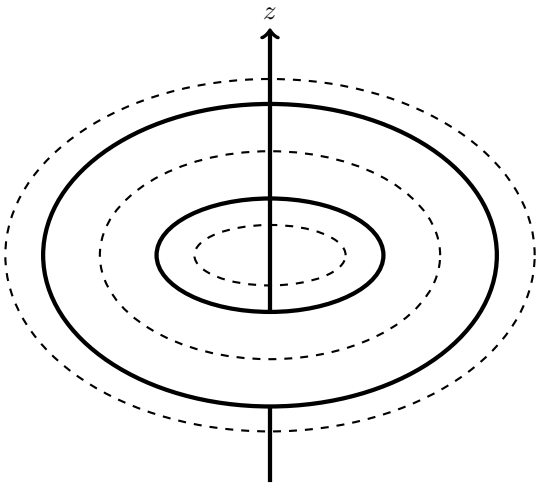


Figure 9.4: Finding the magnetic field inside a toroid

9.3 Magnetic Vector Potential

Theorem 2

The following are equivalent for a vector field, $\boldsymbol{B}\colon \mathbb{R}^3 \to \mathbb{R}^3$:

1. $\nabla \cdot \boldsymbol{B} = 0$, in which case we call the field **solenoidal**.
2. There exists a vector field, $\boldsymbol{A}\colon \mathbb{R}^3 \to \mathbb{R}^3$ such that $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, in which case we call $\boldsymbol{A}$ the **vector potential** of $\boldsymbol{B}$.
3. The surface integral
$$\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}$$
is independent of the shape of the surface, S, for a given boundary curve. In particular
$$\oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0$$
for any closed surface S.

Maxwell's second equation tells us that $\nabla \cdot \boldsymbol{B} = 0$ so there must exist $\boldsymbol{A}$ such that

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}.$$

We call $\boldsymbol{A}$ the **magnetic vector potential**. Compare this to the scalar electric potential, φ for a stationary electric field where we have

$$\boldsymbol{E} = -\nabla\varphi.$$

9.3.1 Poisson's Equation for the Vector Potential

Ampère's law is

$$\nabla \times \boldsymbol{B} = \nabla \times (\nabla \times \boldsymbol{A}) = \mu_0 \boldsymbol{J}.$$

Applying a vector calculus identity we have

$$-\nabla \times (\nabla \times \boldsymbol{A}) = \nabla^2 \boldsymbol{A} - \nabla(\nabla \cdot \boldsymbol{A}) = -\mu_0 \boldsymbol{J}. \tag{9.1}$$

This is the most general form that we can find making no assumptions about the potential. However we have a **gauge freedom** that allows us to modify the potentials in a certain way without effecting the physics. That is the modified potentials given the same electric and magnetic fields. We have already seen this for the electric field. If we have two potentials, $V(\boldsymbol{r})$ and $\tilde{V}(\boldsymbol{r}) = V(\boldsymbol{r}) + C$ where $C \in \mathbb{R}$ then

$$-\nabla V(\boldsymbol{r}) = \boldsymbol{E}$$

by definition but also

$$-\nabla\tilde{V} = -\nabla(V + C) = -\nabla V - \nabla C = -\nabla V = \boldsymbol{E}.$$

Thus the field is unchanged by the addition of a constant to the potential. Similarly with the vector potential if we have two potentials, $\boldsymbol{A}(\boldsymbol{r})$ and $\tilde{\boldsymbol{A}}(\boldsymbol{r}) = \boldsymbol{A}(\boldsymbol{r}) + \nabla\varphi(\boldsymbol{r})$ where $\varphi\colon \mathbb{R}^3 \to \mathbb{R}$ then

$$\nabla \times \boldsymbol{A} = \boldsymbol{B}$$

by definition but also

$$\nabla \times \tilde{\boldsymbol{A}} = \nabla \times (\boldsymbol{A} + \nabla\varphi) = \nabla \times \boldsymbol{A} + \nabla \times \nabla\boldsymbol{A} = \nabla \times \boldsymbol{A} = \boldsymbol{B}.$$

Thus the field is unchanged by the addition of a gradient field to the potential.

We use this gauge freedom to select a gauge (a condition on the potential) that simplifies equations. One of the most common gauges to choose is the **Coulomb gauge**:

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

For all magnetic fields, $\boldsymbol{B}$, it is always possible to define a potential, $\boldsymbol{A}$, such that

$$\nabla \times \boldsymbol{A} = \boldsymbol{B}, \qquad \text{and} \qquad \nabla \cdot \boldsymbol{A} = 0.$$

Given a potential $\boldsymbol{A'}$ that satisfies the first of these condition then $\boldsymbol{A} = \boldsymbol{A'} + \nabla\varphi$ also satisfies the first condition. We can also require that

$$\nabla \cdot \boldsymbol{A} = \nabla \cdot \boldsymbol{A'} + \nabla \cdot \nabla\varphi = \nabla \cdot \boldsymbol{A} + \nabla^2\varphi = 0 \implies \nabla^2\varphi = -\nabla \cdot \boldsymbol{A}.$$

This is simply Poisson's equation (PE) for φ. We know that this has a solution

$$\varphi(\boldsymbol{r}) = \frac{1}{4\pi} \int \frac{\nabla \cdot \boldsymbol{A}(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|^3} \, \mathrm{d}^3 r' \, .$$

In the Coulomb gauge equation 9.1 becomes

$$\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}.$$

This is actually three equations, one for each component:

$$\nabla^2 A_i - -\mu_0 J_i.$$

So we have reduced equation 9.1 to PE by carefully choosing a gauge. This means that we can use all of the methods we have already developed for PE. In particular we have the explicit solution

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

For example using this we can find that the magnetic dipole has a potential of

$$\boldsymbol{A} = \frac{u_0}{4\pi} \frac{\boldsymbol{m} \times \boldsymbol{r}}{r^3}.$$

9.4 Summary of Statics

The electric field, $\boldsymbol{E}$, and magnetic field, $\boldsymbol{B}$, are defined such that for a charge, q, moving at velocity $\boldsymbol{v}$, the force on the charge is given by the Lorentz force law:

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}).$$

9.4.1 Electrostatics

A static electric field is created by a stationary charge distribution, ρ, meaning $\partial_t \rho = 0$. The empirical law from which we derive everything else is Coulomb's law:

$$\boldsymbol{F} = \frac{1}{4\pi\varepsilon_0}\frac{qQ}{r^2}.$$

From this we can derive Maxwell's first and third laws, which are

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0} \iff \oint_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \frac{Q_{\text{enc}}}{\varepsilon_0}, \tag{MI}$$

and

$$\nabla \times \boldsymbol{E} = \boldsymbol{0} \iff \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = 0. \tag{MIII static}$$

This last equation allows us to define an electrostatic potential, V, such that

$$\boldsymbol{E} = -\nabla V.$$

Combining this with Maxwell's first equation we get PE:

$$\nabla^2 V = -\frac{\rho}{\varepsilon}.$$

9.4.2 Magnetostatics

A static magnetic field is created by a steady current density, $\boldsymbol{J}$, meaning $\partial_t \boldsymbol{J} = \boldsymbol{0}$. The empirical law from which we derive everything else is the Biot Savart law:

$$\mathrm{d}\boldsymbol{B}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\mathrm{d}\boldsymbol{\mathcal{I}}(\boldsymbol{r'}) \times (\boldsymbol{r} - \boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|^3}.$$

From this we can derive Maxwell's second and fourth laws, which are

$$\nabla \cdot \boldsymbol{B} = 0 \iff \oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0, \tag{MII}$$

and

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J} \iff \oint_C \boldsymbol{B} \cdot \mathrm{d}l = \mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S}\,. \tag{MIV static}$$

The first of these allows us to define a vector potential, $\boldsymbol{A}$, such that

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}.$$

Combining both of these equations, and working in the Coulomb gauge, $\nabla \cdot \boldsymbol{A} = 0$, we get PE:

$$\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}.$$

Part III

Electromagnetism

10 Electromotance and Faraday's law

10.1 Sources of Steady Current

Consider a conducting loop where Ohm's law, $\boldsymbol{J} = \sigma \boldsymbol{E}$, holds. Then integrating along the loop with Stokes' theorem we have

$$\oint_C \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{l} = \sigma \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = \int_S \nabla \times \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}\,.$$

Here C is the loop and S is the disc it encloses. So far we have seen that $\nabla \times \boldsymbol{E} = \boldsymbol{0}$ and therefore we must have that $\boldsymbol{J} = \boldsymbol{0}$. However this is clearly not true so $\nabla \times \boldsymbol{E} = \boldsymbol{0}$ must not always be correct.

Similarly

$$\oint_C \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{l} = \sigma \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\sigma \oint_C \nabla V \cdot \mathrm{d}\boldsymbol{l} = -\sigma[V(2\pi) - V(0)] = 0.$$

Again this seems to imply that $\boldsymbol{J} = \boldsymbol{0}$. The only assumption made here was that $\boldsymbol{E} = -\nabla V$ which is only valid if $\nabla \times \boldsymbol{E} = \boldsymbol{0}$. So it seems that $\boldsymbol{E} = -\nabla V$ also doesn't hold all the time. The question is what do we have to do to fix these equations so that they apply all the time.

Imagine inserting a battery between the points B and A in the loop such that it creates a potential difference of ΔV. If we then integrate over the whole loop but *not* the battery then we have

$$\Delta V = \mathcal{E} = \int_A^B \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l}\,.$$

Here $\mathcal{E}$ is known as the **electromotance**, or **electromotive force (emf)**. If we assume a unit test charge then we see that

$$\mathcal{E} = \oint \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{l}\,.$$

10.2 Induced EMF

We introduce here the **magnetic flux**,

$$\Phi_B = \int_A \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \oint_C \nabla \times \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \oint_C \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{l}\,.$$

Suppose a loop of wire is placed partly in a magnetic field as shown in figure 10.1. The loop is dragged

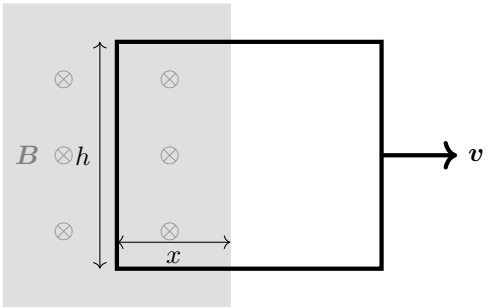


Figure 10.1: Wire loop moving through a magnetic field inducing an emf.

out of the field with a velocity $\boldsymbol{v}$. The force on a charge carrier in the loop due to the magnetic field is given by

$$\boldsymbol{F_M} = q\boldsymbol{v} \times \boldsymbol{B}.$$

The emf is

$$\mathcal{E} = \oint \boldsymbol{F_M} \cdot \mathrm{d}\boldsymbol{l} = hvB.$$

Here we have used that the force along the top and bottom parts of the loop are in opposite directions so cancel, meaning we only need to consider the contribution to the emf of the force along the vertical section in the field.

We can also calculate the magnetic flux:

$$\Phi_B = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = Bhx$$

where we define the surface normal so that the right hand rule with the current direction is satisfied. The flux is clearly not constant as the area of the loop that is in the field is changing:

$$\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}Bhx = Bh\frac{\mathrm{d}x}{\mathrm{d}t} = -Bhx = -\mathcal{E}.$$

Note the minus sign as the area is decreasing. In fact the statement,

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t},$$

holds in general. This is called **Faraday's law of induction**. We say that this emf is an induced emf as it doesn't come from a normal source of emf but rather the interaction of electric and magnetic fields.

10.3 Faraday's Law

Experimentally Faraday showed that altering Φ_B by any of the following:

1. moving the surface,
2. moving the field region,
3. varying the field strength,

produced an induced emf in accordance with Faraday's law of induction.

Note that in the last two of these the charges don't move so $\boldsymbol{v} = 0$ meaning $\boldsymbol{F_M} = \boldsymbol{0}$. Assuming that the Lorentz force law is correct there must be an induced electric field which creates a current.

Faraday;s law has a differential form:

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} \implies \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}\,.$$

Here C is the loop, S is the surface it encloses, and we have simply replaced $\mathcal{E}$ and Φ_B with their definitions. We assume that the surface is constant so

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \int_S \partial_t \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}\,.$$

We can also use Stokes' law to get

$$\oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = \int_S \nabla \times \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}\,.$$

Hence

$$\int_S \nabla \times \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = -\int_S \partial_t \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}\,.$$

Since both integrals are over the same area we must have that

$$\nabla \times \boldsymbol{E} = -\partial_t.$$

We see that $\nabla \times \boldsymbol{E} = \boldsymbol{0}$ only holds for static magnetic fields. This is the full version of this law and is Maxwell's third equation.

$$\mathcal{E} = \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} \iff \nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}. \tag{MIII}$$

10.4 Connection to the Magnetic Vector Potential

If we combine $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$ and $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ then we have

$$\nabla \times \boldsymbol{E} = -\partial_t(\nabla \times \boldsymbol{A}) = -\nabla \times (\partial_t \boldsymbol{A})$$

which means

$$\nabla \times (\boldsymbol{E} - \partial_t \boldsymbol{A}) = \boldsymbol{0}.$$

This means that there must exist a scalar field, V, such that

$$\boldsymbol{E} - \partial_t = -\nabla V.$$

Rearranging this we have

$$\boldsymbol{E} = -\nabla V - \partial_t \boldsymbol{A}.$$

This is the most general way to write $\boldsymbol{E}$ in terms of potentials, it reduces to $\boldsymbol{E} = -\nabla V$ if $\boldsymbol{B}$ is time independent. We have now fixed the two equations that we showed weren't correct in section 10.1. We see that there are two sources of the electric field:

- A stationary charge distribution, ρ, which gives a contribution $-\nabla V$.
- A time dependent magnetic field, $\boldsymbol{B}$, which gives a contribution $-\partial_t \boldsymbol{A}$.

10.5 Lenz's Law

Lenz's law is a rule of thumb that allows us to decide the sign of the flux. It states that the induced emf, $\mathcal{E}$, acts in a way as to oppose the change that created it. For example in the case of removing a wire from a magnetic field the emf causes a current in the wire in a direction such that the force due to the magnetic field's interaction with this current is in the opposite direction to the direction which the wire is being pulled. This is useful as it is often not that easy to see what the direction of the flux is and hence the direction of the induced emf and induced current.

11 Induction

11.1 Induction Examples

11.1.1 AC Generator

A current loop, initially the (x, z)-plane, with area A is placed in a uniform magnetic field, $\boldsymbol{B}$. The loop is rotated at an angular velocity $\boldsymbol{\omega} = \omega \boldsymbol{e_z}$. Since the loop is rotating the flux through the loop is time dependent:

$$\Phi_B = \int_A \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = B\cos(\omega t) \int_A \mathrm{d}S = AB\cos(\omega t).$$

Therefore there is an induced emf:

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = AB\omega \sin(\omega t).$$

The induced current will be an AC current with a frequency ω. The current is $\pi/2$ out of phase with the flux in a way that means that the peak current occurs when the flux is zero.

11.1.2 Rotating Disc of Charge

An insulating disc of radius a with a uniform surface charge density, σ, is rotated around its axis. There is a uniform magnetic field parallel to the axis of the disc. The magnetic force on a charge element, $\mathrm{d}q$, on the disc at a radius r, is

$$\mathrm{d}\boldsymbol{F} = \mathrm{d}q\, vB\boldsymbol{e_\rho} = \mathrm{d}q\, r\omega B\boldsymbol{e_\rho}.$$

This magnetic force is equivalent to a radial electric field:

$$\boldsymbol{E'} = \frac{1}{q}\boldsymbol{F} = r\omega B\boldsymbol{e_\rho}.$$

The induced emf between the centre and edge of the disc is then

$$\mathcal{E} = \int_0^a \boldsymbol{E'} \cdot \boldsymbol{e_\rho} \, \mathrm{d}\rho = \frac{\omega B a^2}{2}.$$

This acts outwards to move the charge to the outside of the disc. If the disc were a conductor then this would actually happen.

While the flux rule,

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t},$$

is still valid there is not a useful current loop to compute the flux through so it isn't of much help in this problem.

11.2 Mutual Inductance

Two current loops, C_1 and C_2 are carrying currents I_1 and I_2 in directions $\mathrm{d}\boldsymbol{l_1}$ and $\mathrm{d}\boldsymbol{l_2}$ respectively at positions $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$. Current I_1 creates a magnetic field, $\boldsymbol{B_1}$, which passes through loop 2. Therefore a current is induced in loop 2. The induced flux in loop 2 is

$$\Phi_2 = \int_{S_2} \boldsymbol{B_1} \cdot \mathrm{d}\boldsymbol{S_2} = M_{21} I_1$$

where we have used that $|\boldsymbol{B_1}| \propto I_1$ and M_{21} is just a constant of proportionality called the **mutual inductance**. If we substitute the vector potential, $\boldsymbol{A_1}$, such that $\boldsymbol{B_1} = \nabla \times \boldsymbol{A_1}$, we get

$$\begin{aligned}
\Phi_2 &= \int_{S_2} \boldsymbol{B_1} \cdot \mathrm{d}\boldsymbol{S_2} \\
&= \int_{S_2} (\nabla \times \boldsymbol{A_1}) \cdot \mathrm{d}\boldsymbol{S_2} \\
&= \oint_{C_2} \boldsymbol{A_1} \cdot \mathrm{d}\boldsymbol{l_1} \\
&= \frac{\mu_0 I_1}{4\pi} \oint_{C_2} \oint_{C_1} \frac{\mathrm{d}\boldsymbol{l_1} \cdot \mathrm{d}\boldsymbol{l_2}}{|\boldsymbol{r_1} - \boldsymbol{r_2}|}.
\end{aligned}$$

In the last step we have used the explicit solution to PE, $\nabla^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}$. This gives us that

$$M_{21} = \frac{\mu_0}{4\pi} \oint_{C_2} \oint_{C_1} \frac{\mathrm{d}\boldsymbol{l_1} \cdot \mathrm{d}\boldsymbol{l_2}}{|\boldsymbol{r_1} - \boldsymbol{r_2}|}.$$

This is called the Neumann formula. It isn't that useful, as it is pretty hard to calculate for arbitrary geometries, it is important to note that it is symmetric in 1 and 2, that is if we perform the equivalent calculation but completely swap the two loops we will get the same answer so

$$M_{21} = M_{12} = M.$$

The flux through loop 1 due to a current I in loop 2 is the same as the flux through loop 2 due to a current I in loop 1.

The induced emf in loop 2 due to current I_2 is

$$\mathcal{E}_2 = -\frac{\mathrm{d}\Phi_2}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}[M I_1] = -M\frac{\mathrm{d}I_1}{\mathrm{d}t}.$$

We can then use Lenz's law to calculate the direction of the current noting that it must oppose any change in I_1.

One application of this is a spark plug in a car. In this loop 1 is wrapped around a core a few times and loop 2 is wrapped around the same core many many times. This leads to a very large value of M and so a small current in loop 1 can create a very large current in loop 2 which is large enough to spark across an air gap and ignite the fuel.

11.3 Self Inductance

If a loop carries a current then it generates a magnetic field. This magnetic field creates a flux through the loop. Therefore if the current is time dependent there will be an induced emf and induced current in the loop that acts to oppose the change in the initial current. The emf will be

$$\mathcal{E} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = -L\frac{\mathrm{d}I}{\mathrm{d}t},$$

where we have defined L, the **self inductance**, to be such that $\Phi_B = LI$.

For example a solenoid of length ℓ and radius a with n loops per unit length has an interior magnetic field of

$$\boldsymbol{B} = B_z\boldsymbol{e_z} = \mu_0 nI\boldsymbol{e_z}.$$

Each loop is approximately a closed circle of area $A = \pi a^2$ and so the total flux through all $n\ell$ loops is

$$\Phi_B = \int \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S} = An\ell B_z = \mu_0 I\pi a^2 n^2\ell.$$

We can then identify the self inductance as

$$L = \mu_0\pi a^2 n^2\ell = \mu_0 n^2 V_s$$

where $V_s = \pi a^2\ell$ is the volume of the solenoid.

11.4 Electronics

This section is non-examinable

We have introduced three material properties which all relate a voltage, ΔV, to the charge, Q, in a different way:

- Resistance, R:
$$\Delta V \sim I \sim \frac{\mathrm{d}Q}{\mathrm{d}t}.$$
- Capacitance, C:
$$\Delta V \sim Q.$$
- Inductance, L:
$$\Delta V \sim \frac{\mathrm{d}I}{\mathrm{d}t} \sim \frac{\mathrm{d}^2Q}{\mathrm{d}t^2}.$$

By constructing a circuit with two of these in series we can model differential equations. Since they are in series the voltage in all components will be the same. We start with a resistor and capacitor. From the fact that the voltages are equal we have that

$$\frac{\mathrm{d}Q}{\mathrm{d}t} \sim Q \implies Q = Q_0 e^{-\alpha t}$$

for some constants Q_0 and α. If we swap the capacitor for an inductor then we will have

$$\frac{\mathrm{d}I}{\mathrm{d}t} \sim I \implies I = I_0 e^{-\beta t}$$

for some constants I_0 and β. Finally if we have a capacitor and inductor in series then

$$\frac{\mathrm{d}^2Q}{\mathrm{d}t^2} \sim Q \implies Q = Q_0'\sin(\omega t + \varphi) \implies I = Q_0'\omega\cos(\omega t + \varphi)$$

for some constants Q_0', ω, and φ.

In the first case the charge on the capacitor decays away exponentially. In the second case the current that is initially induced by the inductor decays away exponentially. In the last case if initially we start with a charged capacitor and with no current ($\varphi = \pi/2$) then as the capacitor discharges a current is is induced and this charges the capacitor and we get an oscillating charge and current.

11.5 Magnetic Energy in Inductors

In electrostatics the electrostatic energy is due to the work done to create a charge distribution, done against Coulomb repulsion. Similarly in magnetostatics the magnetostatic energy is due to the work done to create a steady current, done against the induced emf which opposes the creation of any current.

If we try to create a current, $\mathrm{d}I$, in a loop in time $\mathrm{d}t$ we need to do work against the induced emf, $\mathcal{E}$:

$$\mathrm{d}U_M = -\mathcal{E} I \,\mathrm{d}t = LI\frac{\mathrm{d}I}{\mathrm{d}t}\,\mathrm{d}t = LI\,\mathrm{d}I\,,$$

where L is the self inductance of the loop and we have used that

$$-\mathcal{E} = \frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}[LI] = LI\frac{\mathrm{d}I}{\mathrm{d}t}.$$

Integrating the energy with time we get

$$U_M = \frac{1}{2}LI^2.$$

Compare this to the equivalent statement in electrostatics that

$$U_E = \frac{1}{2}\frac{Q^2}{C}.$$

For two loops with inductances L_1 and L_2 carrying currents I_1 and I_2 with mutual inductance M then the energy becomes

$$U_M = \frac{1}{2}L_1I_1^2 + \frac{1}{2}L_2I_2^2 + MI_1I_2.$$

We can generalise the energy to any current arrangement:

$$\begin{aligned} U_M = \frac{1}{2}LI^2 & \\ &= \frac{1}{2}\Phi_B I \\ &= \frac{1}{2}I\oint_C \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{l} \\ &= \frac{1}{2}\oint_C \boldsymbol{A}\cdot\boldsymbol{I} \end{aligned}$$

for a single loop. Notice that $\oint_C \boldsymbol{A}\cdot\boldsymbol{I}$ is effectively a volume integral over all space of $\boldsymbol{J}$ since this is only non-zero where there is a current, which in the case of a single loop reduces to a contour integral. Hence

$$U_M = \frac{1}{2}\int \boldsymbol{A}\cdot\boldsymbol{J}\,\mathrm{d}V\,.$$

Where the integral is performed over all space. Using Ampère's law we have

$$\begin{aligned} \mu_0\boldsymbol{A}\cdot\boldsymbol{J} &= \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B}) \\ &= |\boldsymbol{B}|^2 - \nabla\cdot(\boldsymbol{A}\times\boldsymbol{B}) \end{aligned}$$

where we have used the vector identity

$$\begin{aligned} \nabla\cdot(\boldsymbol{A}\times\boldsymbol{B}) &= \boldsymbol{B}\cdot\underbrace{(\nabla\times\boldsymbol{A})}_{=\boldsymbol{B}} - \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B}) \\ &= \boldsymbol{B}\cdot\boldsymbol{B} - \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B}) \\ &= |\boldsymbol{B}|^2 - \boldsymbol{A}\cdot(\nabla\times\boldsymbol{B}). \end{aligned}$$

Hence the energy is

$$U_M = \frac{1}{2}\boldsymbol{A}\cdot\boldsymbol{J}\,\mathrm{d}V = \frac{1}{2\mu_0}\int |\boldsymbol{B}|^2\,\mathrm{d}V - \frac{1}{2\mu_0}\int \nabla\cdot(\boldsymbol{A}\times\boldsymbol{B})\,\mathrm{d}V\,.$$

We can write this last integral as a surface integral using the divergence theorem:

$$\int \nabla \cdot (\boldsymbol{A} \times \boldsymbol{B}) \, \mathrm{d}V = \oint_S (\boldsymbol{A} \times \boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{S} \, .$$

If $A \sim 1/r$ and $B \sim 1/r^2$, as is the case for steady, finite, localised currents, then $|\boldsymbol{A} \times \boldsymbol{B}| \sim 1/r^3$. The surface area, $S \sim r^2$. Since the volume integral is over all space the surface integral is over a boundary at infinity and so

$$\int (\boldsymbol{A} \times \boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{S} \sim \frac{1}{r^3} r^2 = \frac{1}{r}$$

which becomes zero as $r \to \infty$. Hence the energy is

$$U_M = \frac{1}{2\mu_0} \int |\boldsymbol{B}|^2 \, \mathrm{d}V \, .$$

We can define an energy density, u_M, such that

$$u_m = \frac{|\boldsymbol{B}|^2}{2\mu_0} \implies U_M = \int u_m \, \mathrm{d}V \, .$$

Compare this to the electrostatic case

$$u_E = \frac{1}{2} \varepsilon |\boldsymbol{E}|^2 .$$

12 Displacement Current

12.1 The Continuity Equation

The global electric charge is conserved. This means that the only way that the charge contained in a volume, V, can change is if charge enters or leaves the volume. In the case that this happens there will be a current $\boldsymbol{J}$ associated with the charge moving. The charge enters the volume at a rate of

$$\frac{\partial Q}{\partial t} = - \oint_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} \, .$$

The minus sign accounts for the fact the surface normal, $\mathrm{d}\boldsymbol{S}$, points out of the volume so $\boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S}$ is the charge leaving the volume through the surface $\mathrm{d}\boldsymbol{S}$, on the other hand ∂_t is the rate of increase of charge. Substituting in the charge density and applying the divergence theorem we have

$$-\frac{\partial}{\partial t} \int_V \rho \, \mathrm{d}V = \int_V \nabla \cdot \boldsymbol{J} \, \mathrm{d}V \, .$$

Since this holds for all volumes we must have

$$-\frac{\partial \rho}{\partial t} = -\nabla \cdot \boldsymbol{J} .$$

This is the **continuity equation**. It actually applies to any conserved quantity with an associated density and current. For example if ρ is mass density and $\boldsymbol{J}$ is mass current which describes how mass is moving, for example it may describe fluid flow, then the continuity equation applies. Another important application is in quantum mechanics where ρ is the probability density and $\boldsymbol{J}$ is a probability current which describes how the probability of different states changes with time.

12.2 Displacement Current

Ampère's law is

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

Taking the divergence of this we have

$$\nabla \cdot (\nabla \times \boldsymbol{B}) = \mu_0 \nabla \cdot \boldsymbol{J} .$$

The left hand side is identically zero by a vector calculus identity. For the right hand side we apply the continuity equation and get

$$0 = -\mu_0 \frac{\partial \rho}{\partial t}.$$

This is obviously wrong for a non-static charge distribution, $\rho(\boldsymbol{r}, t)$, which we know can exist and would expect ρ not to be constant with time. Using the continuity equation we can known

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

and so

$$\nabla \cdot (\nabla \times \boldsymbol{B}) = \mu_0 (\nabla \cdot \boldsymbol{J} + \partial_t \rho)$$

From Maxwell's first law we know that

$$\rho = \varepsilon_0 \nabla \cdot \boldsymbol{E}$$

so

$$\nabla \cdot (\nabla \times \boldsymbol{B}) = \mu_0 (\nabla \cdot \boldsymbol{J} + \varepsilon_0 \partial_t \nabla \cdot \boldsymbol{E}) = \mu_0 \nabla \cdot (\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E})$$

Since $\nabla \cdot (\nabla \times \boldsymbol{K}) = 0$ for all vector fields, $\boldsymbol{K}$, we must have

$$\nabla \times \boldsymbol{B} = \mu_0 (\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E} + \nabla \times \boldsymbol{K}).$$

It turns out that $\boldsymbol{K} = \boldsymbol{0}$, this is required so that in the static case this equation reduces to $\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J}$. Therefore

$$\nabla \times \boldsymbol{B} = \mu_0 (\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E}).$$

This is the **Ampère–Maxwell law** and is the full version of Maxwell's fourth equation with time dependent fields. It also has an integral form, using Stoke's theorem we have

$$\int_S (\nabla \times \boldsymbol{B}) \cdot \mathrm{d}\boldsymbol{S} = \oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l}$$

also

$$\mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} + \varepsilon_0 \int_S \frac{\partial}{\partial t} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} + \mu_0 \varepsilon_0 \frac{\mathrm{d}}{\mathrm{d}t} \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}$$

so

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} + \mu_0 \varepsilon_0 \frac{\mathrm{d}}{\mathrm{d}t} \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S}$$

What this law tells us is that magnetic fields can originate from two sources:

- Steady currents, $\boldsymbol{J}$
- Time dependent electric fields, $\boldsymbol{E}$ (which themselves are created by time dependent charge densities or time dependent magnetic fields that change at a non-constant rate).

We call $\varepsilon_0 \partial_t \boldsymbol{E}$ the **displacement current** because it has dimensions of current density and creates a magnetic field. It is not a real current however.

12.2.1 Capacitor Paradox and Resolution

The capacitor paradox is another example of why Ampère's law alone is not quite correct. The capacitor paradox is as follows. Consider a capacitor charging/discharging with a time dependent current, $I(t)$, What is the magnetic field? Since the current is time dependent then so is the field. By Ampère's law we have

$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S}\,.$$

The paradox arises in the right hand side of this equation. If we take the surface S to cut through the wire then the right hand side is equal to $\mu_0 I(t)$. It is also possible to construct a surface with the same boundary in a way such that the surface goes in between the plates of the capacitor without ever intersecting any of the circuit. Then the right hand side of this equation is zero.

We fix the supposed paradox by using the full Ampère–Maxwell law. We known that for an ideal parallel plate capacitor the electric field is entirely constrained to be between the plates and in this volume it is normal to the plates and the field strength is

$$E(t) = \frac{Q(t)}{\varepsilon_0 A}$$

where A is the area of the plates. If we define the direction of $\boldsymbol{E}$ to be the z direction then

$$\varepsilon_0 \partial_t \boldsymbol{E} = \frac{1}{A}\frac{\partial Q}{\partial t}\boldsymbol{e_z} = \frac{I(t)}{A}\boldsymbol{e_z}.$$

If we also choose to construct the surface so that it is parallel to the plates of the capacitor when it is between them then

$$\mu_0 \int_{S_2} (\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E}) \cdot \mathrm{d}\boldsymbol{S} = \frac{I(t)}{A}\int_{S_2} \mathrm{d}S = \mu_0 I(t)$$

where we have used the fact that the field is zero outside of the capacitors so the surface integral is equivalent to an integral over the area of the capacitor. This is exactly the same result that we get using the other surface so we have fixed the paradox.

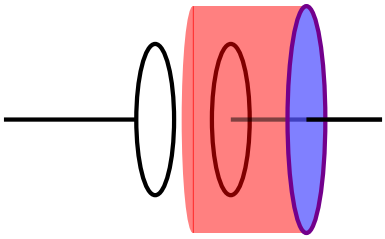


Figure 12.1: The capacitor paradox. The two surfaces used in the capacitor paradox. S_1 in blue goes through the wire and S_2 in red goes through the gap in the capacitor. They share the boundary, C, shown in purple.

12.3 Maxwell's Equations

Maxwell's equations in differential form are

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0}, \tag{MI}$$
$$\nabla \cdot \boldsymbol{B} = 0, \tag{MII}$$
$$\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}, \tag{MIII}$$
$$\nabla \times \boldsymbol{B} = \mu_0(\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E}). \tag{MIV}$$

In integral form these are

$$\oint_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \frac{Q_{\text{enc}}}{\varepsilon_0}, \tag{MI}$$
$$\oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0, \tag{MII}$$
$$\oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = -\frac{\mathrm{d}\Phi_B}{\mathrm{d}t} = \mathcal{E}, \tag{MIII}$$
$$\oint_C \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} + \mu_0\varepsilon_0 \frac{\mathrm{d}}{\mathrm{d}t}\int_S \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{S} = \mu_0 I_{\text{enc}} + \mu_0\varepsilon_0 \frac{\mathrm{d}\Phi_E}{\mathrm{d}t}. \tag{MIV}$$

The first two equations are Gauss' laws for electric and magnetic fields, the third is Faraday's law of induction and the fourth is the Ampère–Maxwell law. Combining Maxwell's first and fourth laws as well as the fact that the divergence of a curl is zero we get

$$0 = \nabla \cdot (\nabla \times \boldsymbol{B}) = \mu_0 \left[\nabla \cdot \boldsymbol{J} + \varepsilon_0 \partial_t \nabla \cdot \boldsymbol{E}\right] = \mu_0 \left[\nabla \cdot \boldsymbol{J} + \partial_t \rho\right]$$

which is the continuity equation again.

12.3.1 Solutions to Maxwell's Equation

We will look for non-trivial solutions to Maxwell's equations in free space, that is with $\rho = 0$ and $\boldsymbol{J} = \boldsymbol{0}$. In this case Maxwell's equations become

$$\begin{aligned} \nabla \cdot \boldsymbol{E} &= 0 && \text{(MI in free space)} \\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(MII)} \\ \nabla \times \boldsymbol{E} &= -\partial_t \boldsymbol{B} && \text{(MIII)} \\ \nabla \times \boldsymbol{B} &= \mu_0 \varepsilon_0 \partial_t \boldsymbol{E} && \text{(MIV in free space)} \end{aligned}$$

Taking the curl of the third equation we have

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\nabla \times (\partial_t \boldsymbol{B}) = -\partial_t \nabla \times \boldsymbol{B} = -\mu_0 \varepsilon_0 \partial_t^2 \boldsymbol{E}.$$

Similarly taking the curl of the fourth equation we have

$$\nabla \times (\nabla \times \boldsymbol{B}) = \mu_0 \varepsilon_0 \nabla \times (\partial_t \boldsymbol{E}) = \mu_0 \varepsilon_0 \partial_t \nabla \times \boldsymbol{E} = -\mu_0 \varepsilon_0 \partial_t^2 \boldsymbol{B}.$$

Or we can use the vector identity

$$\nabla \times (\nabla \times \boldsymbol{K}) = \nabla(\nabla \cdot \boldsymbol{K}) - \nabla^2 \boldsymbol{K}$$

and we get

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

and

$$\nabla \times (\nabla \times \boldsymbol{B}) = \nabla(\nabla \cdot \boldsymbol{B}) - \nabla^2 \boldsymbol{B} = -\nabla^2 \boldsymbol{E}.$$

Thus

$$\partial_t^2 \boldsymbol{E} = \frac{1}{\mu_0 \varepsilon_0} \nabla^2 \boldsymbol{E} \qquad \text{and} \qquad \partial_t^2 \boldsymbol{B} = \frac{1}{\mu_0 \varepsilon_0} \nabla^2 \boldsymbol{B}.$$

Thus we have turned four coupled first order partial differential equations (PDEs) into two seemingly uncoupled second order PDEs. They are only seemingly uncoupled as Maxwell's laws always apply. We can identify these equations as wave equations, which have the general form

$$\partial_t^2 \boldsymbol{u} = v^2 \nabla^2 \boldsymbol{u}$$

where v is the speed of waves in the vector field $\boldsymbol{u}$. We see that for the waves in the electric and magnetic fields

$$v = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = c$$

and indeed these equations relate to electromagnetic waves.

13 Electromagnetic Waves

13.1 The Wave Equation in One Dimension

The wave equation for a scalar field, u, with a wave propagating at speed c, is

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}.$$

This is a second order PDE and typically we have initial conditions of $u(x, 0)$ and $\dot{u}(x, 0)$. Any twice differentiable function, u, of the form

$$u(x, t) = f(kx - \omega t),$$

or equivalently

$$u(x, t) = g(x - ct),$$

satisfies this equation and represents a wave moving in the x direction at speed

$$c = \frac{\omega}{k},$$

with frequency ω and wave number k. For example,

$$u(x,t) = u_0 \exp\left[-\frac{(x-ct)^2}{2\sigma^2}\right]$$

describes a Gaussian wave packet moving in the x direction at speed c. The most important function of this form is

$$u(x,t) = A\exp[i(kx-\omega t)] = A\cos(kx-\omega t) + iA\sin(kx-\omega t).$$

Here A is the amplitude of the waves (it is possible that this is complex). We can see that this is made of two sinusoidal waves which are independent of each other. This wave, u, is monochromatic because, although it is a superposition of two waves, both have the same frequency, ω. Sine and cosine form a basis for a Fourier expansion of some arbitrary solution, $u(x,t)$. By this we mean that given a solution, u, to the wave equation it is always possible to write it as a superposition of sinusoids. It is often convenient to work with complex expressions and then take the real part when we need to get a physical solution. For example,

$$\begin{aligned} u(x,t) &= \mathrm{Re}[A\exp(i(kx-\omega t))] \\ &= \mathrm{Re}[A\cos(kx-\omega t) + iA\sin(kx-\omega t)] \\ &= \mathrm{Re}[A\cos(kx-\omega t)] + \mathrm{Im}[A\sin(kx-\omega t)]. \end{aligned}$$

Here we have used that if $z = a+bi$ for $a, b \in \mathbb{R}$ then

$$\mathrm{Re}[iz] = \mathrm{Re}[ia] - \mathrm{Re}[b] = -b.$$

To solve the wave equation we start by writing the initial conditions as inverse Fourier transforms:

$$u(x,0) = \int_{-\infty}^{\infty} \tilde{g}(k)e^{ikx}\,\mathrm{d}k$$

and

$$\dot{u}(x,0) = \int_{-\infty}^{\infty} \tilde{h}(k)e^{ikx}\,\mathrm{d}k$$

where $g(x) = u(x,0)$ and $h(x) = \dot{u}(x,0)$ and $\tilde{g}$ and $\tilde{h}$ are the Fourier transforms of g and h respectively. We then solve the wave equation for the initial conditions

$$f(x,0) = \tilde{g}(k)e^{ikx}, \qquad \text{and} \qquad \dot{f}(x,0) = \tilde{h}(k)e^{ikx}.$$

We obtain the monochromatic solution

$$f(kx-\omega t) = \left[\tilde{g}(k) + \frac{i}{\omega}\tilde{h}(k)\right]\exp[i(kx-\omega t)].$$

The full solution, $u = u(x,t)$, is then a superposition of all frequencies/wave numbers which we can write as an integral over one of these:

$$u(x,t) = \int_{-\infty}^{\infty} f(kx-\omega t)\,\mathrm{d}k = \int_{-\infty}^{\infty}\left[\tilde{g}(k) + \frac{i}{\omega}\tilde{h}(k)\right]\exp[i(kx-\omega t)]\,\mathrm{d}k\,.$$

13.2 The Wave Equation in Three Dimensions

In three dimensions the wave equation for a scalar field, u, with a wave propagating at speed c, is

$$\nabla^2 u = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}.$$

Again any twice differentiable function of the form

$$u(\boldsymbol{r},t) = f(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)$$

is a solution. Here $\boldsymbol{k}$ is the wave vector which points in the direction of propagation and is such that $c = \omega/k$ still. A similar form that is still a solution is

$$u(\boldsymbol{r},t) = g(\hat{\boldsymbol{k}}\cdot\boldsymbol{r} - ct).$$

We will show here that a plane wave,

$$u(\boldsymbol{r},t) = A\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)],$$

is a solution.

$$\begin{aligned}\nabla u(\boldsymbol{r},t) &= \nabla A\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)]\\ &= \boldsymbol{e_j}\partial_j A\exp[i(k_j x_j-\omega t)]\\ &= \boldsymbol{e_j}k_j kA\exp[i(k_j x_j-\omega t)]\\ &= i\boldsymbol{k}A\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)]\\ &= i\boldsymbol{k}u(\boldsymbol{r},t).\end{aligned}$$

We see here the reason that this is called a plane wave. The gradient gives the direction of fastest increase and in this case ∇u is parallel to $\boldsymbol{k}$. This means that in a plane normal to $\boldsymbol{k}$, that is for the plane

$$\boldsymbol{k}\cdot\boldsymbol{r} = \omega t + a,$$

for some $a\in\mathbb{R}$, $u(\boldsymbol{r},t)$ has the same value everywhere in the plane. Continuing on with showing that this is indeed a wave:

$$\begin{aligned}\nabla^2 u(\boldsymbol{r},t) &= \nabla\cdot\nabla u(\boldsymbol{r},t)\\ &= \nabla\cdot[i\boldsymbol{k}u(\boldsymbol{r},t)]\\ &= i\boldsymbol{k}\cdot\nabla u(\boldsymbol{r},t)\\ &= i\boldsymbol{k}\cdot i\boldsymbol{k}u(\boldsymbol{r},t)\\ &= -k^2u(\boldsymbol{r},t).\end{aligned}$$

Also

$$\begin{aligned}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}u(\boldsymbol{r},t) &= \frac{1}{c^2}\frac{\partial^2}{\partial t^2}A\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)]\\ &= -\frac{1}{c^2}\omega^2 A\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)]\\ &= -\frac{\omega^2}{c^2}u(\boldsymbol{r},t)\end{aligned}$$

so we see that u is a solution as long as $k=\omega/c$.

13.3 The Wave Equation for a Vector Field

The wave equation for a vector field, $\boldsymbol{F}$, is

$$\nabla^2\boldsymbol{F}(\boldsymbol{r},t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\boldsymbol{F}(\boldsymbol{r},t).$$

This has the expected plane wave solution:

$$\boldsymbol{F}(\boldsymbol{r},t) = \boldsymbol{F_0}\exp[i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)].$$

Here $\boldsymbol{F_0}$ is a constant (possibly complex) vector. In a similar way to the previous section we can show that for a plane wave $\boldsymbol{F}(\boldsymbol{r},t)$

$$\nabla\cdot\boldsymbol{F} = i\boldsymbol{k}\cdot\boldsymbol{F}, \qquad \text{and} \qquad \nabla^2\boldsymbol{F} = -k^2\boldsymbol{F}.$$

We also have that

$$\nabla\times\boldsymbol{F} = i\boldsymbol{k}\times\boldsymbol{F}.$$

13.4 Electromagnetic Plane Waves

The wave equations for the electric and magnetic fields in free space are

$$\nabla^2 \boldsymbol{E} = \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \boldsymbol{E}, \qquad \text{and} \qquad \nabla^2 \boldsymbol{B} = \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \boldsymbol{B}.$$

The plane wave solutions to these are

$$\boldsymbol{E} = \boldsymbol{E_0} \exp[i(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)], \qquad \text{and} \qquad \boldsymbol{B} = \boldsymbol{B_0} \exp[i(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)].$$

In free space we know that $\nabla \cdot \boldsymbol{E} = \nabla \cdot \boldsymbol{B} = 0$. We also know the divergence of a plane wave, $\boldsymbol{F}$, is $\nabla \cdot \boldsymbol{F} = i\boldsymbol{k} \cdot \boldsymbol{F}$, so we must have

$$0 = i\boldsymbol{k} \cdot \boldsymbol{E}.$$

This implies that $\boldsymbol{k} \cdot \boldsymbol{E} = 0$ which means that $\boldsymbol{k}$ and $\boldsymbol{E}$ are perpendicular. The exact same logic shows us that $\boldsymbol{k}$ and $\boldsymbol{B}$ must be perpendicular. For this reason we call electromagnetic waves transverse because their amplitude is in a plane perpendicular to the direction of propagation, which is along $\boldsymbol{k}$.

Maxwell's third law in free space gives us $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$. We also know that the curl of a plane wave, $\boldsymbol{F}$, is $\nabla \times \boldsymbol{F} = i\boldsymbol{k} \times \boldsymbol{F}$, and the time derivative of a plane wave, $\boldsymbol{F}$, is $\partial_t \boldsymbol{F} = -i\omega \boldsymbol{F}$. Thus

$$-i\omega \boldsymbol{B} = i\boldsymbol{k} \times \boldsymbol{E}.$$

This means that $\boldsymbol{B}$ and $\boldsymbol{E}$ are perpendicular, further

$$E = \frac{\omega}{k} B = cB.$$

So not only are $\boldsymbol{E}$ and $\boldsymbol{B}$ perpendicular to the direction of propagation, $\boldsymbol{k}$, they are also perpendicular to each other with proportional field strengths with c as the constant of proportionality.

13.5 Polarisation

Let $\boldsymbol{k} = k\boldsymbol{e_z}$. Then the amplitudes of electromagnetic plane waves, $\boldsymbol{E_0}$ and $\boldsymbol{B_0}$, are in the (x, y)-plane. The most general electromagnetic plane wave has

$$\boldsymbol{E_0} = E_0 e^{i\varphi}(\alpha \boldsymbol{e_x} + i\beta \boldsymbol{e_y})$$

where φ is some phase and $\alpha^2 + \beta^2 = 1$. We can parametrise this as $\alpha = \cos\zeta$ and $\beta = \sin\zeta$ for some angle ζ. We say that this wave is elliptically polarised as plotting this of $\zeta \in [0, 2\pi]$ gives an ellipse.

13.5.1 Linear Polarisation

If $|\alpha| = 1$ and $\beta = 0$ (or $\alpha = 0$ and $|\beta| = 1$) then the plane wave solution reduces to

$$\boldsymbol{E_0} = E_0 e^{i\varphi} \boldsymbol{e_x}.$$

To get the wave we take the real part of $\boldsymbol{E}$:

$$\mathrm{Re}[\boldsymbol{E}] = \mathrm{Re}[\boldsymbol{E_0} e^{i\varphi} e^{i(kz - \omega t)} \boldsymbol{e_x}] = E_0 \cos(kz - \omega t + \varphi) \boldsymbol{e_x}.$$

Notice that $\boldsymbol{k} \cdot \boldsymbol{r}$ reduces to kz as $k_x = k_y = 0$. We see that $\boldsymbol{E}$ is polarised along the $\boldsymbol{e_x}$ direction. This means that $\boldsymbol{B}$ must be polarised along the $\boldsymbol{e_y}$ direction.

13.5.2 Circular Polarisation

If $|\alpha| = |\beta| = \sqrt{2}/2$ then

$$\boldsymbol{E_0} = E_0 \frac{\sqrt{2}}{2} e^{i\varphi}(\boldsymbol{e_x} \pm i\boldsymbol{e_y}).$$

Again we take the real part to get the actual wave:

$$\mathrm{Re}[\boldsymbol{E}] = \mathrm{Re}\left[\boldsymbol{E_0} \frac{\sqrt{2}}{2} e^{i\varphi}(\boldsymbol{e_x} \pm i\boldsymbol{e_y})\right] = E_0 \frac{\sqrt{2}}{2} [\cos(kz - \omega t + \varphi) \mp \sin(kz - \omega t - \varphi) \boldsymbol{e_y}].$$

This corresponds to circular polarisation. The solution with the minus sign gives anticlockwise polarisation, also known as positive helicity or left circular polarisation. The solution with the plus sign gives clockwise polarisation, also known as negative helicity or right circular polarisation.

Both linearly and circularly polarised waves provide a basis from which any general (elliptical) polarisation can be described as a superposition of linearly/circularly polarised waves.

14 Energy and the Poynting Vector

Recall that the total energy stored in a static electromagnetic field is

$$U_{EM} = U_E + U_B = \frac{1}{2}\int \varepsilon_0 |\boldsymbol{E}(\boldsymbol{r})|^2 \mathrm{d}^3 r + \frac{1}{2}\int \frac{1}{\mu_0}|\boldsymbol{B}(\boldsymbol{r})|^2 \,\mathrm{d}^3 r\,.$$

This is simply the sum of the energy stored in the electric field and the magnetic field which were derived in sections 6.1 and 11.5 respectively. We will find a general (non-static) expression for U_{EM} in this section using the full version of Maxwell's equations.

Suppose we have a general charge density, $\rho(\boldsymbol{r}, t)$, and current density, $\boldsymbol{J}(\boldsymbol{r}, t)$. What is the work done on a moving charge, q? If the charge has velocity $\boldsymbol{v}$ then in time $\mathrm{d}t$ the charge is displaced by $\mathrm{d}\boldsymbol{\ell} = \boldsymbol{v}\,\mathrm{d}t$. The Lorentz force law still holds for non-static fields as it is how we define the fields. Thus the work done on the charge is

$$\mathrm{d}U = \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{\ell} = q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B})\cdot\boldsymbol{v}\,\mathrm{d}t = q\boldsymbol{E}\boldsymbol{v}\,\mathrm{d}t\,.$$

Where we have used that $\boldsymbol{v}\times\boldsymbol{B}$ is orthogonal to $\boldsymbol{v}$ so the dot product gives zero and no work is done by the magnetic field. Now let $q = \rho\,\mathrm{d}V$ and $\boldsymbol{J} = \rho\boldsymbol{v}$. Then dividing by $\mathrm{d}t$ and integrating in a limiting process we have

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \int_V \boldsymbol{E}\cdot\boldsymbol{J}\,\mathrm{d}V\,.$$

This is the power delivered to the volume, V, so $\boldsymbol{E}\cdot\boldsymbol{J}$ is the power delivered per unit volume. From Maxwell's fourth law we have

$$\nabla\times\boldsymbol{B} = \mu_0(\boldsymbol{J} + \varepsilon_0\partial_t\boldsymbol{E}) \implies \boldsymbol{J} = \frac{1}{\mu_0}\nabla\times\boldsymbol{B} - \varepsilon_0\partial_t\boldsymbol{E}$$

so

$$\boldsymbol{E}\cdot\boldsymbol{J} = \frac{1}{\mu_0}\boldsymbol{E}\cdot(\nabla\times\boldsymbol{B}) - \varepsilon_0\boldsymbol{E}\cdot\partial_t\boldsymbol{E}.$$

Next we use the product rule,

$$\nabla\cdot(\boldsymbol{K}\times\boldsymbol{V}) = \boldsymbol{V}\cdot(\nabla\times\boldsymbol{K}) - \boldsymbol{K}\cdot(\nabla\times\boldsymbol{V}),$$

to write

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

Using Maxwell's third law this gives

$$\boldsymbol{E}\cdot(\nabla\times\boldsymbol{B}) = -\boldsymbol{B}\cdot\partial_t\boldsymbol{E} - \nabla\cdot(\boldsymbol{E}\times\boldsymbol{B}).$$

Next we use the normal product rule:

$$\frac{\partial}{\partial t}(\boldsymbol{V}\cdot\boldsymbol{V}) = \frac{\partial\boldsymbol{V}}{\partial t}\cdot\boldsymbol{V} + \boldsymbol{V}\cdot\frac{\partial\boldsymbol{V}}{\partial t} = 2\boldsymbol{V}\cdot\frac{\partial\boldsymbol{V}}{\partial t} \implies \boldsymbol{V}\cdot\frac{\partial\boldsymbol{V}}{\partial t} = \frac{1}{2}\left(\frac{\partial V}{\partial t}\right)^2.$$

Combining these we have

$$\boldsymbol{E}\cdot\boldsymbol{J} = -\frac{1}{2}\frac{\partial}{\partial t}\left[\varepsilon_0 E^2 + \frac{1}{\mu_0}B^2\right] - \frac{1}{\mu_0}\nabla\cdot(\boldsymbol{E}\times\boldsymbol{B}).$$

Thus

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \int_V \boldsymbol{E}\cdot\boldsymbol{J}\,\mathrm{d}V = -\frac{1}{2}\frac{\partial}{\partial t}\int_V\left(\varepsilon_0 E^2 + \frac{1}{\mu_0}B^2\right)\mathrm{d}V - \frac{1}{\mu_0}\oint_A(\boldsymbol{E}\times\boldsymbol{B})\cdot\mathrm{d}\boldsymbol{A}\,.$$

This is known as **Poynting's theorem**. The left hand side is the power delivered to charge carriers in V, which is the rate of energy gain of these charges. The first term on the right hand side is the loss rate of electromagnetic energy stored in the electric and magnetic fields in V. The second term on the right hand side is the flux rate of energy out of the volume. From this we see that the energy lost by the fields is equal to the energy gained by the charges plus the the energy that leaves V. We introduce the energy flux density,

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

called the **Poynting vector** and we have

$$\frac{\mathrm{d}U}{\mathrm{d}t} = \frac{\mathrm{d}U_{EM}}{\mathrm{d}t} - \oint_A \boldsymbol{S}\cdot\mathrm{d}\boldsymbol{A} \implies \frac{\mathrm{d}}{\mathrm{d}t}(U+U_{EM}) = -\oint_A \boldsymbol{S}\cdot\mathrm{d}\boldsymbol{A}\,.$$

Defining energy density of the charges, which is the mechanical energy density, u_{mech}, and energy density of the fields, u_{EM}, applying the divergence theorem, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(U+U_{EM}) = \frac{\mathrm{d}}{\mathrm{d}t}\int_V (u_{\text{mech}}+u_{EM})\,\mathrm{d}V = -\int_V \nabla\cdot\boldsymbol{S}\,\mathrm{d}V\,.$$

Thus

$$\frac{\partial}{\partial t}(u_{\text{mech}}+u_{EM}) = -\nabla\cdot\boldsymbol{S}.$$

This is an energy continuity equation (that is it implies conservation of energy):

$$\frac{\partial u}{\partial t} + \nabla\cdot\boldsymbol{S} = 0$$

where $u = u_{\text{mech}} + u_{EM}$.

14.1 Energy of Electromagnetic Waves

Choosing axis such that $\varphi = n\pi$ and $\boldsymbol{k} = k\boldsymbol{e_z}$ for linearly polarised electromagnetic waves we have

$$\boldsymbol{E} = E_0\boldsymbol{e_x}\cos(kz-\omega t)$$

and

$$\boldsymbol{B} = B_0\boldsymbol{e_y}\sin(kz-\omega t).$$

Recall also that $B_0 = E_0/c$. The electromagnetic energy density stored is thus

$$\begin{aligned}
u_EM &= \frac{1}{2}\left(\frac{1}{\mu_0}B^2 + \varepsilon_0E^2\right)\\
&= \frac{1}{2}\left(\frac{1}{\mu_0}\frac{1}{c^2}E^2 + \varepsilon_0E^2\right)\\
&= \frac{1}{2}\left(\frac{1}{\mu_0}\mu_0\varepsilon_0E^2 + \varepsilon_0E^2\right)\\
&= \frac{1}{2}\left(\varepsilon_0E^2 + \varepsilon_0E^2\right)\\
&= \varepsilon_0E^2\\
&= \varepsilon_0E_0\cos^2(kz-\omega t).
\end{aligned}$$

We see that the energy is split evenly between $\boldsymbol{E}$ and $\boldsymbol{B}$ for an electromagnetic wave. The Poynting vector is then

$$\boldsymbol{S} = \frac{1}{\mu_0}(\boldsymbol{E}\times\boldsymbol{B}) = \varepsilon_0cE_0^2\cos^2(kz-\omega t)\boldsymbol{e_z} = u_{EM}c\boldsymbol{e_z}.$$

This makes sense as we can think of the product on the right as the amount of energy that can move through a certain area per unit time, consider a pipe of fluid of mass density ρ flowing at velocity v, in time t a mass of ρvt would pass through a cross section of the pipe. This generalises to a wave with general $\boldsymbol{k}$, we then have

$$\boldsymbol{S} = u_{EM}c\hat{\boldsymbol{k}}.$$

The time average of the energy density is defined as the average over one period, T, it is given by

$$\begin{aligned}
\langle u_{EM}\rangle &= \frac{\varepsilon_0E_0^2}{T}\int_0^T \cos^2(kz-\omega t)\,\mathrm{d}t\\
&= \frac{\varepsilon_0E_0^2}{T}\frac{T}{2}\\
&= \frac{1}{2}\varepsilon_0E_0^2\\
&= \frac{1}{2}\frac{B_0^2}{\mu_0}.
\end{aligned}$$

So we see that the energy density of an electromagnetic wave is proportional to the square of the electric or magnetic field.

14.2 Energy of Discharging Capacitor

Consider a circular parallel plate capacitor, C, with plate area A, being discharged through a resistor, R. From Ohm's law we know that

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

using the fact that

$$I = -\frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{Q}{RC}$$

we have

$$Q = Q_0 e^{-t/RC} = Q_0 e^{-t/\tau}$$

and

$$I = I_0 e^{-t/RC} = \frac{Q_0}{RC} e^{-t/\tau}.$$

We assume a quasistatic approximation where we treat the fields as static at any one instant. Thus

$$\boldsymbol{E} = \frac{Q}{A\varepsilon_0}\hat{\boldsymbol{n}} = \frac{Q_0}{A\varepsilon_0} e^{-t/\tau}\hat{\boldsymbol{n}},$$

where $\hat{\boldsymbol{n}}$ is normal to the plates. We can compute $\boldsymbol{B}$ from the Ampère–Maxwell law. The cylindrical symmetry means that the magnetic field must be circumferential. The Ampérian loop that we choose is a circle of radius r between the plates where $\boldsymbol{J} = \boldsymbol{0}$. Thus

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{l} = \mu_0 \int_S (\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E}) \cdot \mathrm{d}\boldsymbol{S} = \mu_0 \pi r^2 \varepsilon_0 \partial_t \left(\frac{Q_0}{A\varepsilon_0} e^{-t/\tau} \right).$$

The left hand side of this is $2\pi r B_\varphi$ so

$$\boldsymbol{B} = -\frac{\mu_0 I(t) r}{2A} \boldsymbol{e}_{\boldsymbol{\varphi}}.$$

Hence

$$\boldsymbol{S} = \frac{1}{\mu_0}\boldsymbol{E} \times \boldsymbol{B} = -\frac{Q_0}{A\varepsilon_0} e^{-t/\tau} I_0 \frac{r}{2A} e^{-t/\tau} \boldsymbol{e}_{\boldsymbol{z}} \times \boldsymbol{e}_{\boldsymbol{\varphi}} = \frac{I_0^2 CR}{2A^2 \varepsilon_0} r e^{-2t/\tau} \boldsymbol{e}_{\boldsymbol{r}}.$$

So the Poynting vector, and hence energy flow, points radially out of the capacitor.

14.3 Momentum of Electromagnetic Radiation

This section is non-examinable.

We can interpret the Poynting vector from a quantum mechanical perspective. Electromagnetic radiation can be viewed as photons travelling with speed c, energy

$$\varepsilon = \hbar\omega = h\nu,$$

and momentum

$$\boldsymbol{p} = \hbar \boldsymbol{k} = \frac{\varepsilon}{c}\hat{\boldsymbol{k}}.$$

For n photons per unit volume travelling at speed c we can interpret the average Poynting vector as the average energy density, $n\varepsilon$, multiplied by the velocity vector, $c\hat{\boldsymbol{k}}$:

$$\langle \boldsymbol{S} \rangle = n\varepsilon c \hat{\boldsymbol{k}} = \langle u_{EM} \rangle c \hat{\boldsymbol{k}}.$$

Thinking of energy transport by photons we have an accompanying momentum flux, $\tilde{\boldsymbol{P}}$. This is defined as the momentum carried across a plane normal to the propagation per unit area and per unit time. For each photon $p = \varepsilon/c$ along the $\hat{\boldsymbol{k}}$ direction so

$$\tilde{\boldsymbol{P}} = \frac{1}{c}\boldsymbol{S}.$$

If light is absorbed on a surface normal to its propagation then this momentum is transferred to the surface which creates a force per unit area equal to the incoming momentum flux. This is called the radiation pressure:

$$p_{\mathrm{rad}} = \tilde{\boldsymbol{P}} \cdot \hat{\boldsymbol{n}} = \frac{S}{c} \implies p_{\mathrm{rad}} = \langle u_{EM} \rangle.$$

If instead the light is reflected then twice the momentum is transferred to the surface to conserve total momentum. However $\langle u_{EM} \rangle$ also doubles so the result holds.

For a classical understanding of radiation pressure we consider a linear polarised wave with $\boldsymbol{k} = k\boldsymbol{e_z}$. The electric field moves charges on the surface where the radiation is absorbed. For a particular charge, q, that ends up moving at velocity $\boldsymbol{v}$ in the $\boldsymbol{E}$ direction the force due to the magnetic field is then $q\boldsymbol{v} \times \boldsymbol{B}$. Since $\boldsymbol{E}$ is perpendicular to $\boldsymbol{e_z}$ the force is parallel to $\boldsymbol{e_z}$ and this is what we can view as creating the radiation pressure.

15 The Electric Field in Media

15.1 Motivation

The electric field in a vacuum is described by Maxwell's equations, in particular

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\varepsilon_0}, \qquad \text{and} \qquad \nabla \times \boldsymbol{B} = \mu_0 \left(\boldsymbol{J} + \varepsilon_0 \partial_t \boldsymbol{E} \right).$$

To use these equations we need to know ρ and $\boldsymbol{J}$ with atomic precision. This is not possible in reality, even before we consider quantum effects important on this scale it is simply impractical to make these measurements. In real materials we have approximately an Avagadro's number of particles which may all have charge and be moving creating currents. It would be better to have a macroscopic effective theory, meaning that it can be used to make predictions without necessarily knowing the underlying mechanism, for electromagnetism (EM) fields interacting with matter.

We classify most materials as one of two types, conductors and insulators. Up until now we have considered conductors to have an unlimited supply of free charge carriers which can move unimpeded. One of the most important consequences of this assumption is that an electric field will create a surface charge density and there will be no field inside the material. On the other hand insulators have been assumed to have no charge carriers. In the context of EM in media we call insulators dielectrics. The assumptions made about conductors and insulators are only an approximation and we will now give a more full treatment of EM in media.

15.2 Dielectric Materials

If we have a dielectric and we apply an electric field then an electric dipole will be created in response. The mechanism and hence exact details of the dipole depend on the nature of the dielectric. An ionic solid can be viewed as a lattice of positive and negative charges on the grid points. The electric field will cause the positive charges to move slightly in the direction of the field and the negative charges in the opposite direction. This causes a charge imbalance which results in a dipole aligned with the electric field.

A single atom can be polarised as an electric field will cause the positive nucleus to move in the direction of the field and the negative electrons to move in the opposite direction. This again causes a charge imbalance which causes a dipole parallel to the electric field.

A polar molecule, such as water, H_2O, has a natural dipole anyway. Applying an external electric field will cause these dipoles to align causing a net dipole parallel to the electric field.

In all three cases detailed above an external electric field causes a dipole aligned with the external field. While the exact details on the atomic scale are, at least in practice, unknowable, we can work with averages and other macroscopic quantities. The average polarisation of a single atom or molecule in an external electric field, $\boldsymbol{E}$, is

$$\langle \boldsymbol{p_{at}} \rangle = \alpha \boldsymbol{E},$$

where $\langle \cdot \rangle$ denotes a time average which accounts for thermal fluctuations. α is the atomic/molecular polarisability. In general α is a tensor and therefore $\langle \boldsymbol{p_{at}} \rangle$ is not necessarily parallel to $\boldsymbol{E}$. In practice for small E we can usually approximate α as a scalar. The effect of each individual atom/molecule being polarised is a net dipole moment per unit volume, or **polarisation**, $\boldsymbol{P} = n\langle \boldsymbol{p_a t} \rangle$ where n is the number density of the atom/molecule. More generally we can define the polarisation field, $\boldsymbol{P}$, through the net dipole moment, $\mathrm{d}\boldsymbol{p}$, in a small volume, $\mathrm{d}V$, by

$$\mathrm{d}\boldsymbol{p} = \boldsymbol{P}\,\mathrm{d}V\,.$$

Suppose now that we have a polarised material. What is the field due to this polarisation. Recall that for a dipole at $\boldsymbol{r'}$ the potential at $\boldsymbol{r}$ is

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\frac{(\boldsymbol{r}-\boldsymbol{r'})\cdot\boldsymbol{p}}{|\boldsymbol{r}-\boldsymbol{r'}|^3}.$$

This generalises by superposition to the potential due to the polarisation field, $\boldsymbol{P}$, in volume V:

$$V(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_0}\int_V \frac{(\boldsymbol{r}-\boldsymbol{r'})\cdot\boldsymbol{P}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\,\mathrm{d}^3r\,.$$

If ∇' is the normal gradient operator that acts only on $\boldsymbol{r'}$ then

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

Thus

$$\begin{aligned}
V(\boldsymbol{r}) &= \frac{1}{4\pi\varepsilon_0}\int_V \boldsymbol{P}(\boldsymbol{r'})\cdot\nabla'\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\mathrm{d}^3r' \\
&= \frac{1}{4\pi\varepsilon_0}\int_V \nabla'\cdot\left(\frac{\boldsymbol{P}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\right)\mathrm{d}^3r' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\nabla'\cdot\boldsymbol{P}(\boldsymbol{r'})\,\mathrm{d}^3r' \\
&= \frac{1}{4\pi\varepsilon_0}\oint_S \frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\boldsymbol{P}(\boldsymbol{r'})\cdot\mathrm{d}\boldsymbol{S'} - \frac{1}{4\pi\varepsilon_0}\int_V \frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\nabla'\cdot\boldsymbol{P}(\boldsymbol{r'})\,\mathrm{d}^3r'\,.
\end{aligned}$$

The first term can be viewed as the potential due to a surface charge distribution, σ_b, defined by

$$\sigma_b = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$$

where $\hat{\boldsymbol{n}}$ is the surface normal. The second term can be viewed as the potential due to a volume charge density, ρ_b, given by

$$\rho_b = -\nabla\cdot\boldsymbol{P}.$$

The subscript bs in these terms refers to the fact that these charge distributions are 'bound' to the atoms.

From this analysis we see that $\boldsymbol{P}$ is both a result of an electric field and a source of an electric field. This causes a recursive problem as to find $\boldsymbol{P}$ we need $\boldsymbol{E}$ and to find $\boldsymbol{E}$ we need to know $\boldsymbol{P}$. We fix this problem by defining a new macroscopic field.

15.3 Electric Displacement Field

We can divide a charge distribution, ρ, into two parts, $\rho = \rho_f + \rho_b$. Here ρ_f is the free charge with which we have dealt in the first part of this course, and ρ_b is the bound charge as defined in the previous section. Maxwell's first equation then becomes

$$\nabla\cdot\boldsymbol{E} - \frac{\rho}{\varepsilon_0} = \frac{\rho_f}{\varepsilon_0} + \frac{\rho_b}{\varepsilon_0} = \frac{\rho_f}{\varepsilon_0} - \frac{1}{\varepsilon_0}\nabla\cdot\boldsymbol{P}.$$

Rearranging this we have

$$\nabla\cdot(\varepsilon_0\boldsymbol{E}+\boldsymbol{P}) = \rho_f.$$

We define the **electric displacement field** as

$$\boldsymbol{D} = \varepsilon_0\boldsymbol{E}+\boldsymbol{P}.$$

Thus Maxwell's first law in media is

$$\nabla\cdot\boldsymbol{D} = \rho_f,$$

of in integral form,

$$\oint_S \boldsymbol{D}\cdot\mathrm{d}\boldsymbol{S} = Q_{f,\mathrm{enc}}$$

where

$$Q_{f,\mathrm{enc}} = \int_V \rho_f\,\mathrm{d}V$$

is the free charge enclosed in the volume V, which is bounded by the surface S.

It is important to note that $\boldsymbol{D}$ is *not* an electric field, despite its seeming similarity. For example there is no equivalent of Coulomb's law for $\boldsymbol{D}$. Also for a static field

$$\nabla \times \boldsymbol{D} = \varepsilon \nabla \times \boldsymbol{E} + \nabla \times \boldsymbol{P} = \nabla \times \boldsymbol{P}$$

which is in general non-zero. This means that there is no scalar potential for $\boldsymbol{D}$.

15.4 Linear Isotropic Homogenous Media

A linear isotropic homogeneous (LIH) media is perhaps the simplest media that we may consider. The three key properties of LIH media are

- Linearity – $E \propto P$, specifically $P = \chi_E \varepsilon_0 E$ where χ_E is the (scalar) **electric susceptibility**.
- Isotropy – There is no preferred direction, so by symmetry $\boldsymbol{P}$ is parallel to $\boldsymbol{E}$.
- Homogeneity – The medium is the same everywhere, which means that χ_E has no position dependence.

The displacement field in LIH media is

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} = \varepsilon_0 \boldsymbol{E} + \varepsilon_0 \chi_E \boldsymbol{E} = \varepsilon_0 (1 + \chi_E) \boldsymbol{E} = \varepsilon_0 \varepsilon_r \boldsymbol{E} = \varepsilon \boldsymbol{E}.$$

where $\varepsilon_r = \chi_E + 1$ is the **relative permittivity**, also known as the **dielectric constant**, and $\varepsilon = \varepsilon_0 \varepsilon_r$ is the **absolute permittivity**. In a vacuum $\varepsilon_r = 1$ so $\varepsilon = \varepsilon_0$. For an insulator ε_r =1.05 to 1.3, mica has $\varepsilon_r = 7$, and a polar fluid such as deionised water has $\varepsilon_r = 80$.

15.5 Dielectric In a Capacitor

A parallel plate capacitor with plate separation d is filled with an LIH dielectric with relative permittivity ε_r. The capacitor is charged so that each plate has a charge of magnitude Q. What is the effect of the dielectric compared to a vacuum?

The charge distribution on the capacitor plates causes a charge separation in the dielectric with a slight positive charge on the side of the dielectric nearest to the negative plate of the capacitor. For a parallel plate capacitor the electric field is simply the superposition of the field due to the free charges on the plate and the bound charges on the surface of the dielectric:

$$\boldsymbol{E} = \boldsymbol{E_0} + \boldsymbol{E_P} = \frac{1}{\varepsilon_0}(\sigma_f - \sigma_b)\hat{\boldsymbol{n}}.$$

Here $\boldsymbol{E_0}$ is the field that we would have from the capacitor without the dielectric, $\boldsymbol{E_P}$ is the field due to the polarisation of the dielectric, σ_f is the surface charge density on the capacitor plates and σ_b is the surface charge density on the dielectric. The negative in this case accounts for the fact that the dipole due to the dielectric is anti-parallel to the dipole due to the charged plates. Finally $\hat{\boldsymbol{n}}$ is normal to the plates and points from the positive plate to the negative plate. We see that the effect of the dielectric is to reduce the net electric field strength in the capacitor.

The displacement field is

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} = \sigma_f \hat{\boldsymbol{n}}.$$

This can be shown by using Gauss' law for the displacement field,

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int_V \rho_f \, \mathrm{d}V = Q_{f,\mathrm{enc}},$$

and the fact that $Q_{f,\mathrm{enc}} = A\sigma_f$ where A is the area of the plate. We use a Gaussian pillbox with area A and we find that $A|\boldsymbol{D}| = A\sigma_f$. The direction is the deduced from the way the charges in the dielectric distribute. The capacitance of the capacitor is then given in terms of the potential difference,

$$V_d = -\int_1^2 \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l}\,,$$

where the bounds are taken to be the two plates. We integrate along the normal to the plates and we get

$$V_d = Ed = \frac{Dd}{\varepsilon_0 \varepsilon_r}$$

so

$$C = \frac{Q}{Ed} = \frac{A\sigma_f}{Ed} = \frac{AD}{Ed} = \frac{A\varepsilon_r\varepsilon_0}{d} = \varepsilon_r C_0$$

where C_0 is the capacitance of the capacitor without the dielectric. Since $\chi_E \geq 0$ we know that $\varepsilon_r = \chi_E + 1 \geq 1$ and therefore the capacitance increases when we add a dielectric. It turns out that for all capacitor geometries the capacitance will increase upon addition of a dielectric however for different geometries the increase may be by an amount other than ε_r.

15.5.1 Partially Filled Capacitors

Suppose we only partially fill the capacitor with dielectric. There are two interesting ways to do this, shown in figure 15.1. For the case of the capacitor half filled as shown on the left the addition of the

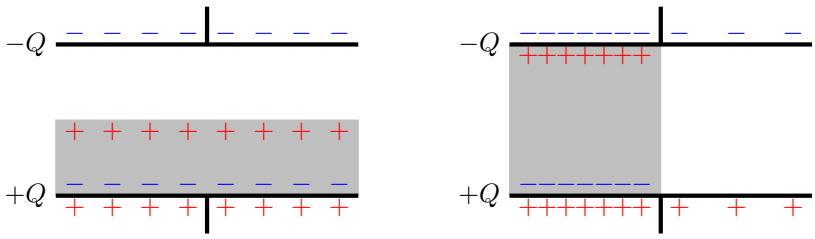


Figure 15.1: The two interesting ways to half fill a capacitor with dielectric.

dielectric does not break the planar symmetry of the situation and therefore σ_f is homogeneous. The high symmetry allows us to apply Gauss' law in media and from this we can get the displacement field and then from that we can get quantities of interest such as the electric field, polarisation field, bound charge density, potential, etc. For the case of the capacitor half filled as shown on the right the planar symmetry is broken and this means that Gauss' law is not as useful. By the symmetry of the situation we know that within the two regions, in the dielectric and outside the dielectric, the electric field is homogeneous. The free charge density, σ_f, is inhomogeneous as the polarisation field distorts it. We therefore need to take a different approach to find the electric field we can construct equipotentials and then from the electric field we can find the displacement field and other interesting quantities.

16 The Magnetic Field in Media

16.1 Types of Magnetisation

When an external magnetic field is applied to a material it will produce a magnetisation response. There are three mechanisms by which this can happen:

- Diamagnetism – All materials have a diamagnetic response but it is only important in materials without an intrinsic magnetic moment as it is such a weak effect. What happens is the external magnetic field alters the angular momentum of the electrons which induces a field that opposes the external field.
- Paramagnetism – This effect is important in materials in which each atom/molecule has an intrinsic magnetic moment and they are free to move. When an external field is applied these individual magnetic moments align with the field and create an induced field parallel to the applied field.
- Ferromagnetism – This effect occurs in only a few materials. In these materials the intrinsic magnetic moments of individual species can align and stay aligned. However this is a local effect and the material will become a mosaic of 'domains' where in each domain the magnetic moments of all species are aligned. When an external field is applied the domain boundaries will change in a way that favours domains where the magnetic moment is aligned with the external field. When the magnetic field is removed the domain boundaries do not return to where they where. This is the process by which a permanent magnet can be created.

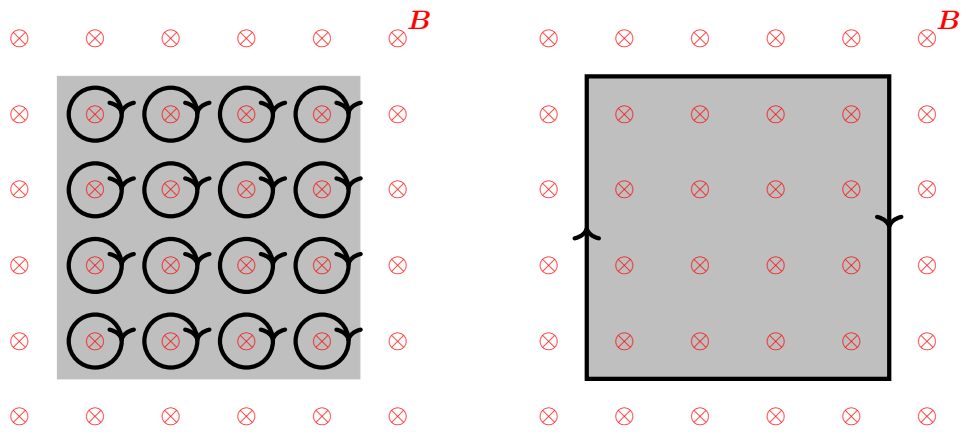


Figure 16.1: Individual magnetic moments viewed as microscopic current loops vs. the net magnetic moment viewed as a macroscopic current loop.

16.2 Magnetisation Field

The **magnetisation field**, $\boldsymbol{M}$, is the net magnetisation dipole density in volume $\mathrm{d}V$:

$$\mathrm{d}\boldsymbol{m} = \boldsymbol{M}\,\mathrm{d}V$$

where $\mathrm{d}\boldsymbol{m}$ is the neg magnetic dipole moment due to the material in $\mathrm{d}V$. This is analogous to the definition of $\boldsymbol{P}$ in section 15.2.

We can think of the magnetic dipole at a point as being the result of a small current loop. These then combine to give a macroscopic effect. Figure 16.1 shows the net effect of many individual magnetic moments. In this case the magnetic field is homogenous and this results in there being no current in the material, $\boldsymbol{J_M} = \boldsymbol{0}$. There is only a surface current, $\boldsymbol{j_M}$. If the field were not homogenous then there would be an internal current. The magnetic moments for an inhomogeneous magnetic field are shown in figure 16.2 In this case the eddy currents do not cancel and we have an internal current, $\boldsymbol{J_M}$, as well as

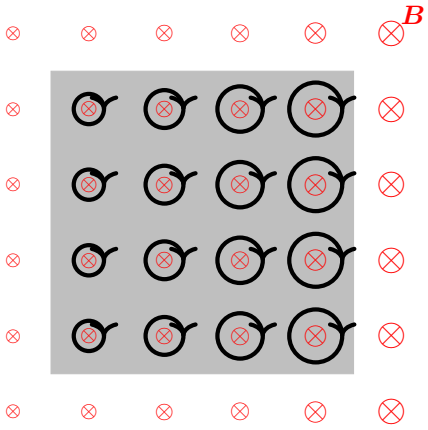


Figure 16.2: Individual magnetic moments due to an inhomogeneous magnetic field.

the surface current, $\boldsymbol{j_M}$.

We now want to quantify these effects. To do this we use the magnetic vector potential at $\boldsymbol{r}$ due to an ideal magnetic dipole, $\boldsymbol{m}$, at $\boldsymbol{r'}$:

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{\boldsymbol{m}\times(\boldsymbol{r}-\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}.$$

This generalises by superposition to

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\int_V \frac{\boldsymbol{M}\times(\boldsymbol{r}-\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3}\,\mathrm{d}^3r'\,.$$

We then use the identity

$$\frac{\boldsymbol{r}-\boldsymbol{r'}}{|\boldsymbol{r}-\boldsymbol{r'}|^3}=\nabla'\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}$$

where ∇' is the gradient operator that acts only on $\boldsymbol{r'}$. Hence

$$\boldsymbol{A}(\boldsymbol{r})=\frac{\mu_0}{4\pi}\int_V \boldsymbol{M}(\boldsymbol{r'})\times\nabla'\left[\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|}\right]\mathrm{d}^3r'\,.$$

We then use the product rule, $\nabla\times(f\boldsymbol{K})=f\nabla\times\boldsymbol{K}-\boldsymbol{K}\times\nabla f$, to get

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{r})&=\frac{\mu_0}{4\pi}\int_V\frac{\nabla'\times\boldsymbol{M}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\mathrm{d}^3r'-\frac{\mu_0}{4\pi}\int_V\nabla\times\left[\frac{\boldsymbol{M}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\right]\mathrm{d}^3r'\\&=\frac{\mu_0}{4\pi}\int_V\frac{\nabla'\times\boldsymbol{M}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\mathrm{d}^3r'-\frac{\mu_0}{5\pi}\oint_S\left[\frac{\boldsymbol{M}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\right]\times\mathrm{d}\boldsymbol{S'}\,.\end{aligned}$$

Recall that

$$\nabla^2\boldsymbol{A}=\mu_0\boldsymbol{J}\implies\boldsymbol{A}=\frac{\mu_0}{4\pi}\int_V\frac{\boldsymbol{J}(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|}\mathrm{d}^3r'\,.$$

This allows us to interpret the numerators of both integrands as currents. First

$$\boldsymbol{J_M}=\nabla\times\boldsymbol{M}$$

is the bulk magnetisation current density, second

$$\boldsymbol{j_M}=\boldsymbol{M}\times\hat{\boldsymbol{n}}$$

is the surface current density where $\hat{\boldsymbol{n}}$ is normal to the surface S' which bounds the volume V'.

Example 16.1. A cylindrical bar magnet has uniform magnetisation, M, along its axis. To what current distribution is this equivalent?

$\boldsymbol{M}$ is uniform so $\nabla\times\boldsymbol{M}=\boldsymbol{0}$ meaning $\boldsymbol{J_M}=0$. The surface current density is then

$$\boldsymbol{j_M}=\boldsymbol{M}\times\hat{\boldsymbol{n}}=M\boldsymbol{e_z}\times\boldsymbol{e_\rho}=M\boldsymbol{e_\varphi}.$$

This has magnitude M and is 'solenoidal', i.e. it resembles a solenoid with current flow around the circumference of a cylinder.

Example 16.2. A long cylindrical bar magnet of uniform magnetisation is bent into a loop. To what current distribution is this equivalent?

The curl in cylindrical coordinates is

$$\nabla\times\boldsymbol{M}=\left[\frac{1}{\rho}\frac{\partial M_z}{\partial\varphi}-\frac{\partial M_\varphi}{\partial z}\right]\boldsymbol{e_\rho}+\left[\frac{\partial M_\rho}{\partial z}-\frac{\partial M_z}{\partial\rho}\right]\boldsymbol{e_\varphi}+\frac{1}{\rho}\left[\frac{\partial}{\partial\rho}(\rho M_\varphi)-\frac{\partial M_\rho}{\partial\varphi}\right]\boldsymbol{e_z}.$$

With this setup $\boldsymbol{M}$ is circumferential so $\boldsymbol{M}=M\boldsymbol{e_\varphi}$ so $M_\rho=M_z=0$ meaning that the curl reduces to

$$\nabla\times\boldsymbol{M}=\frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho M)=\frac{M}{\rho}\boldsymbol{e_z}=\boldsymbol{J_M}.$$

The surface current has magnitude M and wraps around the cylinder still. The surface current is 'toroidal', i.e. it resembles a toroidal solenoid with current flow over the surface. The bulk current density makes up for the fact that due to the geometry the net surface current is greater on the outside of the torus than the inside due to the greater surface area.

16.3 Ampère's Law in Media

The Ampère–Maxwell law in a vacuum is

$$\nabla \times \boldsymbol{B} = \mu_0(\boldsymbol{J} + \varepsilon \partial_t \boldsymbol{E}).$$

We divide $\boldsymbol{J}$ into three parts: $\boldsymbol{J_f}$ which is the current due to free charges, this is the normal current with which we have dealt so far, $\boldsymbol{J_M} = \nabla \times \boldsymbol{M}$ which is the magnetisation current as defined in the previous section, and $\boldsymbol{J_P}$ which is the polarisation current which accounts for the movement of electric dipoles. To rationalise the introduction of $\boldsymbol{J_P}$ we define a new charge density, $\rho_P = -\nabla \cdot \boldsymbol{P}$, which follows the continuity equation

$$\partial_t \rho_P + \nabla \cdot \boldsymbol{J_P} = 0.$$

From this we have

$$\boldsymbol{J_P} = \partial_t \boldsymbol{P}.$$

We aim to write the Ampère–Maxwell law in terms of $\boldsymbol{J_f}$ only:

$$\begin{aligned}\nabla \times \boldsymbol{B} &= \mu_0(\boldsymbol{J_f} + \boldsymbol{J_M} + \boldsymbol{J_P} + \varepsilon_0 \partial_t \boldsymbol{E}) \\ &= \mu_0(\boldsymbol{J_f} + \nabla \times \boldsymbol{M} + \partial_t \boldsymbol{P} + \varepsilon_0 \partial_t \boldsymbol{E}) \\ &= \mu_0(\boldsymbol{J_f} + \nabla \times \boldsymbol{M} + \partial_t \boldsymbol{D})\end{aligned}$$

From this we have

$$\nabla \times \left[\frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}\right] = \boldsymbol{J_f} + \partial_t \boldsymbol{D}.$$

We then define

$$\boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M}$$

so

$$\nabla \times \boldsymbol{H} = \boldsymbol{J_f} + \partial_t \boldsymbol{D}.$$

This is the Ampère–Maxwell law in media. In integral form it reads

$$\oint_C \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \int_S (\boldsymbol{J_f} + \partial_t \boldsymbol{D}) \cdot \mathrm{d}\boldsymbol{S} = I_{f,\mathrm{enc}},$$

where C is the contour bounding the surface S.

Confusingly $\boldsymbol{H}$ is often referred to as the 'magnetic field' and so is $\boldsymbol{B}$. In some texts $\boldsymbol{B}$ is the 'magnetic flux density' and $\boldsymbol{H}$ is the 'magnetic field strength' but in other texts $\boldsymbol{B}$ is the 'magnetic field' and $\boldsymbol{H}$ is the 'auxiliary field'. For this reason it is best to specify 'the magnetic $\boldsymbol{B}$ field' or 'the magnetic $\boldsymbol{H}$ field'.

Maxwell's second and third laws do not need modification in media as they contain no source terms, ρ or $\boldsymbol{J}$. This means that we now have the four Maxwell equations in media:

$$\begin{aligned}\nabla \cdot \boldsymbol{D} &= \rho_f && \text{(MI in media)} \\ \nabla \cdot \boldsymbol{B} &= 0 && \text{(MII)} \\ \nabla \times \boldsymbol{E} &= -\partial_t \boldsymbol{B} && \text{(MIII)} \\ \nabla \times \boldsymbol{H} &= \boldsymbol{J_f} + \partial_t \boldsymbol{D} && \text{(MIV in media)}\end{aligned}$$

We define the **magnetic susceptibility**, χ_M, to describe the relationship between $\boldsymbol{M}$ and $\boldsymbol{H}$:[11]

$$\boldsymbol{M} = \chi_M \boldsymbol{H}.$$

In general χ_M is a tensor however in LIH media it is a scalar. From this we also have

$$\boldsymbol{B} = \mu_r \mu_0 \boldsymbol{H} = \mu_0(1 + \chi_M)\boldsymbol{H} = \mu \boldsymbol{H}$$

where $\mu_r = 1 + \chi_M$ is the **relative permeability** and $\mu = \mu_0 \mu_r$ is the **absolute permeability**. In the absence of magnetisation $\chi_M = 0$ and $\mu_r = 1$. Unlike with dielectrics χ_M can be positive or negative and consequently μ_r is unbounded. For example a diamagnetic material will have $\chi_M < 0$ whereas a paramagnetic material will have $\chi_M > 0$ which corresponds to the fact that the fields induced in these two materials will be in opposite directions.

[11] some texts use $\chi_B \boldsymbol{B} = \mu_0 \boldsymbol{M}$ instead, these are related by $\chi_B = \chi_M/(1 + \chi_M)$

16.4 Media in Solenoids

A solenoid is filled with LIH medium with magnetic susceptibility χ_M. What is the effect compared to a solenoid containing a vacuum?

The magnetisation field is given by $\boldsymbol{M} = \chi_M \boldsymbol{H}$. The inductance of a solenoid is given by $L = \Phi_B / I$. We first get $\boldsymbol{H}$ from

$$\nabla \times \boldsymbol{H} = \boldsymbol{J_f} + \partial_t \boldsymbol{D} \implies \oint_C \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = I_{f,\mathrm{enc}}.$$

By symmetry we know that $\boldsymbol{H}$ is axial and constant inside the solenoid and zero outside the solenoid. Thus if we define an Ampérian loop we only need integrate along the part inside the solenoid parallel to the axis. If this part has length l then

$$\oint_C \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = Hl = I_{f,\mathrm{enc}} = nlI \implies H_z = nI$$

where n is the number of loops the solenoid has per unit length. From this we have

$$\boldsymbol{B} = \mu_0 \mu_r \boldsymbol{H} = \mu_0 (1 + \chi_M) hI \boldsymbol{e_z}.$$

This allows us to calculate the flux through the inductor treating the inductor as n loops of area A per unit length we have

$$\Phi_B = ABnl = \mu_0 (1 + \chi_M) n^2 AlI$$

where A is the cross sectional area of the solenoid. So the inductance is

$$L = \frac{\Phi_B}{I} = \mu_0 (1 + \chi_M) n^2 Al = \mu_0 (1 + \chi_M) n^2 V_s$$

where $V_s = Al$ is the volume of the solenoid.

16.4.1 Partially Filled Solenoids

There are two interesting ways to half fill a solenoid. The first is with a cylindrical core which has a radius smaller than the solenoid. This preserves the cylindrical symmetry and therefore we can use Ampère's law with two different cases, one that reaches all the way to the media and the other which stops in the vacuum.

The second is with a cylindrical core that is of the same radius as the solenoid but doesn't extend for the solenoids entire length. This breaks the cylindrical symmetry. We can view this as two contributions, one from the filled section and one from the empty section. We cannot use Ampère's law in this scenario to find the field.

17 Electromagnetism in Media

17.1 Summary

Maxwell's equations for the macroscopic fields, $\boldsymbol{D}$, $\boldsymbol{B}$, $\boldsymbol{E}$, and $\boldsymbol{H}$, in media with free charge density, ρ_f, and free current density, $\boldsymbol{J_f}$, are

$$\begin{aligned}
\nabla \cdot \boldsymbol{D} &= \rho_f, && \text{(MI in media)} \\
\nabla \cdot \boldsymbol{B} &= 0, && \text{(MII)} \\
\nabla \times \boldsymbol{E} &= -\partial_t \boldsymbol{B}, && \text{(MIII)} \\
\nabla \times \boldsymbol{H} &= \boldsymbol{J_f} + \partial_t \boldsymbol{D}, && \text{(MIV in media)}
\end{aligned}$$

Where the fields, $\boldsymbol{D}$ and $\boldsymbol{H}$, are defined as

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}, \qquad \text{and} \qquad \boldsymbol{B} = \mu_0 (\boldsymbol{H} + \boldsymbol{M}).$$

The free charges and currents satisfy the continuity equation

$$\partial_t \rho_f + \nabla \cdot \boldsymbol{J_f} = 0.$$

In LIH media the relations between the fields are

$$\boldsymbol{P} = \chi_E \varepsilon_0 \boldsymbol{E}, \qquad \boldsymbol{M} = \chi_M \boldsymbol{H}, \qquad \boldsymbol{D} = \varepsilon_0 \varepsilon_r \boldsymbol{E} = \varepsilon \boldsymbol{E},$$

$$\text{and} \qquad \boldsymbol{B} = \mu_0 \mu_r \boldsymbol{H} = \mu \boldsymbol{H}.$$

Where $\varepsilon_r = 1 + \chi_E$ and $\mu_r = 1 + \chi_M$.

17.2 Energy Density and the Poynting Vector

The power delivered by an EM field to a system of charge carriers is $\boldsymbol{E} \cdot \boldsymbol{J_f}$ per unit volume. This means that the energy density, u, obeys

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \boldsymbol{E} \cdot \boldsymbol{J_f}.$$

We aim to express $\boldsymbol{J_f}$ in field related quantities. First we use Maxwell's fourth law to write

$$\boldsymbol{E} \cdot \boldsymbol{J_f} = \boldsymbol{E} \cdot [\nabla \times \boldsymbol{H} - \partial_t \boldsymbol{D}] = \boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) - \boldsymbol{E} \cdot \partial_t \boldsymbol{D}.$$

We then use the product rule,

$$\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) = \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) - \boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) \implies \boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) - \boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) - \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}),$$

to get

$$\boldsymbol{E} \cdot \boldsymbol{J} = \boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) - \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) - \boldsymbol{E} \cdot \partial_t \boldsymbol{D}.$$

Using Maxwell's third law this becomes

$$\boldsymbol{E} \cdot \boldsymbol{J_f} = -\boldsymbol{H} \cdot \partial_t \boldsymbol{B} - \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) - \boldsymbol{E} \cdot \partial_t \boldsymbol{D}.$$

Next we use that in LIH media

$$\boldsymbol{E} \cdot \partial_t \boldsymbol{D} = \boldsymbol{E} \cdot \partial_t (\varepsilon \boldsymbol{E}) = \varepsilon \boldsymbol{E} \cdot \partial_t \boldsymbol{E} = \boldsymbol{D} \cdot \partial_t \boldsymbol{E}$$

and

$$\boldsymbol{H} \cdot \partial_t \boldsymbol{B} = \mu \boldsymbol{B} \cdot \partial_t \boldsymbol{B} = \boldsymbol{B} \cdot \partial_t (\mu \boldsymbol{B}) = \boldsymbol{B} \cdot \partial_t \boldsymbol{H}.$$

This gives us

$$\boldsymbol{E} \cdot \boldsymbol{J_f} = -\boldsymbol{B} \cdot \partial_t \boldsymbol{H} - \boldsymbol{D} \cdot \partial_t \boldsymbol{E} - \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}).$$

Recognising the standard product rule

$$\partial_t (\boldsymbol{E} \cdot \boldsymbol{D}) = \boldsymbol{D} \cdot \partial_t \boldsymbol{E} + \boldsymbol{E} \cdot \partial_t \boldsymbol{D} = 2\boldsymbol{D} \cdot \partial_t \boldsymbol{E}$$

and similarly for the magnetic fields this becomes

$$\boldsymbol{E} \cdot \boldsymbol{J_f} = \frac{1}{2} \partial_t (\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{B} \cdot \boldsymbol{H}).$$

We then integrate over a volume, V, to get the energy from the energy density. Using the divergence theorem on the second term we get

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \int_V (\boldsymbol{E} \cdot \boldsymbol{D} + \boldsymbol{B} \cdot \boldsymbol{H}) \, \mathrm{d}V - \oint_A (\boldsymbol{E} \times \boldsymbol{H}) \cdot \mathrm{d}\boldsymbol{A}\,.$$

We now identify the electric and magnetic energy densities as

$$u_E = \frac{1}{2} \boldsymbol{E} \cdot \boldsymbol{D}, \qquad \text{and} \qquad u_M = \frac{1}{2} \boldsymbol{B} \cdot \boldsymbol{H}$$

respectively. From the second term we identify the macroscopic Poynting vector as

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$$

so the power delivered to the charges can be written as

$$\frac{\mathrm{d}U}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t} [U_E + U_M] - \oint_A \boldsymbol{S} \cdot \mathrm{d}\boldsymbol{A}\,.$$

17.3 Boundary Conditions

17.3.1 Qualitatively

Now that we have characterised fields in and out of media we want to know how the fields behave at the boundary between media. To do this we use the work we have done already with half filled capacitors/inductors and we make the unjustified assumption that the same analysis holds outside of a capacitor/inductor. Figure 17.1 shows qualitatively what happens at a boundary for the four fields, this figure assumes that $\chi_M > 0$. The two cases we consider for each field are what happens to the tangential component and the normal component at the boundary.

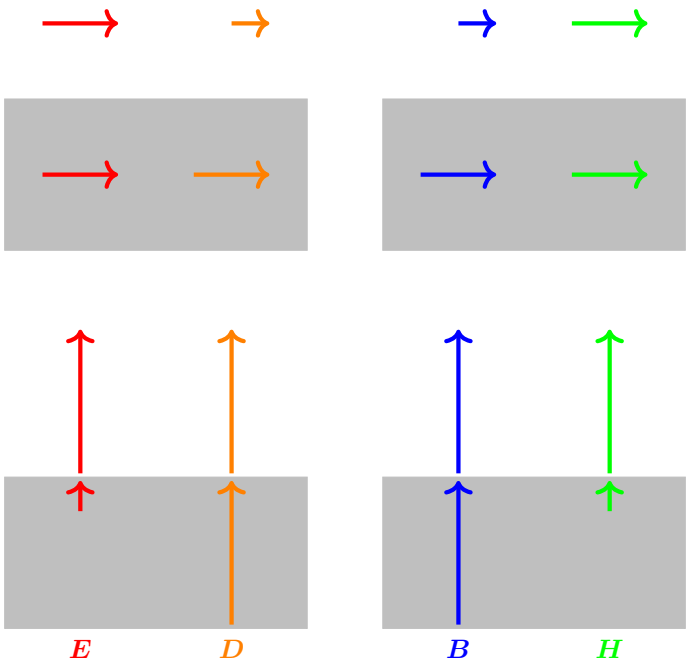


Figure 17.1: The qualitative behaviour of fields at a boundary between media. The normal and tangential components of the four fields are shown both in the media and outside.

17.3.2 Quantitatively

We can be much more rigorous if we use Maxwell's laws. We assume we have a plane and above and below the plane are two different media, media 1 below and media 2 above. We use a Gaussian surface that is a cylindrical pill box of height h with the flat faces parallel to the plane, and a rectangular Ampérian loop of height h with two sides running parallel to the plane. We also define two vectors, $\hat{\boldsymbol{n}}$, and $\hat{\boldsymbol{t}}$, which are normal and tangential to the plane respectively.

The first condition we consider is Maxwell's first law,

$$\nabla \cdot \boldsymbol{D} = \rho_f.$$

From this and the divergence theorem we have

$$\oint_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = Q_{f,\mathrm{enc}}.$$

pill box has a surface normal, on the flat faces, of $\mathrm{d}\boldsymbol{S} = \hat{\boldsymbol{n}}\,\mathrm{d}S$. Allowing $h \to 0$ we can ignore the contribution to the integral from the sides of the pill box and we get

$$\int \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int (\boldsymbol{D_2} - \boldsymbol{D_1}) \cdot \hat{\boldsymbol{n}}\,\mathrm{d}S = \int \sigma_f\,\mathrm{d}S$$

where the last equality arises from the definition of a surface charge density. Note that $\boldsymbol{D_1}$ is the $\boldsymbol{D}$ field below the plane and $\boldsymbol{D_2}$ is the $\boldsymbol{D}$ field above the plane. Since both integrals are over the same area we can conclude that

$$(\boldsymbol{D_2} - \boldsymbol{D_1}) \cdot \hat{\boldsymbol{n}} = \sigma_f.$$

Taking the scalar product with $\hat{\boldsymbol{n}}$ picks out the scalar component of $\boldsymbol{D_2}-\boldsymbol{D_1}$ and so $\boldsymbol{D}_{\mathbf{normal}}$ is continuous *if and only if* $\sigma_f = 0$.

The second condition we consider is Maxwell's second law,

$$\nabla \cdot \boldsymbol{B} = 0 \implies \oint_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0.$$

Using the same Gaussian surface and again allowing $h \to 0$ we have

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = \int (\boldsymbol{B_2} - \boldsymbol{B_1}) \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = 0.$$

This holds for any similarly defined surface so

$$(\boldsymbol{B_2} - \boldsymbol{B_1}) \cdot \hat{\boldsymbol{n}} = 0$$

meaning that $\boldsymbol{B}_{\mathbf{normal}}$ is *always* continuous.

The third condition we consider is Maxwell's third law,

$$\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B} \implies \oint_C \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\partial_t \Phi_B.$$

We use an Ampérian loop of length ℓ and height h. Allowing $h \to -$ we also have $\Phi_B \to 0$ as long as B is finite. Therefore

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = \int (\boldsymbol{E_2} - \boldsymbol{E_1}) \cdot \hat{\boldsymbol{t}} \, \mathrm{d}l = 0$$

where $\hat{\boldsymbol{t}}$ is tangential to the surface meaning that $\hat{\boldsymbol{t}} \cdot \hat{\boldsymbol{n}} = 0$. Since this holds no matter where we put the Ampérian loop, or define $\hat{\boldsymbol{t}}$ we have

$$(\boldsymbol{E_2} - \boldsymbol{E_1}) \cdot \hat{\boldsymbol{t}} = 0$$

meaning that $\boldsymbol{E}_{\mathbf{tangential}}$ is *always* continuous.

The final condition that we consider is Maxwell's fourth law

$$\nabla \times \boldsymbol{H} = \boldsymbol{J_f} + \partial_t \boldsymbol{D}$$

in integral form this becomes

$$\oint_C \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \boldsymbol{j_f} \cdot \hat{\boldsymbol{s}}\ell + (\partial_t \boldsymbol{D}) \cdot \hat{\boldsymbol{s}}\ell h$$

where $\boldsymbol{j_f}$ is the free surface current per unit area and $\hat{\boldsymbol{s}} = \hat{\boldsymbol{t}} \times \hat{\boldsymbol{n}}$ is the surface unit vector perpendicular to the Ampérian loop. Now allowing $h \to 0$ we have

$$\int \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = \boldsymbol{H_1} - \boldsymbol{H_2}) \cdot \hat{\boldsymbol{t}}\ell = \hat{\boldsymbol{j_f}} \cdot \hat{\boldsymbol{s}}\ell.$$

Since $\hat{\boldsymbol{n}}$ is normal to the plane $\hat{\boldsymbol{s}}$ is tangential to the plane so $\boldsymbol{H}_{\mathbf{tangential}}$ is continuous *if and only if* $\boldsymbol{j_f} = \boldsymbol{0}$. This general form can be written in other ways, including

$$\begin{aligned}(\boldsymbol{H_1} - \boldsymbol{H_2}) \cdot \hat{\boldsymbol{t}} &= \boldsymbol{j_f} \cdot \hat{\boldsymbol{s}}, \\ (\boldsymbol{H}_{\mathbf{2,tangential}} - \boldsymbol{H}_{\mathbf{1,tangential}}) &= \boldsymbol{j_f} \times \hat{\boldsymbol{n}}, \\ (\boldsymbol{H_2} - \boldsymbol{H_1}) \times \hat{\boldsymbol{n}} &= -\boldsymbol{j_f}.\end{aligned}$$

In summary at a boundary between media

- $\boldsymbol{D}_{\mathbf{normal}}$ is continuous if and only if $\sigma_f = 0$.
- $\boldsymbol{B}_{\mathbf{normal}}$ is continuous always.
- $\boldsymbol{E}_{\mathbf{tangential}}$ is continuous always.
- $\boldsymbol{H}_{\mathbf{tangential}}$ is continuous if and only if $\boldsymbol{j_f} = \boldsymbol{0}$.

18 Continuity Conditions and Waves in Media

18.1 Applications of Continuity Conditions

18.1.1 Inclined Dielectric

Consider the setup in figure 18.1. It shows a uniform electric field, $\boldsymbol{E^o}$, incident on a dielectric at an

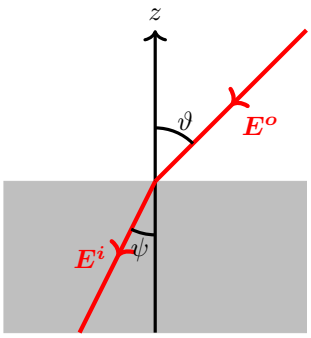


Figure 18.1: An inclined dielectric with an incident electric field.

angle ϑ. The field is then refracted and the field in the dielectric is at an angle ψ. We want to know what the field, $\boldsymbol{E^i}$, inside the dielectric is.

Outside the dielectric we define the displacement field as

$$\boldsymbol{D^o} = \varepsilon_0 \boldsymbol{E^o}.$$

Inside the dielectric the displacement field is

$$\boldsymbol{D^i} = \varepsilon_0 \varepsilon_r \boldsymbol{E^i}.$$

We apply the boundary condition that the normal component of the displacement field, D_n, is continuous at the boundary, since there is no surface charge. This means that $D_z^o = D_z^i$. The second boundary condition is that the tangential components of the electric field, E_t, are continuous at the boundary. This means that $E_x^o = E_x^i$ and $E_y^o = E_y^i$. We choose a coordinate system such that $E_y^o = E_y^i = 0$ meaning that we can work in the two-dimensional (x, z)-plane. In this plane the full fields are

$$\boldsymbol{E^o} = (E^o \sin\vartheta, E^o \cos\vartheta), \qquad \text{and} \qquad \boldsymbol{E^i} = (E^i \sin\psi, E^i \cos\psi).$$

Applying the first continuity condition we have

$$D_z^o = D_i^z \implies \varepsilon_0 E^o \cos\vartheta = \varepsilon_0 \varepsilon_r E^i \cos\psi \implies E^i = \frac{1}{\varepsilon_r} E^o \frac{\cos\vartheta}{\cos\psi}.$$

Applying the second continuity condition we have

$$E_x^o = E_x^i \implies E^o \sin\vartheta = E^i \sin\psi \implies E^i = E^o \frac{\sin\vartheta}{\sin\psi}.$$

Combining these we have

$$\frac{1}{\varepsilon_r}\frac{\cos\vartheta}{\cos\psi} = \frac{\sin\vartheta}{\sin\psi} \implies \frac{\sin\psi}{\cos\psi} = \tan\psi = \varepsilon_r \frac{\sin\vartheta}{\cos\vartheta} = \varepsilon_r \tan\vartheta \implies \psi = \arctan(\varepsilon_r \tan\vartheta).$$

It is worth checking two basic case here. First we consider the case when $\vartheta = 0$. In this case $\psi = \arctan(\varepsilon_r \tan 0) = \arctan 0 = 0$ and so there is no refraction which is what we would expect. Going a step further back in the calculation if $\psi = \vartheta = 0$ then we recover that $D_z^o = D_z^i$. The second case we consider is $\vartheta = \pi/2$. In this case $\psi = \arctan(\varepsilon_r \tan(\pi/2)) = \arctan(\infty) = \pi/2$ and so there is no refraction as the field travels along the boundary. Going a step further back in the calculation if $\psi = \vartheta = \pi/2$ then we recover $E^o = E^i$.

18.1.2 Spherical Cavity in a Dielectric

Suppose we have a large block of LIH dielectric which contains a spherical cavity. The electric field far from the cavity is uniform and has magnitude E_0. What are $\boldsymbol{E}$ and $\boldsymbol{D}$ in the cavity?

We will use spherical coordinates with their origin at the centre of the cavity and aligned so that the z-axis is parallel to the field at large r. See figure 18.2 for a diagram. At the surface of the cavity a

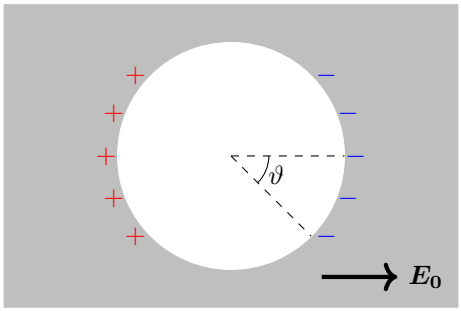


Figure 18.2: A spherical cavity in a dielectric

charge distribution, $\sigma_P = \boldsymbol{P} \cdot \hat{\boldsymbol{n}}$ forms. Note that $\hat{\boldsymbol{n}}$ is the outward normal of the dielectric, meaning it points *towards* the origin. Since this is an LIH media $\boldsymbol{P}$ is parallel to $\boldsymbol{E}$ and so the field inside is enhanced by the charge distribution which is a function of the angle, ϑ, since $r = 1.5$ on the surface of the cavity and the symmetry under changing φ means that σ_P cannot depend on φ.

The charge distribution, $\sigma_P(\vartheta)$, forms an effective dipole. Outside of the cavity the field lines are locally distorted by this charge distribution. We want to determine V, the electrostatic potential, and from this we can calculate $\boldsymbol{E}$. Within the cavity we assume a uniform $\boldsymbol{E}$ field in the z direction. This corresponds to a potential

$$V(r) = -E_{\text{in}} z = -E_{\text{in}} r \cos\vartheta$$

for $r < a$ where a is the radius of the cavity. Outside of the cavity we use a superposition of a uniform $\boldsymbol{E_0}$ field plus a dipole field so

$$V(r) = -E_0 r \cos\vartheta + \frac{A\cos\vartheta}{r^2}.$$

Here A is a constant to be found which gives the relative strength of the dipole field.

Thanks to the uniqueness theorem for Poisson's equation we need only show that this potential fulfils the boundary conditions and Poisson's equation and then we know that the electric field derived from this potential is unique. Since the charges are only on the boundary away from the boundary we only need the potential to satisfy Laplace's equation, $\nabla^2 V = 0$. We then only need to check that the boundary conditions are satisfied. In spherical coordinates

$$\nabla V = \boldsymbol{e_r}\frac{\partial V}{\partial r} + \boldsymbol{e_\vartheta}\frac{1}{r}\frac{\partial V}{\partial \vartheta} + \boldsymbol{e_\varphi}\frac{1}{r\sin\vartheta}\frac{\partial V}{\partial \vartheta} = -\boldsymbol{E}.$$

The first boundary condition we check is that E_t is continuous. This requires E_ϑ to be continuous at $r = a$. So

$$-E_0 a\sin\vartheta + \frac{A\sin\vartheta}{a^2} = -E_{\text{in}} a\sin\vartheta.$$

The second boundary condition is that D_n is continuous (note the surface density is due to polarisation, there is no free charge distribution). This requires $D_r = \varepsilon_r E_r = -\varepsilon_r \partial_r V$ to be continuous. So

$$\varepsilon_r E_o\cos\vartheta + \frac{2A\varepsilon_r\cos\vartheta}{a^3} = E_{\text{in}}\cos\vartheta.$$

Combining these we have

$$E_{\text{in}} = \varepsilon_r\left(E_0 + \frac{2A}{a^3}\right) = E_0 - \frac{A}{a^3}.$$

Thus

$$E_{\text{in}} = E_0\frac{3\varepsilon_r}{1+2\varepsilon_r}.$$

Combining this with the requirement that $\boldsymbol{E_{\text{in}}} = E_{\text{in}}\boldsymbol{e_z}$ and $\boldsymbol{D_{\text{in}}} = \varepsilon_0\boldsymbol{E_{\text{in}}}$ (since inside the cavity $\varepsilon_r = 1$), we have the full field. Notice that $E_{\text{in}} > E_0$ so the field in the cavity is stronger than the field outside.

18.2 Waves in Media

18.2.1 Non-conduction Media

In a non-conducting media $\rho_f = 0$ and $\boldsymbol{J_f} = 0$. Thus

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \qquad \text{and} \qquad \boldsymbol{B} = \mu \boldsymbol{H}.$$

We can derive macroscopic wave equations as before:

$$\nabla^2 \boldsymbol{E} = \varepsilon\mu \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \qquad \text{and} \qquad \nabla^2 \boldsymbol{B} = \varepsilon\mu \frac{\partial^2 \boldsymbol{B}}{\partial t^2}.$$

Clearly these have the normal plane wave solutions

$$\boldsymbol{E} = \boldsymbol{E_0} e^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)}, \qquad \text{and} \qquad \boldsymbol{B} = \boldsymbol{B_0} e^{i(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t)}.$$

Where $k^2 = \mu\varepsilon\omega^2$. We interpret this as the wave phase velocity being

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\mu\varepsilon}}.$$

Recall that $c = (\mu_0\varepsilon_0)^{-1/2}$ so

$$v^2 = \frac{1}{\mu_r\varepsilon_r} c^2 = \frac{1}{n^2} c^2$$

where we have defined $n = \mu_r\varepsilon_r$ which is called the **refractive index** and is a property of the medium. As we did with waves in a vacuum it can be shown that

$$i\boldsymbol{k} \cdot \boldsymbol{E_0} = 0, \qquad \text{and} \qquad i\boldsymbol{k} \cdot \boldsymbol{B_0} = 0.$$

This means that $\boldsymbol{E}$ and $\boldsymbol{B}$ are perpendicular to the direction of propagation, $\boldsymbol{k}$, and therefore the wave is transverse. We see that for LIH media all of the extra complications that come with having media are neatly hidden away in ε and μ and have the net effect of changing the velocity of the wave.

18.2.2 Waves in Conductors

In conductors in general $\rho_f \neq 0$ and $\boldsymbol{J_f} \neq \boldsymbol{0}$. We will start with Ohm's law, $\boldsymbol{J} = \sigma \boldsymbol{E}$ and assume linear media so

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \qquad \text{and} \qquad \boldsymbol{B} = \mu \boldsymbol{H}.$$

Combining these with Maxwell's fourth equation gives us

$$\nabla \times \boldsymbol{H} = \boldsymbol{J_f} + \partial_t \boldsymbol{D} \implies \nabla \times \boldsymbol{B} = \mu\sigma \boldsymbol{E} + \mu\varepsilon \partial_t \boldsymbol{E}. \tag{18.1}$$

Taking the curl of Maxwell's third equation we have

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E} = -\nabla \times (\partial_t \boldsymbol{B}) = -\partial_t (\nabla \times \boldsymbol{B}).$$

Substituting for $\nabla \times \boldsymbol{B}$ from equation 18.1 we have

$$\partial_t \left(\mu\sigma \boldsymbol{E} + \mu\varepsilon \partial_t \boldsymbol{E}\right) = \nabla^2 \boldsymbol{E} - \nabla\left(\frac{\rho}{\varepsilon}\right) = \nabla^2 \boldsymbol{E}$$

where in the last equality we assume a uniform charge density meaning that $\nabla\rho = 0$. Rearranging this we have

$$\nabla^2 \boldsymbol{E} = \mu\varepsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu\sigma \frac{\partial \boldsymbol{E}}{\partial t}.$$

Notice that this is the wave equation with an additional term on the right. The origin of this additional term is the free current.

Taking the curl of Maxwell's fourth equation gives us

$$\nabla \times (\nabla \times \boldsymbol{B}) = \nabla(\nabla \cdot \boldsymbol{B}) - \nabla^2 \boldsymbol{B} = \mu\sigma \nabla \times \boldsymbol{E} + \mu\varepsilon \partial_t (\nabla \times \boldsymbol{E}).$$

Substituting for Maxwell's second and third laws we have

$$\nabla^2 \boldsymbol{B} = \mu\varepsilon \partial_t \frac{\partial^2 \boldsymbol{B}}{\partial t^2} + \mu\sigma \frac{\partial \boldsymbol{B}}{\partial t}.$$

19 Waves In Conductors

We saw in the last section that we could derive wave equations in conductors which, if we assume linear media, have the form

$$\nabla^2 \boldsymbol{E} = \mu\varepsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu\sigma \frac{\partial \boldsymbol{E}}{\partial t}.$$

We make the ansatz that this has the plane wave solution

$$\boldsymbol{E} = \tilde{\boldsymbol{E}} e^{i(\tilde{k}z - \omega t)}$$

where we assume the wave is travelling in the z direction. If we substitute this into the wave equation we get

$$\tilde{k}^2 = \mu\varepsilon\omega^2 + i\mu\sigma\omega.$$

This is the **dispersion relation**. To solve this we clearly need to consider some complex numbers. Let $\tilde{k} = k + i\kappa$ for some $k, \kappa \in \mathbb{R}$. Then if we equate real and imaginary parts in the dispersion relation we have

$$k^2 - \kappa^2 = \mu\varepsilon\omega^2, \qquad \text{and} \qquad 2k\kappa = \mu\sigma\omega.$$

The second of these has the solution $\kappa = \mu\sigma\omega/2k$ which allows us to eliminate κ from the first giving

$$k^4 - \left(\frac{\mu\sigma\omega}{2}\right)^2 = \mu\varepsilon\omega^2 k^2.$$

This is a quadratic in k^2 and has the solution

$$k^2 = \frac{1}{2}\mu\varepsilon\omega^2 = \frac{1}{2}\left[(\mu\varepsilon\omega^2)^2 + (\mu\sigma\omega)^2\right]^{1/2} = \frac{\mu\varepsilon\omega^2}{2}\left[\left(1 + \left[\frac{\sigma}{\varepsilon\omega}\right]^2\right)^{1/2} + 1\right],$$

where we choose the positive square root so that k^2 is positive, since $k \in \mathbb{R}$. We can then use this to obtain

$$\kappa^2 = \frac{\mu\varepsilon\omega^2}{2}\left[\left(1 + \left[\frac{\sigma}{\varepsilon\omega}\right]^2\right)^{1/2} - 1\right].$$

Hence

$$\boldsymbol{E} = \tilde{\boldsymbol{E}}_{\boldsymbol{0}} e^{-\kappa z} e^{i(kz - \omega t)}.$$

This describes exponential decay of the wave along z, which is the propagation direction. The wave is attenuated over a characteristic distance, called the **skin depth**, which is given by

$$\delta = \frac{1}{\kappa}.$$

This gives the depth at which the amplitude of the wave is e^{-1} times the amplitude at which the wave enters the conductor.

A good check to do here is consider the case when $\sigma = 0$, this corresponds to a vacuum. In this case $k^2 = \varepsilon\mu\omega^2$ and $\kappa^2 = 0$ so we revert to the vacuum solution.

19.1 Good and Poor Conductors

The ratio $\sigma/\varepsilon\omega$ is important in this result. Both $1/\omega$ and ε/σ have units of time so this quantity is a ratio of two timescales. The question now is what do these timescales represent?

Consider the continuity equation for free charges,

$$\partial_t \rho_f + \nabla \cdot \boldsymbol{J_f} = 0.$$

Using Ohm's law, $\boldsymbol{J_f} = \sigma \boldsymbol{E}$ we have

$$\nabla \cdot \boldsymbol{J_f} = \sigma \nabla \cdot \boldsymbol{E} = \frac{\sigma}{\varepsilon} \nabla \cdot \boldsymbol{D} = \frac{\sigma}{\varepsilon} \rho_f$$

so the continuity equation becomes

$$\partial_t \rho_f = -\frac{\sigma}{\varepsilon} \rho_f.$$

This has as a solution

$$\rho_f(t) = \rho_f(0)e^{-\sigma t/\varepsilon}.$$

So ε/σ is the characteristic timescale for how fast charge decays in a conductor. This is known as the **relaxation time**, τ. This quantity tells us how fast charges migrate to the surface and how fast the electric field disappears in a conductor. In an ideal conductor $\sigma \to \infty$ and $\tau \to 0$ which corresponds to our assumption for an ideal conductor that charge is always concentrated on the surface and $\boldsymbol{E} = \boldsymbol{0}$ inside an ideal conductor.

The interpretation of $1/\omega$ is, as one would expect, the period of the wave, $T = 2\pi/\omega$. Thus $\sigma/\varepsilon\omega = T/2\pi\tau$. We use this ratio to characterise conductors as "good" and "bad". For a "good" conductor $\sigma/\varepsilon\omega \gg 1$. For a "bad" conductor $\sigma/\varepsilon\omega \ll 1$. Notice that how "good" a conductor is depends on the frequencies that we care about transmitting. For example it is possible that a conductor can be a good conductor of radio waves, which have a relatively small value of ω, and a bad conductor of ultraviolet, which has a large value of ω.

It can be shown that for a good conductor

$$\delta \approx \sqrt{\frac{2}{\mu\omega\sigma}},$$

and for a bad conductor

$$\delta \approx \sqrt{\frac{4\varepsilon}{\mu\sigma^2}}.$$

Typical metals are good conductors up to about 1 MHz with $\delta \approx 1\,\text{cm}$ at 50 Hz (mains power frequency) and $\delta \approx 10\,\mu\text{m}$ at 50 MHz.

Some consequences of the skin depth are

- Cables greater than about 1 cm in width are wasted as the current is mostly in the skin layer around the outside and there is a 'dead zone' in the centre. Cables that initially seem thicker than 1 cm are often actually multiple narrower cables bound together.
- Submarines cannot use radio as at the typical depth of a submarine radio waves simply cannot penetrate the water.
- Mobile phones don't work inside metal boxes as they use radio waves which have frequencies in the gigahertz range and therefore don't penetrate very far through metal.
- Microwave oven doors have a metal mesh with holes much smaller than the wavelength of the microwaves. This allows visible light through but not microwaves.

19.2 Phase Relations of Fields

As with waves in a vacuum or insulator Maxwell's first and second laws imply that

$$i\tilde{\boldsymbol{k}} \cdot \tilde{\boldsymbol{E}}_{\boldsymbol{0}} = 0, \qquad \text{and} \qquad i\tilde{\boldsymbol{k}} \cdot \tilde{\boldsymbol{B}}_{\boldsymbol{0}} = 0$$

so the waves are transverse. In the case that $\tilde{\boldsymbol{k}} = \tilde{k}\boldsymbol{e_z}$ and $\boldsymbol{e}_{\tilde{\boldsymbol{E}}_{\boldsymbol{0}}} = \tilde{E}_0\boldsymbol{e_x}$ if we substitute into Maxwell's third equation we get

$$i\tilde{\boldsymbol{k}} \times \tilde{\boldsymbol{E}}_{\boldsymbol{0}} = i\omega\tilde{\boldsymbol{B}}_{\boldsymbol{0}}$$

hence

$$\tilde{\boldsymbol{B}}_{\boldsymbol{0}} = \frac{\tilde{k}\tilde{E}_0}{\omega}\boldsymbol{e_y}. \tag{19.1}$$

In general $\tilde{k}$ is complex and therefore so are $\tilde{E}_0$ and $\tilde{B}_0$. We write

$$\tilde{k} = Re^{i\varphi}, \qquad \text{where} \qquad R = \sqrt{k^2 + \kappa^2}, \qquad \text{and} \qquad \varphi = \arctan\left(\frac{\kappa}{k}\right).$$

Then

$$\tilde{E}_0 = E_0 e^{i\delta_E}, \qquad \text{and} \qquad \tilde{B}_0 = B_0 e^{i\delta_B}.$$

Substituting these into equation 19.1 we get

$$B_0 e^{i\delta_B} = \frac{Re^{i\varphi}}{\omega} E_0 e^{i\delta_E} \implies \delta_B - \delta_E = \varphi.$$

What this means is that the magnetic field has a phase of φ behind the electric field since when we take the real part to get the physical fields we get

$$\boldsymbol{E} = E_0 e^{-\kappa z}\cos(kz - \omega t + \delta_E)\boldsymbol{e_x},$$
$$\boldsymbol{B} = B_0 e^{-\kappa z}\cos(kz - \omega t + \delta_E + \varphi)\boldsymbol{e_y}.$$

In terms of physical constants we have

$$R = \sqrt{k^2 + \kappa^2} = \omega\sqrt{\mu\varepsilon}\left[1 + \left(\frac{\sigma}{\varepsilon\omega}\right)^2\right]^{1/4},$$

and

$$\varphi = \arctan\left(\frac{\kappa}{k}\right) = \arctan\left(\left[\frac{\sqrt{1 + (\sigma/\varepsilon\omega)^2} - 1}{\sqrt{1 + (\sigma/\varepsilon\omega)^2} + 1}\right]^{1/2}\right).$$

For a good conductor

$$\varphi \to \arctan(1) = \frac{\pi}{4}$$

and

$$\tilde{k} \approx e^{i\pi/4}\sqrt{\mu\omega\sigma}.$$

19.3 Intrinsic Impedance

We define the **intrinsic impedence**, or **wave impedence** as the ratio

$$Z = \frac{\tilde{E}_0}{\tilde{H}_0}.$$

This is a property of the medium. It has dimensions of ohms. It can be thought of as a generalised resistance. In a vacuum

$$\frac{E_0}{H_0} = \frac{E_0\mu_0}{B_0} = c\mu_0 = Z_{\text{vac}} = 377\,\Omega.$$

This is the **vacuum impedance**. It is real as $\boldsymbol{E}$ and $\boldsymbol{H}$ are in phase in a vacuum.

In a linear dielectric

$$Z = \frac{E_0}{H_0} = \frac{E_0\mu}{B_0} = \sqrt{\frac{\mu_r}{\varepsilon_r}} Z_{\text{vac}}.$$

In a good conductor $\tilde{k} \approx e^{i\pi/4}\sqrt{\mu\omega\sigma}$ and so

$$Z = \frac{\tilde{E}_0}{\tilde{H}_0} = \frac{\tilde{E}_0\mu}{\tilde{B}_0} = \frac{\omega\mu}{\tilde{k}} \approx e^{-i\pi/4}\sqrt{\frac{\mu\omega}{\sigma}}.$$

This is complex as $\boldsymbol{E}$ and $\boldsymbol{H}$ are out of phase.

20 Waves at Interfaces

20.1 Summary of Plane Waves and Interfaces

A plane polarised wave propagating in the $\boldsymbol{e_z}$ direction will have

$$\boldsymbol{E} = \boldsymbol{E_0} e^{i(kz-\omega t)}, \qquad \text{and} \qquad \boldsymbol{B} = \boldsymbol{B_0} e^{i(kz-\omega t)}.$$

We have seen that combining this with Maxwell's third law gives us $ik\boldsymbol{e_z} \times \boldsymbol{E_0} = i\omega\boldsymbol{B_0}$. We are often free to choose $\boldsymbol{E_0} = E_0\boldsymbol{e_x}$ meaning that the waves are plane polarised in the $\boldsymbol{e_x}$ direction. Thus

$$\boldsymbol{B_0} = \frac{kE_0}{\omega}\boldsymbol{e_y}.$$

In general $E_0, B_0 \in \mathbb{C}$. Previously we drew attention to this with a tilde but we will drop that from here on referring to E_0 (previously $\tilde{E_0}$) as the complex amplitude and the (real) amplitude as $|E_0|$ (previously E_0).

Recall that the complex impedance of the medium is defined as

$$Z = \frac{E_0}{H_0} = \frac{\mu E_0}{B_0}$$

assuming a linear medium. A complex value of Z corresponds to a phase shift between $\boldsymbol{E}$ and $\boldsymbol{H}$.

Consider a plane polarised wave, $\boldsymbol{E_{\text{inc}}}$, propagating in the $\boldsymbol{e_z}$ direction and crossing between media at $z = 0$ across a boundary that is orthogonal to the wave. Suppose the wave starts in a medium with impedance Z_1 and ends in a medium with impedance Z_2. We take $\boldsymbol{e_x}$ to be along $\boldsymbol{E_{\text{inc}}}$ and $\boldsymbol{e_y}$ to be along $\boldsymbol{H_{\text{inc}}}$ and $\boldsymbol{e_z}$ to be along $\boldsymbol{e_{k_1}}$. We expect that there will be three important waves, the incoming wave, the reflected part of the wave, and the transmitted part of the wave.

20.2 Interfaces Between Two Dielectric Media

In a linear dielectric media

$$Z_i = v_i \mu_i$$

is real and there is no phase lag between $\boldsymbol{E}$ and $\boldsymbol{H}$. We take the amplitude E_I to be real and we write

$$\boldsymbol{E_{\text{inc}}} = E_I \boldsymbol{e_x} e^{i(k_1 z - \omega t)},$$

and

$$\boldsymbol{H_{\text{inc}}} = \frac{E_I}{\mu_1 v_1} \boldsymbol{e_y} e^{i(k_1 z - \omega t)}.$$

For the transmitted wave the propagation is still in the $\boldsymbol{e_z}$ direction and now

$$\boldsymbol{E_{\text{trans}}} = E_T \boldsymbol{e_x} e^{i(k_2 z - \omega t)},$$

and

$$\boldsymbol{H_{\text{trans}}} = \frac{E_T}{\mu_2 v_2} \boldsymbol{e_y} e^{i(k_2 z - \omega t)}.$$

The reflected wave propagates in the $-\boldsymbol{e_z}$ direction so

$$\boldsymbol{E_{\text{ref}}} = E_R \boldsymbol{e_x} e^{i(-k_1 z - \omega_t)},$$

and

$$\boldsymbol{H_{\text{ref}}} = -\frac{E_R}{\mu_1 v_1} \boldsymbol{e_y} e^{i(-k_1 z - \omega t)}.$$

Note that $\boldsymbol{H_{\text{ref}}}$ has a minus sign to ensure that the three components, $\boldsymbol{E_{\text{ref}}}$, $\boldsymbol{H_{\text{ref}}}$, and $-\boldsymbol{e_z}$, form a right handed system.

We now apply continuity conditions. Both $\boldsymbol{e_x}$ and $\boldsymbol{e_y}$ are tangential to the boundary and therefore $\boldsymbol{E}$ and $\boldsymbol{H}$ are continuous at the boundary, since we are assuming no surface charges/currents. From the assumption that $E_t = E_x$ is continuous we get

$$E_I + E_R = E_T.$$

From the assumption that $H_t = H_y$ is continuous we get

$$\frac{E_I - E_R}{\mu_1 v_1} = \frac{E_T}{\mu_2 v_2}.$$

For a give E_I we can solve for E_T and E_R from which we define the **amplitude transmission coefficient**,

$$t = \frac{E_T}{E_I} = \frac{2}{1 + \beta},$$

where

$$\beta = \frac{\mu_1 v_1}{\mu_2 v_2} = \frac{Z_1}{Z_2},$$

and the **amplitude reflection coefficient**,

$$r = \frac{E_R}{E_I} = \frac{1-\beta}{1+\beta}.$$

If the media is non-magnetic, i.e. $\mu_i = \mu_0$, which is often a good approximation, then $\beta = v_1/v_2 = n_2/n_1$ where we have used $v_i = 1/\sqrt{\mu_i \varepsilon_i} = c/n_i$. This allows us to write

$$t = \frac{2v_2}{v_1 + v_2} = \frac{2n_1}{n_1 n_2}$$

and

$$r = \frac{v_2 - v_1}{v_1 + v_2} = \frac{n_1 - n_2}{n_1 + n_2}.$$

A good sanity check here is if $Z_1 = Z_2$ then $t = 1$ and $r = 0$ so the wave is 100 % transmitted, which is what we would expect since $Z_1 = Z_2$ means that there isn't really a boundary so there is no reason for the wave to be reflected.

Notice that t can be greater than 1 and r can be negative. It is also possible that $t > 1$ and $r > 0$, this seems to be generating energy as the amplitude of both the transmitted and reflected waves is greater than the amplitude of the incoming wave. To show that this doesn't violate energy conservation we need to more carefully consider the energy.

20.2.1 Energy Flow Across a Boundary

The Poynting vector is

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} = \frac{1}{\mu} \boldsymbol{E} \times \boldsymbol{B}.$$

This gives the energy flux across the boundary. The energy flux per unit volume, averaged over one period, is what we define as the **intensity** of the wave. It is given by

$$|\langle \boldsymbol{S} \rangle| = \frac{1}{\mu} |\langle \boldsymbol{E} \times \boldsymbol{B} \rangle| = \frac{1}{2\mu v} E_0^2 = \frac{\varepsilon v}{2} E_0^2.$$

Note that this is proportional to the square of the amplitude. We define the ratio of reflected to incident intensity, R, and the ratio of transmitted to incident intensity, T, as

$$R = \frac{|\langle \boldsymbol{S_R} \rangle|}{|\langle \boldsymbol{S_I} \rangle|} = \frac{E_R^2}{E_I^2} = r^2 = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2$$

and

$$T = \frac{|\langle \boldsymbol{S_T} \rangle|}{|\langle \boldsymbol{S_I} \rangle|} = \frac{\varepsilon_2 v_2 E_T^2}{\varepsilon_1 v_1 E_I^2} = \frac{\varepsilon_2 v_2}{\varepsilon_1 v_1} t^2 = \frac{4n_1 n)2}{(n_1 + n_2)^2}.$$

The fact that

$$R + T = 1$$

implies that energy is conserved. The paradox that arose before was because we considered the amplitude, not the square of the amplitude, as a measure of energy.

20.3 General Media

We repeat the above calculations but now allowing for a complex impedance, Z_i, which may lead to a phase lag between $\boldsymbol{E}$ and $\boldsymbol{H}$. Now

$$\begin{aligned}
\boldsymbol{E}_{\mathbf{inc}} &= E_I \boldsymbol{e_x} e^{i(k_1 z - \omega t)} \\
\boldsymbol{H}_{\mathbf{inc}} &= \frac{E_I}{Z_1} \boldsymbol{e_y} e^{i(k_1 z - \omega t)} \\
\boldsymbol{E}_{\mathbf{trans}} &= E_T \boldsymbol{e_x} e^{i(k_2 z - \omega t)} \\
\boldsymbol{H}_{\mathbf{trans}} &= \frac{E_T}{Z_2} \boldsymbol{e_y} e^{i(k_2 z - \omega t)} \\
\boldsymbol{E}_{\mathbf{ref}} &= E_R \boldsymbol{e_x} e^{i(-k_1 z - \omega t)}
\end{aligned}$$

$$\boldsymbol{H}_{\mathbf{ref}} = -\frac{E_R}{Z_1}\boldsymbol{e_y}e^{i(-k_1 z-\omega t)}$$

Again assuming no surface currents/charges we have that $E_t = E_x$ and $H_t = H_y$ are continuous so

$$E_I + E_R = E_T$$

and

$$\frac{E_I - E_R}{Z_1} = \frac{E_T}{Z_2}$$

solving for a given E_I we have

$$t = \frac{E_T}{E_I} = \frac{2Z_2}{Z_2 + Z_1}$$

and

$$r = \frac{E_R}{E_I} = \frac{Z_2 - Z_1}{Z_2 + Z_1}.$$

Note that in general these will be complex.

20.3.1 Energy Flow Across a Boundary

We need to be slightly more careful with the Poynting vector when we have complex impedances involved. We work with the time averaged Poynting vector,

$$\langle \boldsymbol{S} \rangle = \hat{\boldsymbol{k}}\frac{1}{2}\,\mathrm{Re}\left[\frac{1}{Z}\right]|E_0|^2.$$

The intensity of the wave is then

$$|\langle \boldsymbol{S} \rangle| = \frac{1}{2}\,\mathrm{Re}\left[\frac{1}{Z}\right]|E_0|^2.$$

20.4 Reflection at Conducting Surfaces, or Why are Metals Shiny?

Consider the same set up as before but now let the first medium be a vacuum, meaning $Z_1 = Z_{\text{vac}} = 377\,\Omega$, and let the second medium be a conductor meaning

$$Z_2 = e^{-i\pi/4}\sqrt{\frac{\mu\omega}{\sigma}} = \frac{1-i}{\sigma\delta},$$

where

$$\delta = \sqrt{\frac{2}{\mu\sigma\omega}}$$

is the skin depth. In general Z_2 is complex and ω dependant. However for a conductor the magnitude of Z_2 is very small. For example for copper at $\omega = 10\,\text{GHz}$ $|Z_2| = 0.036\,\Omega = 10^{-4}Z_{\text{vac}}$. At 100 THz, visible light frequencies, $|Z_2| = 3.6\,\Omega = 10^{-2}Z_{\text{vac}}$. The amplitude reflection is then

$$r = \frac{Z_2 - Z_1}{Z_2 + Z_1} = -0.98 \approx -1.$$

This corresponds to almost 100 % reflection with a phase reversal. The physical origin of shininess is skin effect. The transmitted wave decays as $e^{-z/\delta}$ and almost all energy that is put in comes back out as δ is so small. Notice that this is still ω dependent, for example gamma radiation, with a very high frequency, can penetrate far into a metal and is therefore not reflected as much.

Part IV

Electromagnetic Waves

21 Waves Recap

21.1 Waves in One Dimension

The archetypal one dimensional wave is a wave travelling along a string. We can characterise this wave by the displacement of the string from its start position. We will call this quantity ψ and in general it is a function of position and time. If the wave travels without changing shape then the wave can be written as

$$\psi(x,t) = f(x - vt)$$

for some arbitrary (twice differentiable) function f. This corresponds to a wave travelling in the $+x$ direction. A wave travelling in the $-x$ direction will have the form

$$\psi(x,t) = g(x + vt).$$

This all assumes that the media (in this case the string) is 'transparent' i.e. it doesn't absorb any energy, isotropic i.e. it doesn't favour any particular direction, and homogenous i.e. it is the same everywhere.

It is easy to show that ψ satisfies

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2}.$$

This is the wave equation in one dimension. In the future we will define a wave as something that satisfies this equation.

21.1.1 Pulse Wave

The simplest case we might consider is a single pulse that travels in the $+x$ direction at some speed. This is shown in figure 21.1 which shows a wave propagating in the $+x$ direction at a speed of 2. The

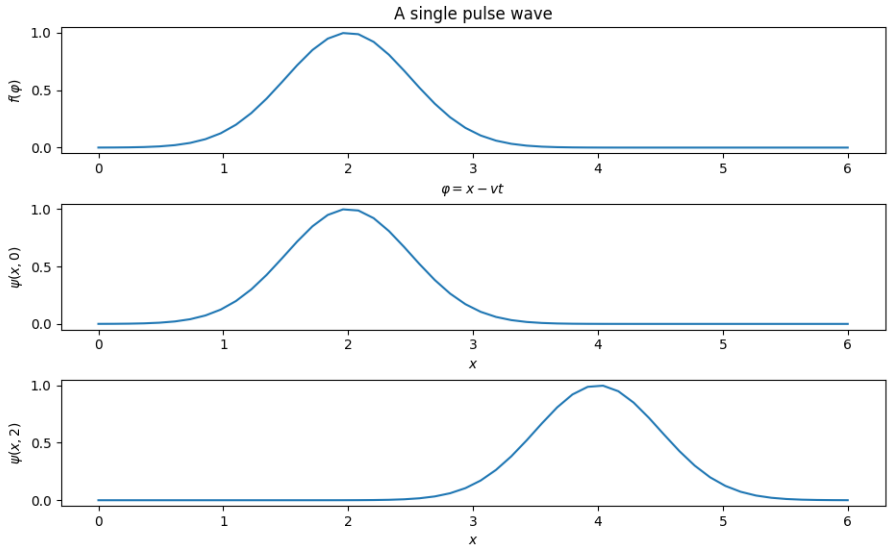


Figure 21.1: A function f, the wave $\psi(x,t) = f(x - vt)$ at times $t = 0, 2$ for $v = 1$.

peak of f in this figure occurs at $\varphi = x - vt = 2$. At $t = 0$ we have $\psi(x,0) = f(x - 0)$ which has a peak at $x = 2$. At $t = 0$ we have $\psi(x,2) = f(x - 1 \cdot 2)$ which still has a peak at $\varphi = 2$ which means the peak is at $x = 4$. We can also view this as a transformation of the original wave as $x \to x - 2$ represents a translation $+2$ units in the x direction.

21.1.2 Harmonic Wave

Often we think of waves as being periodic and repeating regularly. Of all the waves of this type the most important are the harmonic waves which are described mathematically by sines and cosines. We will see later that through Fourier series we can actually use a superposition of harmonic waves to describe any wave.

The most simple harmonic wave is

$$\psi(x,t) = A\sin(k(x-vt)-\Phi)$$

where A is the amplitude, v is the velocity, and Φ is some constant phase shift. The phase of this wave is $\varphi = k(x-vt)-\Phi$. If we consider ψ at some fixed point as time varies we see it is periodic in time with period τ. We know that sin is periodic with period 2π so we must have that at $x = x_0$

$$k(x_0 - vt_0) - \Phi - k(x_0 - v(t_0+\tau)) + \Phi = 2\pi \implies kv = \frac{2\pi}{\tau} = \omega.$$

ω is the temporal angular frequency, or frequency for short. Similarly if we consider the wave at $t = 0$ then we see that it is periodic in space with period λ, known as the wavelength. Again sin is periodic with period 2π so

$$k\lambda = 2\pi \implies k = \frac{2\pi}{\lambda}.$$

k is the spatial angular frequency, or wave number for short.

It is common to write waves in a way which explicitly gives k and ω, such as

$$\psi(x,t) = A\sin(kx - kvt - \Phi) = A\sin(kx - \omega t - \Phi).$$

This is beneficial as it puts space and time on an equal footing. We can easily recover $v = \omega/k$. The relationship between spatial and temporal parts is called a **dispersion relation**.

Another notational convenience is to use complex exponentials, such as

$$\psi(x,t) = Ae^{i(kx-\omega t-\Phi)} = A\cos(kx-\omega t-\Phi) + i\sin(kx-\omega t-\Phi).$$

The complex numbers simplify a lot of calculations but it is the real part of ψ that corresponds to the physical disturbance.

21.1.3 Phase Velocity

Strictly speaking the velocity of the wave that we have spoken of so far is the **phase velocity**. It is the velocity of a peak of the wave. For example the peak of our sinusoidal wave occurs at phase $\varphi = \pi/2$. If there is a crest at (x_1, t_1) then we must have

$$kx_1 - \omega t_1 - \Phi = \frac{\pi}{2}.$$

At time $t_1 + \delta t$ the crest must then be at $x_1 + \delta x$ which satisfies

$$k(x_1+\delta x) - \omega(t_1+\delta t) - \Phi = \frac{\pi}{2}.$$

Subtracting the first of these from the second we get

$$k\delta x - \omega\delta t = 0 \implies \frac{\delta x}{\delta t} = \frac{\omega}{k}.$$

It is this velocity that we refer to when we say the phase velocity. Note that we don't have to use a crest here, we could just as well have used a trough or a point of zero displacement or any other condition of constant phase.

21.2 Waves in Three Dimensions

The wave equation generalises to three dimensions as

$$\nabla^2\psi = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2},$$

and the condition $\psi(x,t) = f(kx - \omega t)$ becomes

$$\psi(\boldsymbol{r},t) = f(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t),$$

where $\boldsymbol{k}$ is the **wave vector**.

21.2.1 Plane Waves

The quantity $\boldsymbol{k}\cdot\boldsymbol{r}$ is the displacement in the direction $\boldsymbol{k}$. There is a whole plane of points, $\boldsymbol{r}$, which give the same value for $\boldsymbol{k}\cdot\boldsymbol{r}$. These planes which satisfy $\boldsymbol{k}\cdot\boldsymbol{r} = \text{const}$ are called wavefronts. These planes have the same phase everywhere on the plane at any given time. As well as the shape of the wavefronts we also have to consider the shape of the wave. one of the most common plane waves is

$$\psi(\boldsymbol{r},t) = Ae^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t-\Phi)}.$$

Three-dimensional waves are hard to draw as we really need four dimensions to draw them, one for each spatial direction and one for the amplitude, even if we fix t as some constant value. Instead we will often consider a two-dimensional wave and use the third dimension for plotting the amplitude. For example see figure 21.2.

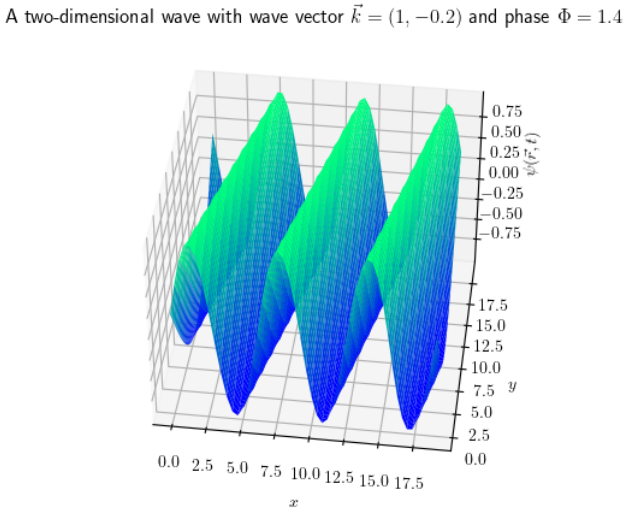


Figure 21.2: A plane wave in two dimensions with $\boldsymbol{k} = (1, -0.2)$ and $\Phi = 1.4$. The wave is $\psi(\boldsymbol{r},t) = \exp(i[x - 0.2y - 1.4])$.

21.2.2 Spherical Waves

Another common case of three-dimensional waves is something radiating from a point. In this case the wavefronts are spherical and we are best to work with spherical coordinates. The Laplacian of a central function in spherical coordinates is

$$\nabla^2\psi(r) = \frac{1}{r}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) = \frac{\partial^2\psi}{\partial r^2} + \frac{2}{r}\frac{\partial\psi}{\partial r} = \frac{1}{r}\frac{\partial^2}{\partial r^2}(r\psi).$$

Using the last version we find that $r\psi$ satisfies the one-dimensional wave equation and therefore

$$\psi(r,t) = \frac{1}{r}f(r - vt).$$

For example a harmonic spherical wave might be given by

$$\psi(r,t) = \frac{A}{r}\sin(k(r-vt)). \tag{21.1}$$

The factor of $1/r$ is important. It ensures that the intensity of the wave decreases as the wave spreads out. Far from the origin a spherical wave will approximate a plane wave. This can be seen in figure 21.3

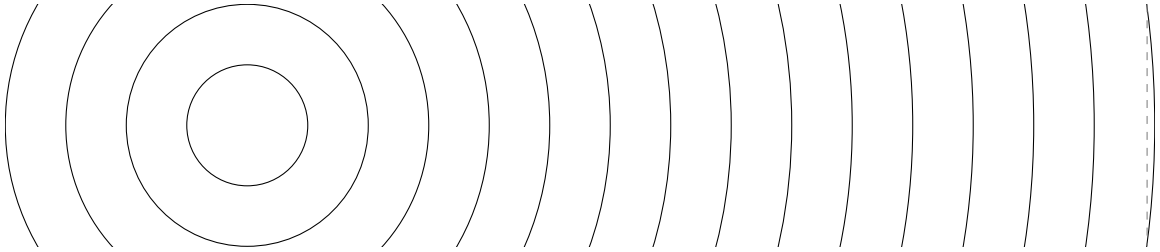


Figure 21.3: A spherical wave at large distances approximates a plane wave. Notice how at the far right the wavefront is almost parallel to the straight dashed line.

22 Electromagnetic Waves

22.1 Energy Density and Optical Intensity

Recall that the Poynting vector is defined as

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

and that the energy density of an electromagnetic wave in a vacuum is

$$u = \frac{1}{2}\varepsilon_0 E^2 + \frac{1}{2\mu_0}B^2\varepsilon_0 E^2.$$

For a harmonic wave polarised in the x direction (that is $\boldsymbol{E}\propto\boldsymbol{e_x}$ so $\boldsymbol{B}\propto\boldsymbol{e_y}$ and $\boldsymbol{k}\propto\boldsymbol{e_z}$) the energy density is

$$u = \varepsilon_0 E^2 = \varepsilon_0 E_0^2\cos^2(kz-\omega t-\Phi)$$

and the Poynting vector is

$$\boldsymbol{S} = c\varepsilon_0 E_0^2\cos^2(kz-\omega 6-\Phi)\boldsymbol{e_z} = cu\boldsymbol{e_z}.$$

For an optical wave $\omega\approx 10^{15}\,\mathrm{s}^{-1}$. It is therefore not possible to measure an instantaneous value of $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{S}$, or u. Instead we can measure the rate of energy transferred over a long period. This corresponds to averaging over many periods. We define the **intensity** as

$$I = \langle S\rangle = c\varepsilon_0\langle E\rangle$$

where $\langle f\rangle$ denotes an average of f over many periods which can be computed as

$$\langle f\rangle = \frac{1}{t_{\max}}\int_0^{t_{\max}} f(t)\,\mathrm{d}t$$

where $t_{\max}\gg T$ where $T = 2\pi/\omega$ is the period of f.

22.2 Violation of Newton's Third Law?

Consider the two charges shown in figure 22.1. It shows two charges of charge q. One is moving horizontally along the x-axis with velocity $\boldsymbol{v_1}$ and the other is moving in the $-z$ direction with velocity $\boldsymbol{v_2}$. Suppose that $v_1 = v_2 = v$. The force due to the electric field from charge 1 on charge 2 is

$$\boldsymbol{F_{E2}} = \frac{q^2}{4\pi\varepsilon_0 d^2}\boldsymbol{e_x}.$$

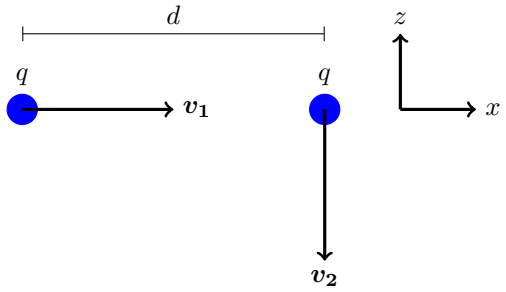


Figure 22.1: Two moving charges.

Similarly the force due to the electric field from charge 2 on charge 1 is

$$\boldsymbol{F_{E1}} = -\frac{q^2}{4\pi\varepsilon_0 d^2}\boldsymbol{e_x}.$$

These forces are equal and opposite as Newton's third law would predict.

The force due to the magnetic field from charge 1 on charge 2 is

$$\boldsymbol{F_{B2}} = q\boldsymbol{v_2} \times \boldsymbol{B_1} = q\boldsymbol{v_2} \times \left[\frac{q}{4\pi\varepsilon_0 c^2}\frac{\boldsymbol{v_1} \times \boldsymbol{r_2}}{r_2^3}\right] = \boldsymbol{0}.$$

This is $\boldsymbol{0}$ since $\boldsymbol{r_2}$, the vector from charge 1 to charge 2, is parallel to $\boldsymbol{v_1}$. The force due to the magnetic field form charge 2 on charge 1 is

$$\boldsymbol{F_{B1}} = q\boldsymbol{v_1} \times \boldsymbol{B_2} = q\boldsymbol{v_1} \times \left[\frac{q}{4\pi\varepsilon_0 c^2}\frac{\boldsymbol{v_2} \times \boldsymbol{r_1}}{r_1^3}\right] \neq \boldsymbol{0}.$$

So the forces due to the magnetic fields are *not* equal and opposite. This seems to violate Newton's third law. In the next system we will see why this is.

22.3 Radiation Pressure

Light incident on a positive charge, q, at some moment when $\boldsymbol{E}$ is in the positive x direction will initiate movement of the charge in the positive x direction doing work at a rate

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \boldsymbol{v} \cdot \boldsymbol{F_E} = v_x q E_x.$$

At this point the charge is moving and therefore experiences a force due to the magnetic component of the light. At the same moment this component is in the positive y direction and the force is

$$\boldsymbol{F_B} = q\boldsymbol{v} \times \boldsymbol{B} = qv_x\boldsymbol{e_x} \times B_y\boldsymbol{e_y} = qv_xB_y\boldsymbol{e_z}.$$

At a later time when $\boldsymbol{E}$ is in the negative x direction the rate of work that $\boldsymbol{E}$ does will be

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \boldsymbol{v} \cdot \boldsymbol{F_E} = -v_x q E_x.$$

At the same moment $\boldsymbol{B}$ will be in the negative y direction and the charge will be moving now in the negative x direction so the force due to the magnetic field is

$$\boldsymbol{F_B} = q\boldsymbol{v} \times \boldsymbol{B} = q(-v_x\boldsymbol{e_x}) \times (-B_y\boldsymbol{e_y}) = qv_xB_y\boldsymbol{e_z}.$$

Notice that one period the average force due to $\boldsymbol{E}$ is

$$\langle \boldsymbol{F_E} \rangle = q\langle \boldsymbol{E} \rangle = \boldsymbol{0}.$$

However the average net force is

$$\langle \boldsymbol{F} \rangle = \langle \boldsymbol{F_E} + \boldsymbol{F_B} \rangle = \langle \boldsymbol{F_B} \rangle = q\langle v_xB_y \rangle \boldsymbol{e_z} = \frac{q}{c}\langle v_xE_x \rangle \boldsymbol{e_z},$$

using the fact that for an electromagnetic wave $B_y = E_x/c$. From this we see that there is a net force which does work at a rate

$$\left\langle \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle = \langle \boldsymbol{v} \cdot \boldsymbol{F} \rangle = \langle \boldsymbol{v} \cdot (\boldsymbol{F_E} \cdot \boldsymbol{F_B} \rangle = \langle \boldsymbol{v} \cdot \boldsymbol{F_E} \rangle = q \langle v_x E_x \rangle,$$

where we have used the fact that $\boldsymbol{v} \cdot \boldsymbol{F_B} = \boldsymbol{v} \cdot (\boldsymbol{v} \times \boldsymbol{B}) = 0$. We see that there is a net gain in energy while the light is incident on the charge. Notice that

$$\left\langle \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle \boldsymbol{e_z} = c \langle \boldsymbol{F} \rangle = \left\langle \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} \right\rangle.$$

So there is a net gain in momentum. This gain of momentum is inversely proportional to c so is often too small to notice. The pressure due to this force is called **radiation pressure**. For example on Earth the energy density from solar radiation is approximately $1\,\mathrm{kW\,m^{-}2}$. This is an appreciable amount of energy, the sun feels warm. However the radiation pressure is then on the order of $10^{-6}\,\mathrm{Pa}$ which is 11 orders of magnitude smaller than atmospheric pressure.

While this effect is small it is still important. For example there have been a few space probes that use the radiation pressure of the sun to accelerate. To do this they use large sails to maximise the are over which sunlight is incident. Even traditionally driven probes have to account for the radiation pressure changing their course slightly.

This effect also explains the seeming violation of Newton's third law earlier. If we properly account for the radiation pressure then we will find that Newton's third law is not violated.

We saw with radiation pressure that the momentum is equal to the energy divided by c. If we consider for a moment quantum mechanics then we have

$$p = \frac{h}{\lambda} \qquad \text{and} E = hf = \frac{hc}{\lambda} = pc \implies p = \frac{E}{c}.$$

If instead we consider special relativity then we have

$$E^2 = m^2c^4 + p^2c^2$$

if $m = 0$ then we have

$$E = pc \implies p = \frac{E}{c}.$$

So we have agreement between electromagnetism, quantum mechanics, and special relativity as to $p = E/c$.

One experimental verification of radiation pressure comes from Compton scattering. This is a process by which an incoming photon of wavelength λ_i hits an electron which deflects away at angle φ and the photon deflects at angle ϑ. The wavelength of the scattered photon is λ_f. It can be shown that

$$\lambda_f - \lambda_i = \Delta\lambda = \frac{h}{m_e c}(1 - \cos\vartheta).$$

Since some momentum must be transferred to the electron the momentum of the photon must decrease and therefore the wavelength changes.

23 Dipole Radiation

23.1 Light From Maxwell's Equations

Recall that we can combine Maxwell's equations to get the wave equations

$$\nabla^2 \boldsymbol{E} = \mu_0 \varepsilon_0 \partial_t^2 \boldsymbol{E}, \qquad \text{and} \qquad \nabla^2 \boldsymbol{B} = \mu_0 \varepsilon_0 \partial_t \boldsymbol{B}.$$

We see that each component of the electric and magnetic fields must satisfy the three-dimensional wave equation for a wave travelling at speed $(\mu_0\varepsilon_0)^{-1/2}$, which is the speed of light. This was one of the facts that originally convinced Maxwell that light could be explained as an electromagnetic phenomenon.

Suppose we have an electromagnetic wave which propagates such that

$$E_i(\boldsymbol{r}, t) = f(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t).$$

These are plane waves propagating along $\boldsymbol{k}$. Now we choose $\boldsymbol{e_z}$ to be the same direction as $\boldsymbol{k}$ so that $E_i(\boldsymbol{r}, t) = f(kz - \omega t)$. We see that

$$\nabla \cdot \boldsymbol{E} = \partial_z E_z$$

if we further assume that the wave is propagating in free space then we have $\nabla \cdot \boldsymbol{E} = 0 = \partial_z E_z$. This means that $\boldsymbol{E}$ doesn't oscillate in the direction of propagation meaning that it is a transverse wave. Suppose at some point $\boldsymbol{E}$ is aligned with $\boldsymbol{e_x}$. Then We must have $\boldsymbol{B}$ parallel to $\boldsymbol{e_y}$ as we know that it is perpendicular to $\boldsymbol{E}$.

Suppose

$$\boldsymbol{E} = E_x(\boldsymbol{r}, t)\boldsymbol{e_x} = E_{x0}\cos(kz - \omega t - \Phi)\boldsymbol{e_x}.$$

That is $\boldsymbol{E}$ is a harmonic wave polarised in the x direction. We then find from Faraday's law that

$$-\partial_t B_y = \partial_z E_x = -E_{x0}k\sin(kz - \omega t - \Phi).$$

From this we have

$$B_y = E_{x0}k\int \sin(kz - \omega t - \Phi)\,\mathrm{d}t = E_{x0}\frac{k}{\omega}\cos(kz - \omega t - \Phi) = \frac{1}{c}E_{x0}\cos(kz - \omega t - \Phi) = \frac{E_x}{c}.$$

So the $\boldsymbol{B}$ oscillates perpendicular to $\boldsymbol{E}$ and has a magnitude $B = E/c$.

23.2 Dipole Radiation

Suppose we have an atom at rest. If we displace the electron cloud slightly we will have an effective dipole. This dipole turns out to be critical to how light interacts with matter. Suppose that we distort an atom to create a dipole moment $\boldsymbol{p_0}$. Approximating this as an ideal dipole the field we expect is

$$E_r = \frac{2p_0\cos\vartheta}{4\pi\varepsilon_0 r^3}, \qquad \text{and} \qquad E_\vartheta = \frac{p_0\sin\vartheta}{4\pi\varepsilon_0 r^3}.$$

Due to rotational symmetry about $\boldsymbol{p_0}$ we expect that there will be no φ dependence of the field.

If we slightly distort the dipole and then allow it to change freely then to second order we expect harmonic oscillation so we expect the dipole to vary as

$$\boldsymbol{p}(t) = \boldsymbol{p_0}\cos(\omega t).$$

However we cannot use this new dipole in the equations for a dipole field. The problem is that these equations assume the dipole is static, which is no longer the case. The reason this assumption is necessary is the finite speed of light. While the dipole oscillates it takes time for the change in the field to propagate and by the time it has the dipole is different again. There is a time lag in the response to the oscillation and the lag increases with distance.

23.2.1 Retarded Time

The solution to the dipole field of an oscillating dipole uses a concept called **retarded time**. The retarded time at a distance x from an accelerating charge is $t' = t - x/c$ where t is the actual time. We use square brackets, $[\cdot]$, to denote a quantity that is to be calculated at the retarded time. For example if $a = a(t)$ is the acceleration of a charge then $[a] = a(t')$ is the acceleration of the charge at the retarded time. If we have a single charge accelerating along the z-axis then it can be shown that the electric field at some distance, x, along the x-axis at time t is

$$\boldsymbol{E}(x, y = 0, z = 0, t) = \boldsymbol{e_x}\frac{q}{4\pi\varepsilon_0 x^2} - \boldsymbol{e_z}\frac{q[a]}{4\pi\varepsilon_0 x c^2}.$$

The first term is the normal Coulomb term due to the presence of the charge. The second term, called the retarded term, is due to the accelerating charge causing a changing magnetic field which in turn causes an electric field. We see that in the limit $c \to \infty$ the retarded term disappears which makes sense since the retarded term only appears due to the fact that c is *not* infinite. In the limit $x \to \infty$ the retarded term dominates.

23.2.2 Full Dipole Radiation Equations

It can be shown that the electric and magnetic fields for a time dependent dipole, with magnitude $p(t)$ oriented along the z-axis, are given by

$$E_r = \frac{2}{4\pi\varepsilon_0}\left(\frac{[p]}{r^3} + \frac{[\mathrm{d}p/\mathrm{d}t]}{cr^2}\right),$$
$$E_\vartheta = \frac{1}{4\pi\varepsilon_0}\left(\frac{[p]}{r^3} + \frac{[\mathrm{d}p/\mathrm{d}t]}{cr^2} + \frac{[\mathrm{d}^2p/\mathrm{d}t^2]}{c^2r}\right)\sin\vartheta,$$
$$B_\varphi = \frac{1}{4\pi\varepsilon_0}\left(\frac{[\mathrm{d}p/\mathrm{d}t]}{c^2r^2} + \frac{[\mathrm{d}^2p/\mathrm{d}t^2]}{c^3r}\right)\sin\vartheta$$

and

$$E_\varphi = B_r = B_\vartheta = 0.$$

The terms including a factor of $[p]$ are the fields due to the static field. Again if we take the limit $c \to \infty$ then this reduces to the equations for a static dipole as propagation time becomes zero. Far from the dipole the $1/r^3$ terms dominate and we have

$$E_r \approx 0, \qquad E_\vartheta \approx \frac{1}{4\pi\varepsilon_0}\frac{[\mathrm{d}^2p/\mathrm{d}t^2]}{c^2r}\sin\vartheta, \qquad \text{and} \qquad B_\varphi \approx \frac{1}{4\pi\varepsilon_0}\frac{[\mathrm{d}^2p/\mathrm{d}t^2]}{c^3r}\sin\vartheta.$$

These equations are much simpler and we will work with them. We see that $B = E/c$ as we would expect and also that $\boldsymbol{E}$ and $\boldsymbol{B}$ are perpendicular to each other and to $\boldsymbol{e_r}$ which is the propagation direction. The two fields are in phase and the $\sin\vartheta$ term means that the fields are strongest around the 'equator' of the dipole and are zero along the z axis. The $\boldsymbol{B}$ field is azimuthal (in the $\boldsymbol{e_\varphi}$ direction) which is also the case with a current carrying wire along the z-axis. Far from the origin $\boldsymbol{E}$ and $\boldsymbol{B}$ have a form similar to equation 21.1 which means that they look like harmonic spherical waves with an extra $\sin\vartheta$ term which decreases the magnitude towards the poles. As one last observation note that the magnitude of the pointing vector is $S = \left|\mu_0^{-1}\boldsymbol{E}\times\boldsymbol{B}\right| \propto 1/r^2$, so intensity decays as $I = \langle S\rangle \propto 1/r^2$ as we would expect for light which famously follows an inverse square law for intensity.

All of these points hold for the approximate version of the equations 'far' from the dipole. So when is this a good approximation? What counts as far? The answer turns out to be that far is not very far at all. In fact the $1/r$ terms dominate enough that the approximation is valid for r being only a few wavelengths which is very small. Since the approximation is valid so close and we aren't doing quantum mechanics we really needn't consider anything other than the far field approximation here.

24 Electromagnetic Waves in Dielectrics

An electromagnetic field in a dielectric with permittivity $\varepsilon = \varepsilon_0\varepsilon_r$ and permeability $\mu = \mu_0\mu_r$ can be shown to lead to the wave equations

$$\nabla^2\boldsymbol{E} = \varepsilon\mu\partial_t^2\boldsymbol{E}, \qquad \text{and} \qquad \nabla^2\boldsymbol{B} = \varepsilon\mu\partial_t^2\boldsymbol{B}.$$

We assume that the dielectric is linear, isotropic and homogenous. That is the polarisation is linearly proportional to the applied electric field and the material properties are the same in all directions and everywhere in space.

These corresponds to waves with phase velocity

$$v_p = \frac{\omega}{k} = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{c}{\sqrt{\mu_r\varepsilon_r}}.$$

We see that the speed of light in the medium is attenuated by a factor of $1/\sqrt{\mu_r\varepsilon_r}$. We call this factor, $n = \sqrt{\mu_r\varepsilon_r}$, the **refractive index**. For a material that is not a ferromagnet we find that $\mu_r \approx 1$ and so $n = \sqrt{\varepsilon_r}$ is usually a good approximation for our purposes and

$$v_p = \frac{c}{\sqrt{\mu_r\varepsilon_r}} = \frac{c}{n} \approx \frac{c}{\sqrt{\varepsilon_r}}.$$

Recall that an electric field causes a polarisation $\boldsymbol{P} = \chi_E\varepsilon_0\boldsymbol{E}$ where $n^2 \approx \varepsilon_r = 1 + \chi_E$. Therefore if we can calculate the polarisation, $\boldsymbol{P}$, we can find ε_r and from this we can find the refractive index, n. We can calculate $\boldsymbol{P}$ from the polarisation of a single molecule, $\boldsymbol{p}$, and then $\boldsymbol{P} = N\boldsymbol{p}$ where N is the number density of molecules.

24.1 Snell's Law Derivation From Fermat's Principle

Fermat's principle of stationary time states that a light ray travelling between two points takes a path such that the time taken is stationary with respect to variations in the path. Roughly speaking this means that a slight variation in the path to a nearby path will cause, at most, second order changes in the traversal time. While the principle simply states 'stationary' it is often assumed that the time taken is in fact minimal. After all there is an infinite number of paths that take arbitrarily long to reach the point.

Fermat's principle makes no assumptions about light as an electromagnetic wave but we will see that it leads to many correct predictions. For example, we will use it here to derive Snell's law. This principle is the basis of geometric optics, which is the field of optics where light is treated as a ray and its path through space calculated based on various rules for how this ray interacts with the mediums it travels through.

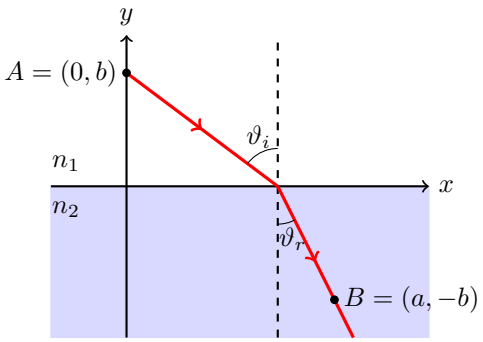


Figure 24.1: Light being refracted at the boundary between two media.

Consider light travelling from point A to B and along the way it changes from one medium with refractive index n_1 to another with refractive index n_2. Within a medium the only stationary path is a straight line, which is what we expect. What we don't know is what angles the light will meet the medium at. It is traditional to consider the angle to the normal with the incoming angle being called the angle of incidence, ϑ_i, and the outgoing angle being called the angle of refraction, ϑ_r Set up coordinates as in figure 24.1. Notice that the diagram shows $\vartheta_r < \vartheta_i$ but this is not necessarily the case. We can fix both angles by choosing the point $C = (x, 0)$, along the x-axis at which the light passes from one medium to the other.

The total time taken for the light to travel from A to B is $T = T_1 + T_2$ where T_1 is the time to travel the path in medium 1 and T_2 is the time to travel the path in medium 2. Using the fact that the speed of light in a given medium is c/n_i and the fact that time is distance over speed we have

$$\begin{aligned} T &= T_1 + T_2 \\ &= \frac{\left|\overrightarrow{AC}\right|}{c/n_1} + \frac{\left|\overrightarrow{CB}\right|}{c/n_2} \\ &= \frac{\sqrt{b^2 + x^2}}{c/n_1} + \frac{\sqrt{b^2 + (a-x)^2}}{c/n_2}. \end{aligned}$$

For the time to be stationary we require

$$\frac{\mathrm{d}T}{\mathrm{d}x} = 0.$$

Calculating the derivative we have

$$\frac{\mathrm{d}T}{\mathrm{d}x} = \frac{n_1}{c}\frac{x}{\sqrt{b^2 + x^2}} - \frac{n_2}{c}\frac{a-x}{\sqrt{b^2 + (a-x)^2}} = 0.$$

Hence

$$\frac{n_1 x}{\sqrt{b^2 + x^2}} = \frac{n_2(a-x)}{\sqrt{b^2 + (a-x)^2}}.$$

Now simply applying the definition of $\sin\vartheta$ as opposite over hypotenuse we have

$$n_1 \sin\vartheta_i = n_2 \sin\vartheta_r.$$

This is **Snell's law** of refraction. This derivation was based on geometric optics and Fermat's principle of stationary time. In later sections we will derive this same result with Huygen's principle and full electromagnetic theory.

24.2 Polarisation and Refractive Index

We can use Snell's law to measure the refractive index of a material. If we do this we find that n decreases as frequency decreases. If we ask why this is the first thing we may think of is since $n^2 = \varepsilon_r$ is related to polarisation the polarisation must also be frequency dependent. To find out why we may ask what it is that causes polarisation. There are three components that we might consider:

- The orientation of polar molecules leads to polarisation when the molecules align their dipoles with the electromagnetic field. This requires the entire molecule to rotate so happens on the time scales of rotational frequencies. For example in water it takes approximately 10 ps.
- The polarisation of ions leads to polarisation when the molecules distort to align the resulting dipole with the electromagnetic field. This requires only part of the molecule to move and so happens on the time scale of vibrational frequencies. For example in water it takes approximately 1 ps.
- The polarisation of non-polar molecules leads to polarisation when the electron cloud is distorted to create a dipole aligned with the electromagnetic field. This requires only electrons to move and so happens on very short time scales of about 1 fs.

From all of these mechanisms the key point is that polarisation is not instantaneous. If the field is oscillating back and forth we expect the polarisation to lag behind. How much it lags behind will depend on how fast the field oscillates. There si just one problem. This explanation would lead to us expecting that refractive index decreases as frequency increases. This is the exact opposite of what we see experimentally. To find out why this isn't what happens we will need a more careful treatment of the oscillation of the dipoles with changes in the field. This is what we will do in the next section.

25 Oscillator Model

For simplicity we assume that interactions between electrons are negligible and that we can model the electron cloud as simply being stuck in a potential well caused by the Coulomb interaction with the nucleus. If we expand this potential to second order then, by definition, we have a quadratic potential. We can therefore model the motion of the electron cloud as a harmonic oscillator obeying

$$m_\mathrm{e}\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = q_\mathrm{e}E_x - m_\mathrm{e}\omega_0^2 x - m_\mathrm{e}\gamma\frac{\mathrm{d}x}{\mathrm{d}t}.$$

Here m_e and q_e are the mass and charge of the electron cloud, x is the displacement of the centre of mass of the electron cloud from the centre of mass of the atom and γ is a damping coefficient. We also assume that any radiation is x polarised for simplicity so

$$E_x = E_0\cos(\omega t).$$

If we include this we then have a damped harmonic oscillator with resonant frequency ω being driven at frequency ω.

In the undamped ($\gamma = 0$) case this has the solution

$$x(t) = x_0\cos(\omega t) = \frac{q_\mathrm{e}/m_\mathrm{e}}{\omega_0^2 - \omega^2}E_i\cos(\omega t) = \frac{q_\mathrm{e}/m_\mathrm{e}}{\omega_0^2 - \omega^2}E(t).$$

Notice that this motion of the electron cloud sets up a dipole with dipole moment $p = q_\mathrm{e}x$. Notice also that $q_\mathrm{e} < 0$ and the dipole vector, $\boldsymbol{p}$, is conventionally defined to point from negative to positive, meaning that $\boldsymbol{p}$ points in the opposite direction to $\boldsymbol{x}$. From this we can draw conclusions based on two different cases for the values of ω and ω_0:

- For the case of $\omega < \omega_0$, i.e. below the natural frequency, the electron shell oscillates exactly π out of phase with the electric field, $\boldsymbol{E}(t)$. That is $\boldsymbol{x}$ points in the opposite direction to $\boldsymbol{E}$ and so $\boldsymbol{p}$ points in the same direction to $\boldsymbol{E}$.
- For the case of $\omega > \omega_0$, i.e. above the natural frequency, the electron cloud oscillates exactly in phase with the electric field, $\boldsymbol{E}(t)$. That is $\boldsymbol{x}$ points in the same direction as $\boldsymbol{E}$ and so $\boldsymbol{p}$ points in the opposite direction to $\boldsymbol{E}$.

We can use the solution for $x(t)$ to find the atomic dipole moment, $\boldsymbol{p}_{\mathbf{atom}}$ and the polarisation of the atom, which is simply the dipole moment per unit volume. From this we can calculate the refractive index:

$$\begin{aligned} n^2 &= \varepsilon_r \\ &= 1 + \frac{P}{\varepsilon_0 E} \\ &= 1 + \frac{N p_{\text{atom}}}{\varepsilon_0 E} \\ &= 1 + \frac{N q_\text{e} x}{\varepsilon_0 E} \\ &= 1 + \frac{N q_\text{e}^2}{\varepsilon_0 m_\text{e}(\omega_0^2 - \omega^2)} \end{aligned}$$

where N is the number density (i.e. the number of atoms per unit volume).

This is a simple model so it won't give exactly the right solution. In particular it has the following shortcomings:

- In reality there won't be one single resonant frequencies but different components of motion will have different resonant frequencies, ω_{0i}, and we will have to sum over these.
- We ignore the interaction between atoms and in particular the effect that atomic dipole has on its neighbours. This can be accounted for with something called the Claussius–Mossotti correction but we won't take it into account.
- There is a singularity at $\omega = \omega_0$.

In the case where we have non-zero damping the mathematics is harder but we can still find a solution. The correct damping regime to consider is light damping where $0 < \gamma \ll 1/(2\omega_0)$. Recall that $\gamma = 2\omega_0$ leads to critical damping. For non-zero damping the initial displacement at $t = 0$ has explicit ω dependence and a phase shift as well. The solution turns out o be of the form $x = x_0 \cos(\omega t + \Phi)$ leading to an atomic dipole of the form

$$p_x = q_\text{e} x = p_0 \cos(\omega t + \Phi)$$

which can be shown to give solutions of the form

$$p_0(\omega) = q_\text{e} x_0(\omega) = \frac{q_\text{e}^2 E_0/m_\text{e}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}}$$

and

$$\Phi = \arctan\left[\frac{-\gamma\omega}{\omega_0^2 - \omega^2}\right].$$

We can solve this for x and it is easier to do this using complex exponentials for trig and we simply implicitly take the real part when necessary. Using $x = x_0 \exp[-i(\omega t + \Phi)]$ we find that

$$x(\omega) = x_0 e^{-i(\omega t + \Phi)} = x_0 e^{-i\Phi} e^{-i\omega t} = \left[\frac{q_\text{e} E_0/m_\text{e}}{\omega_0^2 - \omega^2 - i\gamma\omega}\right] e^{-i\omega t}$$

and the dipole moment is then

$$p_x(\omega) = \frac{q_\text{e}^2 E_x/m_\text{e}}{\omega_0^2 - \omega^2 - i\gamma\omega}.$$

We then have

$$\varepsilon_r = \varepsilon_r' + i\varepsilon_r'' = 1 + \frac{N q_\text{e}^2}{\varepsilon_0 m_\text{e}(\omega_0^2 - \omega^2 - i\gamma\omega)} \tag{25.1}$$

where ε_r' and ε_r'' are the real and imaginary components of ε_r. Using a similar notation for the refractive index, $n = n' + in''$ we have

$$\varepsilon_r = n^2 \implies \varepsilon_r' = n'^2 - n''^2, \qquad \text{and} \qquad \varepsilon_r'' = 2n'n''.$$

Since ε_r depends on the frequency, ω, of the radiation we find that the refractive index also depends on the frequency of the radiation meaning that different colours of light are diffracted/refracted/slowed down by a different amount. We can also replace ω_0^2 with $\omega_{01}^2 + \omega_{02}^2 + \cdots$ where ω_{0i} are the relevant natural frequencies of the oscillations.

A few features of this new equation show us that it is a good model or raise some more questions:

- If $\gamma \neq 0$ then we don't have any singularities, even at $\omega = \omega_0$.
- The real part of the refractive index, n', increases with the frequency as ω becomes closer to ω_{0i} which is what we observe experimentally.
- Ignoring increases for $\omega \approx \omega_{0i}$ we have a general decreasing trend in n' with ω.
- For $\omega > \max\{\omega_i\}$ we have $n' < 1$ which means that light at these frequencies travels faster than $3 \times 10^8\,\mathrm{m\,s^{-1}}$. This seems initially like it breaks the rules of special relativity until you recall that the speed of light applies only in a vacuum.
- Using this model we can measure refractive index dependent things at some frequencies and extrapolate to other frequencies.

25.1 Interpreting the Oscillator Model

Consider the case of a free electron. This corresponds to setting $\omega_0 = 0$ and so $\omega_0 < \omega$ which we saw in the previous section meant that the electron oscillates in phase with $\boldsymbol{E}$. This initially seems incorrect as the electron has a negative charge and so should move in the *opposite* direction to $\boldsymbol{E}$. This would indeed be the case if $\boldsymbol{E}$ was constant, but it isn't. The *acceleration* of the electron is what must be in the opposite direction of $\boldsymbol{E}$. The displacement is then given by integrating twice and integrating a sinusoid twice gives, up to a positive constant factor, the same sinusoid back but negative. So the acceleration is in the opposite direction to $\boldsymbol{E}$ but the position is in the same direction as $\boldsymbol{E}$.

If instead $\boldsymbol{E}$ is static, or equivalently $\omega = 0$, then the electron is displaced in the opposite direction to $\boldsymbol{E}$ which agrees with the case of $\omega < \omega_0$ (even though $\omega = \omega_0 = 0$).

For the case of $\omega \approx \omega_0$ we have a large spike in amplitude. In the case of the undamped oscillator this is a singularity and the amplitude becomes infinite. In the lightly damped case the amplitude is finite but still much larger than in other regions. This is due to resonance, recall that a harmonic oscillator has its maximum amplitude at some resonant frequency, ω_r, which is just slightly lower than ω_0.

Consider the case of very light damping such that γ becomes infinitesimal. In this case the imaginary part of x or ε_r becomes a delta distribution at ω_0 and the real part becomes discontinuous. This is because there is no limit on vibration amplitude in this mathematical model. In a real material there is damping as oscillating dipoles will lose energy due to random thermal collisions. The dipoles will also re-radiate energy. This gives a physical significance the the complex parts of ε_r and n which up until now we have treated purely as a mathematical convenience. These terms correspond to how much a given frequency is absorbed by the medium. For example glass has $\omega_0 \approx 10^6\,\mathrm{rad\,s^{-1}}$, which corresponds to light in the UV region of the spectrum. Glass also strongly absorbs UV radiation at this frequency which corresponds to the high peak at this point in ε_r'' and n''.

For low frequencies, $\omega \ll \omega_0$, $\tan\Phi$ is small and negative, which means that Φ is small and negative. This means that the oscillations have a negligible phase lag. For $\omega = \omega_0$ we have $\Phi = -\pi/2$ and for $\omega \gg \omega_0$ we Φ approaches $-\pi$. So the phase lag increases as ω increases. This corresponds to ε_r dropping from greater than 1 for $\omega \ll \omega_0$ to being smaller than 1 for $\omega \gg \omega_0$. At low frequencies the polarisation of matter opposes the changing electric field and at higher frequencies it reinforces the changing field.

26 Huygens' Principle and Colour

26.1 Huygens' Principle

So far we have considered electromagnetic radiation at a point propagating outwards from a dipole. We haven't seen how the fields from many dipole oscillators can combine to form an electromagnetic wave that fills space. The simplest explanation is called **Huygens' principle**:

> Every point on a primary wavefront serves as the source of spherical secondary wavelets which travel at the speed of light such that the primary wavefront at some time later is the envelope of these wavelets.

What this means is that every point emits waves which combine to form a new wave front. This can

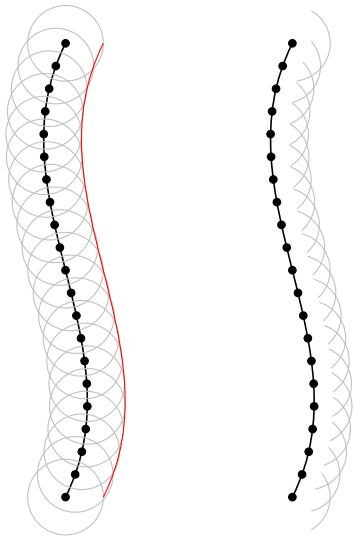


Figure 26.1: Many points along a wavefront emit wavelets which form a new wavefront.

be seen in figure 26.1. What isn't clear is why we only consider wavelets propagating in the forward direction. The reason for this will be given later when we have discussed superposition and interference.

26.2 Colour

As human beings we give particular attention to the visible part of the spectrum because, well, we can see it. The physics isn't fundamentally different for the rest of the spectrum but it is worth spending some time considering the visible spectrum and how we see light.

There are two types of colour. There are **pure/spectral/chromatic** colours such as red, blue, green, yellow, etc. which are formed of a single wavelength and there are colours, such as purple or brown, which are a mixture of other colours. The pure colours are the ones we see when we pass light through a prism and split it up into a rainbow.

Wavelength (nm)	Colour		Complementary Colour
400–435	Violet	Yellow-Green	
435–480	Blue	Yellow	
480–490	Green-Blue	Orange	
490–500	Blue-Green	Red	
500–560	Green	Magenta	
560–580	Yellow-Green	Violet	
580–595	Yellow	Blue	
595–605	Orange	Green-Blue	
605–700	Red	Cyan	

While many colours exist as a pure, single wavelength this is not how humans perceive colour. The human eye has three types of cones which are cells that can detect light. Each type preferentially detects light at

one of three different wavelengths corresponding to red, green, and blue light. The brain then interprets both which cones are activated and also how much they are activated and combines this information to give us colour vision. For example if the red and green cones are activated a lot and the blue cones aren't activated at all then we see yellow:

■ + ■ = ■.

If instead red is activated a lot and blue is activated about half as much then we see a sort of pink-purple colour:

■ + ■ = ■.

We use this to produce colour in screens. Each pixel of a screen actually produces red, green, and blue, colour (RGB) in varying amounts which combine to give the colours we see. The standard way for this to happen is for the amount of each colour to be given by an 8 bit number which allows for values from 0 to 255. So for three different colours there are $256^3 = 2^{8^3} = 2^{24} = 16777216$ different combinations that can be shown. This is just one of many ways of telling a computer what colour to show, an equivalent way uses hexadecimal instead and each colour is a six digit hexadecimal number ranging from #000000 = 0 for none of any colour (black) to #FFFFFF = 16777215 for all of each colour (white). This isn't really that different as we can split the hexadecimal number into three two digit hexadecimal numbers each of which can be translated into binary to give the same three number RGB colour definition.

Another aspect that needs to be considered is that the cones don't all only activate at a single wavelength but at a range of wavelengths and these ranges all overlap to some extent so light of a single wavelength can stimulate more than one type of cone at once which is how we see pure colours that aren't red, green, or blue.

There are many mechanisms by which an object that doesn't produce its own light can become coloured in white light. We will discuss two right now. The simplest is colour by absorption. For example Cu^{2+} absorbs most light except blue-green light and so when we see something with lots of Cu^{2+} ions most of the light that bounces off of it is blue-green so this is the colour we see the object. The second mechanism is more complicated and involves scattering light at different wave lengths different amounts in such a way that most colours are scattered away before reaching us. This is the mechanism by which the sky appears blue. Scattering in the atmosphere scatters higher frequencies the most the light leaving the sun is pretty much white and so if you look directly at the sun[12] you will see white light. Looking near the sun you see red/yellow light that has been scattered only a little bit. Looking far from the sun you see blue light which has been scattered far from the sun. The only reason that the sky isn't purple (as this light is the most scattered) is because the sun doesn't actually produce much purple light.

Part V

Light at Boundaries

Reflection and Refraction

27 Laws of Reflection and Refraction

We are mostly interested in this part in light incident on a sudden boundary between materials with different refractive indices. We will assume that the boundaries are smooth (on the scale of the wavelength). Recall that the angle of incidence is defined as the angle the incoming light makes to the surface normal pointing out of the surface. Similarly the angle of reflection is the angle that the reflected light makes to the same normal and the angle of transmission is the angle the transmitted light makes to the surface normal pointing into the surface.

27.1 Snell's Law

We have seen already in section 24.1 a derivation of Snell's law which depended on Fermat's principle—that light minimises the time taken to travel between points. We can also derive the same law with a

[12] don't do this

full treatment of the equations for electromagnetic waves.

Consider a boundary between two media with surface normal $\boldsymbol{e}$. Incident on this surface is light with wave vector $\boldsymbol{k_i}$. We would expect in general that there will be both reflection and transmission. The wave vector, $\boldsymbol{k_i}$, and the surface normal, $\boldsymbol{e}$, define a plane, called the plane of incidence. For a smooth surface all reflected and transmitted light will still be in this plane. The equation of the incoming light is

$$\boldsymbol{E_i} = \boldsymbol{E_{0i}} \cos(\boldsymbol{k_i} \cdot \boldsymbol{r} - \omega t)$$

where $\boldsymbol{E_{0i}}$ is the amplitude and ω is the frequency of the light. The energy of light is proportional to ω and so to conserve energy the reflected and transmitted light must have the same frequency as the incident light. This means that we can express the electric field of the reflected light as

$$\boldsymbol{E_r} = \boldsymbol{E_{0r}} \cos(\boldsymbol{k_r} \cdot \boldsymbol{r} 0\omega t - \Phi_r)$$

where Φ_r is some phase difference to be calculated. Similarly the transmitted field is given by

$$\boldsymbol{E_t} = \boldsymbol{E_{0t}} \cos(\boldsymbol{k_t} \cdot \boldsymbol{r} - \omega t - \Phi_t).$$

We will start by computing the direction of propagation and the time dependence which means finding the wave vectors and phase factors. Later we will work out how the amplitudes are related.

In order to satisfy Maxwell's equations the components of $\boldsymbol{E}$ and $\boldsymbol{H}$ which are parallel to the interface must be continuous across the boundary. We choose to define our coordinate system such that $\boldsymbol{e_z} = \boldsymbol{e}$ is the surface normal and $\boldsymbol{e_x}$ lies in the plane of incidence. This means that none of the wave vectors have a component in the $\boldsymbol{e_y}$ direction. For example the incoming wave vector can be written as $\boldsymbol{k_i} = k_i \sin\vartheta_i \boldsymbol{e_x} - k_i \cos\vartheta_i \boldsymbol{e_z}$ where ϑ_i is the angle of incidence defined as the angle $\boldsymbol{k_i}$ makes to $\boldsymbol{e_z}$. In order for the boundary conditions at the surface to hold at all times we must have that

$$\boldsymbol{k_i} \cdot \boldsymbol{r} = \boldsymbol{k_r} \cdot \boldsymbol{r} - \Phi_r = \boldsymbol{k_t} \cdot \boldsymbol{r} - \Phi_t.$$

Rearranging this we have that $(\boldsymbol{k_i} - \boldsymbol{k_r}) \cdot \boldsymbol{r} = \Phi_r$. Compare this to the equation $\boldsymbol{n} \cdot \boldsymbol{r} = a$ for some constant a which defines a plane with surface normal $\boldsymbol{n}$. Since we are considering the fields at the boundary we have already restricted $\boldsymbol{r}$ to be on the surface which means that the surface normal to the plane we are defining with this equation is $\boldsymbol{k_i} - \boldsymbol{k_r}$. This means that $\boldsymbol{k_i} - \boldsymbol{k_r} \propto \boldsymbol{e_z}$. This means that the x and y components must cancel so $k_{ix} = k_{rx}$ and $k_{iy} = k_{ry}$. We know that $k_{iy} = 0$ from how we define the coordinate system so $k_{ry} = 0$ also. Simple geometry shows that $k_{ix} = k_i \sin\vartheta_i$ and $k_{rx} = k_i \sin\vartheta_r$. Converting wave numbers to wavelengths we have

$$\frac{2\pi}{\lambda_i} \sin\vartheta_i = \frac{2\pi}{\lambda_r} \sin\vartheta_r.$$

Since ω is the same in all cases and the incident and reflected wave are in the same medium so they must have the same wavelength, $\lambda_i = \lambda_r$. This means that $\vartheta_i = \vartheta_r$. This is called the **law of reflection**, that the incident and reflected angle are the same. Similar calculations give imply that

$$k_i \sin\vartheta_i = k_t \sin\vartheta_t.$$

Introducing the refractive indices which relate the wavelengths by $n_i\lambda_i = n_t\lambda_t$ and cancelling common factors we get

$$n_i \sin\vartheta_i = n_t \sin\vartheta_t.$$

This is **Snell's law** again.

27.2 Total Internal Reflection

Consider a light incident on a boundary from a high refractive index to a low refractive index. From Snell's law we have

$$\sin\vartheta_l = \frac{n_h}{n_l} \sin\vartheta_h$$

where subscript l and h refer to quantities measured in the low and high refractive index media respectively. We expect that $\vartheta_l > \vartheta_h$ and this is indeed the case up to some critical angle $\vartheta_l = \vartheta_\mathrm{c}$ at which

point $\sin\vartheta_c = 1$ and instead of being transmitted the incident light is reflected. The critical angle is simply given by

$$\vartheta_c = \arcsin\left(\frac{n_l}{n_h}\right).$$

We call this total internal reflection. This is the mechanism by which an optical fibre works, light is shone in at one end at such an angle that it always meets the outside (or in real use the outer layer of cladding) above the critical angle and is reflected back inside the fibre.

Another use of this effect is measuring the refractive index of a material as it is easy to measure the critical angle with n_l being some known value, such as air so $n_l \approx 1$. For example if we measure the critical angle of a glass block to be 42° then

$$n_h = \frac{1}{\sin\vartheta} = \frac{1}{\sin 42^\circ} = 1.5$$

which is a fairly standard refractive index for glass.

27.3 Dispersion

We know that the refractive index of a material differs with frequency. This means that different colours of light are refracted a different amount. This is the mechanism by which a prism is able to split white light into a rainbow. This effect also allows us to see which wavelengths *aren't* present in light which gives us the absorption spectra of a source which we can tell us about the chemical composition of the source.

The dispersion of a material is simply how much n varies with λ. For example diamond is highly dispersive and can also be cut to have lots of flat planes. If these planes are cut at the correct angles then we also get lots of total internal reflection. These two effects combine to mean that most light that enters a diamond leaves out the front and also is separated slightly into different colours. This makes diamonds very nice to look at. In part this is due to the high refractive index of diamond, $n = 2.419$, which ensures a low critical angle. Thus light can only leave at certain angles which gives diamond its characteristic sparkle.

27.4 Lenses

Lenses rely on the refraction of light to bend incident light such that it converges on a focal point at focal length f. Simple lenses can be described by three numbers. The refractive index of the material it is made from and the radius of curvature of the two faces. The radius of curvature is the radius of a sphere which has the same curvature as the lens and it is a signed number with the sign being positive if the edges of the lens bend in the positive direction along the optical axis and negative if the centre of the lens bulges along the optical axis (see figure 27.1). We say that lens is thin if its thickness, Δ, satisfies $\Delta < |R_1|$ and $\Delta < |R_2|$ where R_i are the radii of curvature. For a thin lens the focal distance is given by the **lens makers formula**:

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2}\right]$$

One problem is that since n varies with wavelength the focal point also varies with wavelength. This is called a chromatic aberration.

27.5 Optical Illusions

27.5.1 Snell's Window

Suppose that you are underwater and you look up. If the surface is sufficiently flat you will see only be able to see out of the water in some circle (see figure 27.2). Outside of this circle the angle is such that instead total internal reflection would occur and you would see back down into the water. This is best demonstrated by looking at figure 27.3.

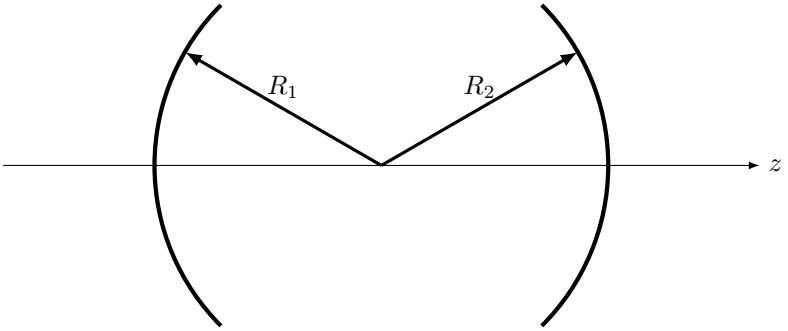


Figure 27.1: Two possible faces of a lens with the same absolute radius of curvature but different signs, $R_1 > 1$ and $R_2 < 1$.

Figure 27.2: Snell's window in real life. Image credit: `https://commons.wikimedia.org/wiki/File:US_Navy_110607-N-XD935-191_Navy_Diver_2nd_Class_Ryan_Arnold,_assigned_to_Mobile_Diving_and_Salvage_Unit_2,_snorkels_on_the_surface_to_monitor_multi.jpg` accessed on 27/04/2021.

27.5.2 Mirages

The refractive index of air depends on its density and hence on the temperature of the air. This means that on a hot day the refractive index changes a non-negligible amount between the ground and a few metres up as the air is hotter higher up. Imagine light coming off the top of a tree towards the ground. We can consider each step along the path the light takes to be going through a boundary with two different refractive indexes. If the conditions are just right this can cause the light to refract slightly at each step until it is travelling upwards. If we see this light that came from the top of a tree but is travelling upwards towards us our brains are unable to comprehend the process that has actually taken place and instead assume that the light has reflected off of something. Our monkey brains are also trained to see anything reflective as a source of water and therefore it is common for it to look like there

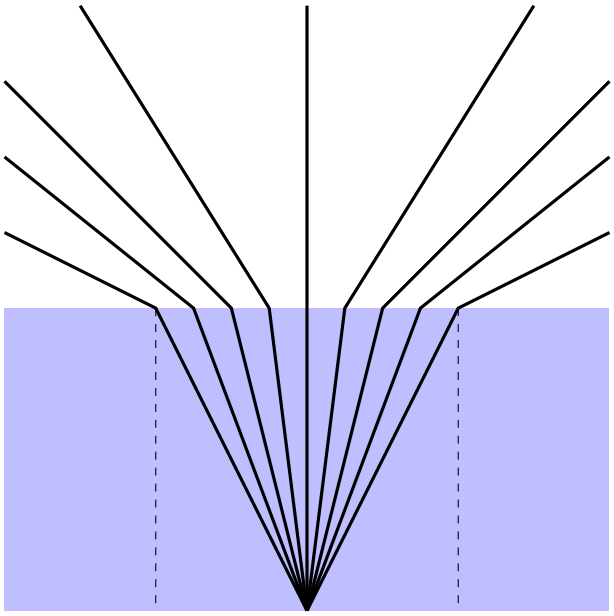


Figure 27.3: Looking up from underwater you only see out of a circle, known as Snell's window. Outside of this circle total internal reflection occurs and you see back down into the water.

is water on the ground even though there isn't. This is the mechanism behind heat hazes and mirages.

28 The Fresnel Equations

In the last section we discussed reflection and refraction with a focus on the angles that the rays make to the surface normals. We found that an incident ray is reflected at the same angle it is incident and that the transmitted ray will be at an angle satisfying Snell's law:

$$n_i \sin \vartheta_i = n_t \sin \vartheta_t.$$

To find this we solved for the wave vectors $\boldsymbol{k_r}$ and $\boldsymbol{k_t}$. We didn't however, find the amplitudes of the reflected and transmitted waves. This is what we will do in this section. To find the amplitudes we need to be slightly more cautious in our calculations and consider the direction the waves oscillate. There are two linearly independent directions that an incident wave can oscillate and all other cases are simply a superposition of these two cases:

- If $\boldsymbol{E_{0i}}$ is perpendicular to the plane of incidence then we call the wave **transverse electric** or S-polarised[13].
- If $\boldsymbol{E_{0i}}$ is parallel to the plane of incidence then we call the wave **transverse magnetic** or P-polarised[14].

We will treat these two cases separately.

28.1 S-Polarised Light

The boundary conditions for $\boldsymbol{E}$ are that the components parallel to the boundary are continuous across the boundary. For S-polarised light the incoming amplitudes are perpendicular to the plane of incidence which necessarily makes them parallel to the boundary. Thus we must have at the boundary that

$$E_{0i} + E_{0r} = E_{0t}. \tag{28.1}$$

The boundary conditions are slightly more tricky for $\boldsymbol{H}$. It is conventional to define $\boldsymbol{k}$, $\boldsymbol{E}$, and $\boldsymbol{H}$ such that they form a right handed system. Considering an incoming wave being reflected the direction of $\boldsymbol{k}$ is set by the direction of travel and the direction of $\boldsymbol{E}$ is set by demanding S-polarised light. The result is that we have to choose to define the direction of $\boldsymbol{H}$ to keep a right handed system as shown in figure 28.1. In particular we have $\boldsymbol{k} \times \boldsymbol{E} = v\boldsymbol{B}$. We can then show that the continuity equation for $\boldsymbol{H}$ is

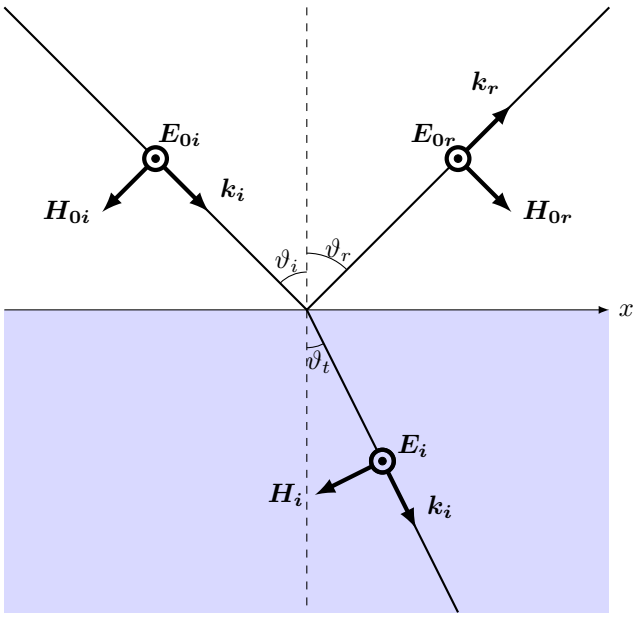


Figure 28.1: S-polarised light reflection and transmission.

$$H_{0i} \cos \vartheta_i - H_{0r} \cos \vartheta_r = H_{0t} \cos \vartheta_t.$$

[13] S for *senkrecht*, German for perpendicular.
[14] P for *parallel*, German for parallel.

Writing $H = B/\mu = E/(v\mu) = En/(c\mu)$ and $\vartheta_r = \vartheta_i$ we have

$$\begin{aligned}\frac{B_{0i}}{\mu_i}\cos\vartheta_i - \frac{B_{0r}}{\mu_i}\cos\vartheta_i &= \frac{B_{0t}}{\mu_t}\cos\vartheta_t \\ \frac{E_{0i}n_i}{c\mu_i}\cos\vartheta_i - \frac{E_{0r}n_i}{c\mu_i}\cos\vartheta_i &= \frac{E_{0t}n_t}{c\mu_t}\cos\vartheta_t \\ (E_{0i} - E_{0r})\frac{n_i\cos\vartheta_i}{c\mu_i} = E_{0t}\frac{n_t\cos\vartheta_t}{c\mu_t} &= (E_{0i} + E_{0r})\frac{n_t\cos\vartheta_t}{c\mu_t}\end{aligned}$$

where in the last equality we have used the continuity equation for the electric field (equation 28.1). Since we can link ϑ_i and ϑ_t by Snell's law and we can measure the material properties μ_i, μ_t, n_i, and n_t this equation gives us a relationship between the amplitudes of the incident magnetic field, E_{0i}, and the reflected field, E_{0r}.

We further assume that the magnetic field is not that strong and so it is a reasonable approximation to assume that $\mu_i = \mu_t = \mu_0$. We define the **Fresnel reflection coefficient** as the ratio E_{0r}/E_{0i} which for S-polarised light we find to be

$$r_\mathrm{S} = \frac{E_{0r}}{E_{0i}} = \frac{n_i\cos\vartheta_i - n_t\cos\vartheta_t}{n_i\cos\vartheta_i + n_t\cos\vartheta_t}.$$

We can use the same equations and instead eliminate E_{0r} leaving us with a relationship between E_{0t} and E_{0i}. Similarly we can then define the **Fresnel trnasmission coefficeint** for S-polarised light:

$$t_\mathrm{S} = \frac{E_{0t}}{E_{0i}} = \frac{2n_i\cos\vartheta_i}{n_i\cos\vartheta_i + n_t\cos\vartheta_t}.$$

28.2 P-Polarised Light

The case of P-polarised light is very similar with slightly different boundary conditions. This time the $\boldsymbol{H}$ continuity equation is simple:

$$H_{0i} + H_{0r} = H_{0t}$$

and the $\boldsymbol{E}$ continuity equations are the more complicated

$$E_{0i}\cos\vartheta_i - E_{0r}\cos\vartheta_i = E_{0t}\cos\vartheta_t.$$

We can similarly define **Fresnel coefficients** for reflection and transmission of P-polarised light:

$$\begin{aligned}r_\mathrm{P} &= \frac{E_{0r}}{E_{0i}} = \frac{n_t\cos\vartheta_i - n_i\cos\vartheta_t}{n_i\cos\vartheta_t + n_t\cos\vartheta_i}, \\ t_\mathrm{P} &= \frac{E_{0t}}{E_{0i}} = \frac{2n_i\cos\vartheta_i}{n_i\cos\vartheta_t + n_t\cos\vartheta_i}.\end{aligned}$$

28.3 Fresnel Coefficients With Incidence Angle

Consider light incident on glass ($n = 1.5$) travelling from air ($n = 1$). For normal, or almost normal incidence ($\vartheta_i \approx 0$) we see that for both S and P-polarised light we have $|t| \gg |r|$. Comparing this to our daily experiences this makes sense as we are used to light passing straight through glass. For grazing incidence ($\vartheta_i \approx \pi/2$) we see that $|t| \approx 0$ and $|r| \approx 1$. Again this should be familiar that glass viewed at a large incidence angle is almost entirely reflective.

For all angles of incidence it turns out that $r_\mathrm{S} < 0$. This corresponds to the incident and reflected $\boldsymbol{E}$ fields being antiparallel. That is S-polarised light undergoes a phase shift of π when reflecting off of a higher refractive index medium.

For P-polarised light at some point r_P becomes zero. The angle, ϑ_B, at which this occurs is called **Brewster's angle**. At this angle there is no reflection.

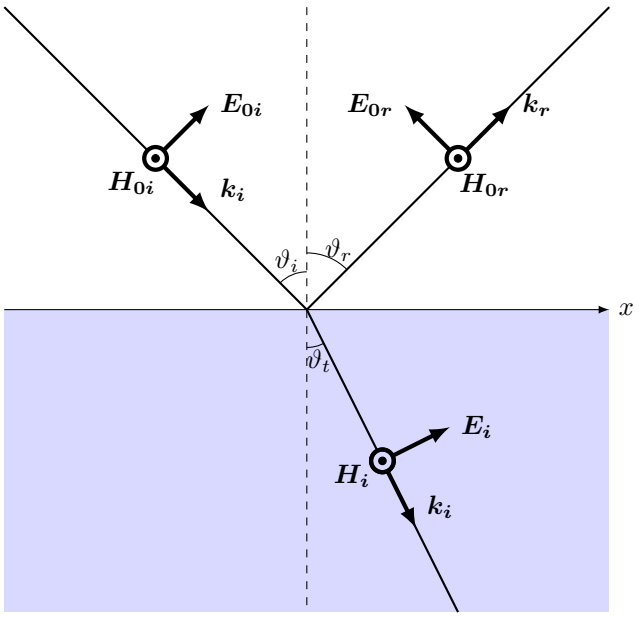


Figure 28.2: P-polarised light reflection and transmission.

The transmission coefficients are similar (but not identical, note the refractive indices in the denominator swap) for both S and P-polarised light. It is mostly in reflection where the polarisation is important.

Now consider the opposite case of light incident on a glass-air boundary from within the glass. For near normal incidence almost all light is transmitted. For P-polarised light there is an angle at which no light is reflected. This is sometimes referred to as the **internal Brewster's angle**. Note that this angle is *not* the same angle as for the air-glass boundary.

There is no phase shift for reflection of S-polarised light (as opposed to the π phase shift for the air-glass case).

Above some critical angle, ϑ_c, both r_S and r_P are 1 and t_S and t_P are 0. This corresponds to total internal reflection.

Perhaps the seemingly weirdest thing is that $t > 1$ which means that, even if some light is reflected, the incident electric field has a smaller amplitude than the transmitted electric field. We will see in the next section why this must be the case.

28.4 Energy Flow

The amplitude of the transmitted field being greater than the amplitude of the incident field initially seems to violate energy conservation. As is often the case in electromagnetism this is because energy scales with the *square* of the amplitude. For an oscillating field the most useful quantity is the intensity, I, defined as the time average of the magnitude of the Poynting vector, $\boldsymbol{S}$:

$$I = \langle S \rangle = \frac{v\varepsilon}{2} E_0^2.$$

Here $v = c/n$ and we replace ε_0 with ε when inside a medium. This gives the power density per unit area normal to the Poynting vector.

Suppose we illuminate an area, A, on a surface. To conserve energy we have to have the energy in per unit time be equal to the energy out per unit time. That is

$$I_i A \cos\vartheta_i = I_r A \cos\vartheta_r + I_t A \cos\vartheta_t$$

where I_i, I_r, and I_t are the intensities of the incident, reflected, and transmitted, field respectively. We can rearrange this and use the fact that $\vartheta_i = \vartheta_r$ to get

$$\frac{I_r}{I_i} + \frac{I_t}{I_i}\frac{\cos\vartheta_t}{\cos\vartheta_i} = 1.$$

The first term gives the fraction of incident energy that is reflected and the second gives the fraction of incident energy that is transmitted. We call the first fraction the **reflectivity**:

$$R = \frac{I_r}{I_i} = \frac{E_{0r}^2}{E_{0i}^2} = r^2.$$

The second term we call the **transmissivity**:

$$T = \frac{I_t}{I_i}\frac{\cos\vartheta_t}{\cos\vartheta_i} = \frac{E_{0t}^2}{E_{0i}^2}\frac{\mu_{ri}^2}{\mu_{rt}^2}\frac{n_t}{n_i}\frac{\cos\vartheta_t}{\cos\vartheta_i} \approx \frac{E_{0t}^2}{E_{0i}^2}\frac{n_t}{n_i}\frac{\cos\vartheta_t}{\cos\vartheta_i} = t^2\frac{n_t}{n_i}\frac{\cos\vartheta_t}{\cos\vartheta_i}.$$

It can be shown that $R + T = 1$ for all incident angle and this is the condition that energy is conserved. This explains why we can have $t > 1$ so long as r is such that $R + T = 1$.

At normal incidence for an air-glass boundary 4 % of the light is reflected. For a single boundary this is negligible and we treat it as pure transmission. The issue arises with multiple boundaries, such as in a telescope or microscope, where there are multiple lenses and therefore the light we get out will be of significantly lower intensity than the light that we put in. The way that the unwanted reflected beams affect the image we get out is also non-trivial.

So far in this section we have assumed that no energy is absorbed by the medium. If this isn't the case then the quantities r, t, n, and ε become complex. We will see this more later.

For a glass-air boundary above the critical angle R becomes 1 and T becomes 0. This is due to total internal reflection.

29 Consequences of the Fresnel Equations

29.1 Total Internal Reflection

In the calculation of the Fresnel coefficients the transmission angle, ϑ_t, appears explicitly. The problem is that we can't define ϑ_t for incidence angles greater than the critical angle. The way we solve this is by using Snell's law and trig identities:

$$\left.\begin{aligned} n_i\sin\vartheta_i = n_t\sin\vartheta_t \implies \sin\vartheta_t = \tfrac{n_i}{n_t}\sin\vartheta_i \\ \sin^2\vartheta_t + \cos^2\vartheta_t \implies \cos\vartheta_t = \sqrt{1-\sin^2\vartheta_t} \end{aligned}\right\} \implies \cos\vartheta_t = \sqrt{1-\frac{n_i^2}{n_t^2}\sin^2\vartheta_i}.$$

Using this we can write the reflection coefficient as

$$r_\mathrm{S} = \frac{\cos\vartheta_i - \sqrt{n_t^2/n_i^2 - \sin^2\vartheta_i}}{\cos\vartheta_i + \sqrt{n_t^2/n_i^2 - \sin^2\vartheta_i}}.$$

For cases where we have total internal reflection we have $n_t/n_i < \sin\vartheta_i$ which means that the square roots result in complex numbers. We can generalise the reflectivity and transmissivity from the last section. In doing so we have to treat the S and P-polarised cases differently:

$$R_\mathrm{S} = r_\mathrm{S}^* r_\mathrm{S}, \qquad \text{and} \qquad R_\mathrm{P} = r_\mathrm{P}^* r_\mathrm{P}.$$

It can be shown that both reflectivities become 1 above the critical angle.

When total internal reflection occurs the transmitted field has intensity 0. However in order for the correct continuity rules to apply we need to have a wave still. It can be shown that the if the transmitted wave vector is $\boldsymbol{k_t}$ then the component parallel to the boundary is $k_t\sin\vartheta_t$ which can be expressed as $k_t(n_i/n_t)\sin\vartheta_i$ using Snell's law. The component perpendicular to the boundary can be expressed as

$$k_t\cos\vartheta = k_t\sqrt{1-\frac{n_i^2}{n_t^2}\sin^2\vartheta_i}.$$

For total internal reflection we can write this as $\pm i\beta$. This imaginary component to the wave vector corresponds to exponential decay of this component of the $\boldsymbol{E}$ field. So for total internal reflection the transmitted ray decays to zero. These are called **evanescent** or **boundary waves**. If we make the low refractive index layer very thin then it is possible to get the transmitted wave out the other side before it decays to zero. This is called **frustrated total internal reflection**.

29.2 Metals and Plasmas

We saw something similar to evanescent waves when we considered electromagnetic fields incident on metal. We saw that the fields penetrate only a small amount, called the skin depth.

Equation 25.1 gives the relative permittivity derived from the oscillator model. If we allow for free electrons as well then we get instead

$$\varepsilon_r = \varepsilon_r' + i\varepsilon_r'' = 1 + \frac{Nq_\mathrm{e}^2}{\varepsilon_0 m_\mathrm{e}}\left[\frac{f_\mathrm{e}}{-\omega^2 - i\gamma_\mathrm{e}} + \sum_j \frac{f_j}{\omega_{0j}^2 - \omega^2 - i\gamma_j\omega}\right]$$

The second term corresponds to the bound electrons summing over all relevant natural frequencies and the first term is the same but for unbound electrons so $\omega_0 = 0$. The numerators, f_e and f_j give the fraction of electrons that fall into each of these categories.

The addition of the extra term changes the optical behaviour of the materials. For transparent dielectrics, such as glass, the natural frequency is $\omega_0 \sim 10^{16}\,\mathrm{rad\,s^{-1}}$, which corresponds to the UV part of the spectrum. This means that visible light has $\omega < \omega_0$ which we saw corresponds to the electron cloud oscillating π out of phase with $\boldsymbol{E}$ and so the dipoles oscillate in phase with $\boldsymbol{E}$. For free electrons the frequency of the driving force is above ω_0, even for visible light, which means that the oscillating electrons radiate wavelets that cancel the incident wave.

Ignoring the bound electron's contribution and neglecting damping, $\gamma_\mathrm{e} = 0$, the relative permittivity is given by

$$\varepsilon_r = 1 - \frac{Nq_\mathrm{e}^2}{\varepsilon_0 m_\mathrm{e}\omega^2} = 1 - \frac{\omega_p^2}{\omega^2}$$

where this equation defines ω_p, called the **plasma frequency**. For $\omega < \omega_p$ we have $\varepsilon_r < 0$ and hence the refractive index is imaginary, $n = i\sqrt{\omega_p^2/\omega^2 - 1}$. The reflectivity at normal incidence can then be shown to be 1. If instead $\omega > \omega_p$ then $\varepsilon_r > 0$ and $n = \sqrt{1 - \omega_p^2/\omega^2} \in \mathbb{R}$. The reflectivity falls rapidly to zero as ω increases.

While these results only apply exactly for free electrons (a plasma) lots of metals can be reasonably approximated as a plasma of free electrons and some bound electrons and these results still approximately apply. This means that metals are transparent to light with $\omega > \omega_p$. For most metals ω_p is in the deep-UV part of the spectrum.

The upper atmosphere is subject to intense solar radiation and is therefore highly ionised. Calculating ω_p based on the charge density of the ionosphere we predict ω_p to be in the MHz range meaning that lower frequency electromagnetic waves should be reflected back down to Earth. This is indeed the case and we use it to send radio messages from one place to another without direct line of site.

Part VI

Superposition

30 Superposition of Waves with the Same Wave Vector

So far we have only considered one wave at a time. If we want to consider more complicated phenomena such as multiple reflections or interference we need a solid mathematical understanding of what happens when two waves occupy the same region of space. We need superposition.

The simplest case is two waves with the same wave vector and frequency. These waves differ only in their amplitude, E_{0i}, and phase, Φ_i. For simplicity we define our coordinate system such that the common wave vector, k, is aligned with the z-axis. Then each wave has the functional form

$$E_i = E_{0i}\cos(kz - \omega t - \Phi_i).$$

The resulting field then has amplitude $E = E_1 + E_2$ given by

$$\begin{aligned} E &= E_{01}\cos(kz - \omega t - \Phi_1) + E_{02}\cos(kz - \omega t - \Phi_2) \\ &= E_{01}[\cos(kz - \omega t)\cos\Phi_1 + \sin(kz - \omega t)\sin\Phi_1] + E_{02}[\cos(kz - \omega t)\cos\Phi_2 + \sin(kz - \omega t)\sin\Phi_2] \\ &= [E_{01}\cos\Phi_1 + E_{02}\cos\Phi_2]\cos(kz - \omega t) + [E_{01}\sin\Phi_1 + E_{02}\sin\Phi_2]\sin(kz - \omega t). \end{aligned}$$

So we see that the resultant field, E, is also a harmonic wave,

$$E = E_0\cos(kz - \omega t - \Phi)$$

where

$$E_0\cos\Phi = E_{01}\cos\Phi_1 + E_{02}\cos\Phi_2, \qquad \text{and} \qquad E_0\sin\Phi = E_{01}\sin\Phi_1 + E_{02}\sin\Phi_2$$

which can be achieved by setting

$$E_0^2 = E_{01}^2 + E_{02}^2 + 2E_{01}E_{02}\cos(\Phi_1 - \Phi_2), \qquad \text{and} \qquad \Phi = \arctan\left[\frac{E_{01}\sin\Phi_1 + E_{02}\sin\Phi_2}{E_{01}\cos\Phi_1 + E_{02}\cos\Phi_2}\right]. \tag{30.1}$$

See figure 30.1.

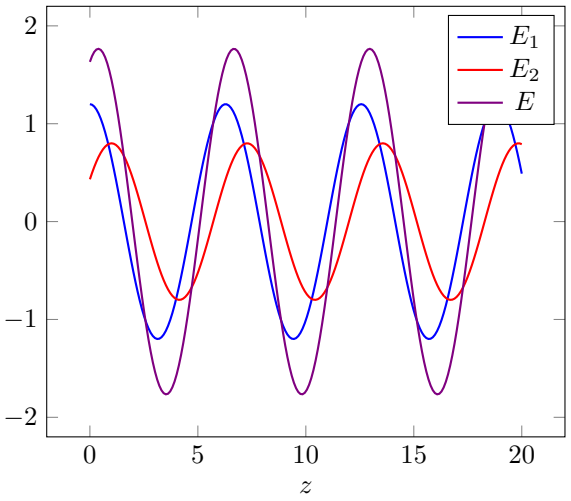


Figure 30.1: A snapshot at some fixed time t showing two harmonic waves, E_i, with the same wave vector but different amplitudes and phases and their superposition, E. Notice that E is a harmonic wave.

Now that we have seen the superposition of two harmonic waves with the same frequency and wave vector the superposition of N harmonic waves all with the same frequency and wave vector is simply given by

$$E = \sum_{i=1}^{N} E_{0i}\cos(kz - \omega t - \Phi_i) = E_0\cos(kz - \omega t - \Phi).$$

30.1 Alternative Computations

We computed the superposition using lots of trig identities. This is perhaps the slowest and least enlightening way to compute superpositions. The most efficient way is probably using complex exponentials and simply discarding the imaginary part when we need the actual physical wave. The starting waves are then given by

$$E_i = E_{0i}\exp[i(kz - \omega t - \Phi_i)] = E_{01}\exp(-i\Phi_i)\exp[i(kz - \omega t)].$$

The superposition of two waves is then

$$E = E_1 + E_2 = E_0\exp(-i\Phi)\exp[i(kz - \omega t)]$$

where E_0 and Φ are given in equation 30.1.

The same computation can also be done graphically with phasors which are simply vectors in the complex plane (viewed as a two dimensional real vector space) and we then simply perform normal vector addition to get the amplitude and phase which are simply the magnitude and angle of the resulting vector. See figure 30.2

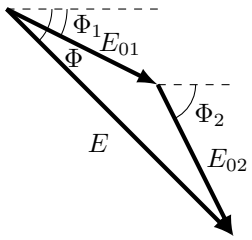


Figure 30.2: Addition of two phasors.

30.2 Intensities

The intensity of the field is proportional to the amplitude squared. In fact $I = v\varepsilon E_0^2/2$ however it is common to just work with $I = E_0^2$ when we only want to compare intensities within a single medium. From equation 30.1 we see that

$$E_0^2 = E_{01}^2 + E_{02}^2 + 2E_{01}E_{02}^2 \cos(\Phi_1 - \Phi_2).$$

So the total intensity is the sum of the intensities plus an interference term which may be positive or negative depending on how in phase the waves are.

The maximum (resp. minimum) intensity is achieved when the phase difference $\Phi_1 - \Phi_2$ is an even (resp. odd) multiple of π and we have

$$E_{0\,\text{min/max}}^2 = E_{01}^2 + E_{02}^2 \pm 2E_{01}E_{02} = (E_{01} \pm E_{02})^2.$$

So for the special case of $E_{01} = E_{02}$ it is possible that the intensity of the resulting wave will be anywhere from 0 to $4E_{01}$.

30.3 Phases

There are a few mechanisms by which a phase difference can appear:

- The first crests of the two waves leave the emitters at different times by a non-integer multiple of the period.
- The two emitters are a non-integer number of wavelengths apart.
- The two waves start in phase but then one passes through a more optically dense material.

To be rigorous about the third of these we introduce the idea of an **phase thickness** which is defined to be

$$\delta = \frac{2\pi d}{\lambda}$$

where d is the actual thickness of the material and λ is the wavelength of the light in the medium. In terms of the vacuum wavelength, λ_0, we have

$$\delta = \frac{2\pi nd}{\lambda_0}.$$

This quantity is related to the optical thickness, nd, but the phase thickness accounts for the phase difference between light that travels through the medium and the same light if it had been unimpeded. If the difference in phase thickness is caused purely by a difference in refractive indices then it is sometimes referred to as **retardation**.

30.4 Coherence

We have so far assumed that the phase difference, $\Phi_1 - \Phi_2$, is constant. We will see later that this isn't always the case and we will introduce notions of coherence time and coherence length which are scales over which the phase difference is approximately constant.

31 More Superposition

31.1 Standing Waves

Suppose two waves of the same frequency, ω, are travelling in opposite directions with different amplitudes. The resultant field has amplitude

$$E(z,t) = E_{0L}\cos(kz - \omega t) + E_{0R}\cos(kz + \omega t).$$

If $E_{0L} = E_{0R}$ then some basic trig identities lead us to conclude that

$$E(z,t) = 2E_{0L}\cos(kz)\cos(\omega t).$$

This is different to waves we have studied so far in that it is not of the form $f(z \pm vt)$ or $f(kz \pm \omega t)$. Instead we have oscillations based on position, z, and oscillations with time, t. All points in the wave rise and fall at the same time. This is a **standing wave**.

It can be shown that for a standing wave the magnetic field is

$$B(z,t) = \frac{E(z',t')}{c}$$

where $z' = z - \lambda/4$ and $t' = t - T/4$ which corresponds to $\boldsymbol{B}$ and $\boldsymbol{E}$ being perpendicular but put of phase by $\pi/2$. An important quantity for standing waves is the Poynting vector, $\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{B}$. This calculation is complicated by the fact that $\boldsymbol{B} \neq \boldsymbol{E}/c$ which we have relied on in the past. It can be shown that

$$\boldsymbol{S} = c\varepsilon_0 E_0^2 \sin(2kz)\sin(2\omega t)\boldsymbol{e_z}.$$

So $\boldsymbol{S}$ is always along the z direction but its amplitude and sign oscillate in time and space in such a way that

$$I = \langle S \rangle = 0.$$

So for a standing wave there is no flow of energy. There is movement of energy but it moves forward and back so the net result is no energy flow.

31.2 Superposition with Similar Frequencies

Consider two waves that are *almost* the same. For example these waves may be produced by the same source and have the same amplitude, E_{01}, but slightly different wave vectors and frequencies. These waves are given by

$$E_1 = E_{01}\cos(k_i z - \omega_i t).$$

We can use the trig identity

$$\cos\alpha + \cos\beta = 2\cos\left(\frac{1}{2}(\alpha+\beta)\right)\cos\left(\frac{1}{2}(\alpha-\beta)\right).$$

The superposition of these waves is then

$$E = 2E_{01}\cos\left[\left(\frac{k_1+k_2}{2}z - \left(\frac{\omega_1-\omega_2}{2}\right)\right)\right]\cos\left[\left(\frac{k_1-k_2}{2}\right)z - \left(\frac{\omega_1-\omega_2}{2}\right)t\right].$$

Notice that this appears to be the product of two different waves. We call the first the **carrier** wave and it has frequency and wave vector

$$\omega_c = \frac{\omega_1+\omega_2}{2}, \qquad \text{and} \qquad k_c = \frac{k_1+k_2}{2}.$$

The second is called the **modulation** wave and it has frequency and wave vector

$$\omega_m = \frac{\omega_1 - \omega_2}{2}, \qquad \text{and} \qquad k_m = \frac{k_1 - k_2}{2}.$$

So the resulting wave is

$$E = 2E_{01}\cos(k_c z - \omega_c t)\cos(k_m z - \omega_t).$$

Since $\omega_1 \approx \omega_2$ and $k_1 \approx k_2$ we have that $\omega_c \gg \omega_m$ and $k_c \gg k_m$. The result is a wave with approximate frequency and wave vector ω_c and k_c which is attenuated by a wave with frequency ω_m and wave vector k_m. This can be seen in figure 31.1.

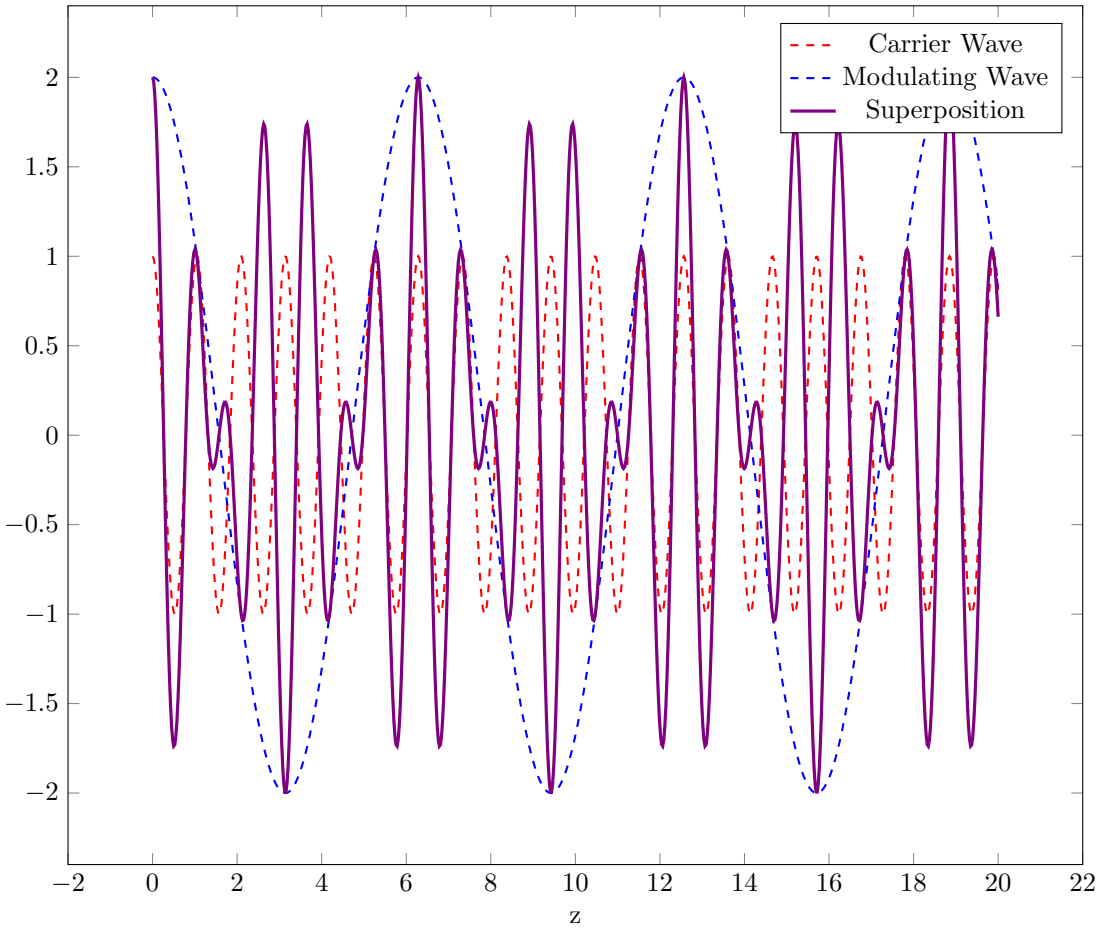


Figure 31.1: The superposition of two waves with similar frequency and wave vector results in a carrier wave attenuated by a lower frequency wave.

The intensity of this wave is found by averaging $|\boldsymbol{S}|$ over one period of the carrier wave. This gives

$$I \propto 4E_{01}^2\cos^2(k_m z - \omega_m t) = 2E_{01}^2[1 + \cos(2(k_m z - \omega_m t))].$$

We see that the intensity varies from 0 to $4E_{01}^2$ with frequency $2\omega_m$, which is known as the beats frequency. We can actually hear the beats frequency if we choose two notes close in frequency but differing by a few hertz.

31.2.1 Group and Phase Velocity

The dispersion of electromagnetic waves in media leads to different phase velocities at different frequencies. Notice we introduce the modifier here 'phase' as we will see there is a second velocity of interest in this scenario.

The modulation wave moves with velocity v_m given by the dispersion relationship:

$$v_m = \frac{\omega_m}{k_m} = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\Delta\omega}{\Delta k}.$$

Assuming that $\omega_c \gg \omega_m$, i.e. that $\omega_1 \approx \omega_2$ and similarly $k_c \gg k_m$ this ratio of differences can be approximated with a derivative:

$$v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k}$$

where v_g is known as the **group velocity** and is the velocity of the envelope (modulation wave). We can relate the group velocity and the phase velocity (which up until this point we have simply referred to as velocity) by noting that $\omega = ck/n$ and also refractive index depends on frequency so $n = n(\omega)$ which means that

$$\omega = \frac{ck}{n(\omega)} = v_p(\omega)k.$$

Differentiating this and rearranging we get

$$v_g = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{1}{n + \omega \mathrm{d}n/\mathrm{d}\omega} = \frac{v_p}{1 - (\omega/v_p)\mathrm{d}v_p/\mathrm{d}\omega}.$$

One interesting consideration here is that it is possible that v_p can be greater than c. This seems to disagree with special relativity but it turns out that no information can be transmitted at the phase velocity, the only way information can be sent in this case is in the resulting wave which travels at the group velocity, which is always lower than c. So special relativity isn't violated.

32 Fourier Analysis

Up until now in this part we have taken two (or more) waves and computed the resulting field. In this section we will look at the resulting wave and attempt to recover the individual waves. Assuming that the only waves present are harmonic we can do this using Fourier analysis.

Consider some periodic wave, f, which is not necessarily harmonic. By periodic we mean there exists some $\lambda \in \mathbb{R}$ such that $f(z) = f(z + \lambda)$. Then we can decompose f (under some fairly week conditions) as

$$f(z) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos(mkz) + \sum_{m=1}^{\infty} \sin(mkz)$$

where $k = 2\pi/\lambda$, and A_m and B_m are real constants given by

$$A_m = \frac{2}{\lambda}\int_0^{\lambda} f(z)\cos(mkz)\,\mathrm{d}z\,, \qquad \text{and} \qquad B_m = \frac{2}{\lambda}\int_0^{\lambda} f(z)\sin(mkz)\,\mathrm{d}z\,.$$

For a non-periodic wave we can consider it to be a periodic function which repeats at infinity. A proper limiting process[15] gives us

$$f(z) = \frac{1}{\pi}\left[\int_0^{\infty} A(k)\cos(kz)\,\mathrm{d}k + \int_0^{\infty} B(k)\sin(kz)\,\mathrm{d}k\right]$$

where A and B are now functions of $k \in \mathbb{R}$

Example 32.1. Consider the function f, defined by

$$f(z) = \begin{cases} 1/L, & \text{if } z \in [-L/2, L/2], \\ 0, & \text{else.} \end{cases}$$

The Fourier transform of this is

$$\begin{aligned} \hat{f}(k) &= \mathcal{F}[f](k) \\ &= \int_{-\infty}^{\infty} f(z)e^{-ikz}\,\mathrm{d}z \\ &= \frac{1}{L}\int_{-L/2}^{L/2} [\cos(kz) + i\sin(kz)]\,\mathrm{d}z \end{aligned}$$

[15] See the Fourier analysis part of the Fourier analysis and statistics course

$$
\begin{aligned}
&= \frac{1}{Lk}[\sin(kz)]_{-L/2}^{L/2} \\
&= \frac{1}{Lk}\sin\left(\frac{kL}{2}\right) - \frac{1}{Lk}\sin\left(-\frac{kL}{2}\right) \\
&= \frac{2}{Lk}\sin\left(\frac{kL}{2}\right) \\
&= \operatorname{sinc}\left(\frac{kL}{2}\right)
\end{aligned}
$$

where sinc is the un-normalised sinc function defined as

$$\operatorname{sinc} x = \begin{cases} \frac{\sin x}{x}, & x \neq 0, \\ 1, & x = 0. \end{cases}$$

The untransformed function, f, has width L and the transformed function, $\hat{f}$, has width $4\pi/L$ (width here being the distance between the first two zeros) so the width of f is inversely proportional to the width of $\hat{f}$. This is known as the reciprocal relation and is common in Fourier analysis.

Example 32.2. Another important example of a Fourier transform is a Gaussian with variance $\sigma^2 = 2L$ and mean $\mu = 0$. The Fourier transform is again a Gaussian with $\mu = 0$ but now $\sigma^2 = 2\pi/L$, which again shows the reciprocal relation.

Example 32.3. The Dirac delta distribution is defined by

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0, \end{cases}$$

and[16]

$$\int_a^b \delta(x)\,\mathrm{d}x = 1_{(a,b)}(0) = \begin{cases} 1, & 0 \in (a,b), \\ 0, & 0 \notin (a,b). \end{cases}$$

It has the property that

$$\int_a^b f(x)\delta(x-x')\,\mathrm{d}x = \begin{cases} f(x'), & x' \in (a,b), \\ 0, & x' \notin (a,b). \end{cases}$$

This sifting property makes it very easy to compute the Fourier transform:

$$
\begin{aligned}
\hat{\delta}(k) &= \mathcal{F}[\delta](x - x') \\
&= \int_{-\infty}^{\infty} \delta(x - x')e^{-ikx}\,\mathrm{d}x \\
&= e^{ikx'}.
\end{aligned}
$$

Notice that δ has zero width and $\hat{\delta}$ has infinite width.

Some useful properties of the Fourier transform that can simplify computations are

- If f is an even function then $\mathcal{F}[f]$ is real.
- If f is an odd function then $\mathcal{F}[f]$ is imaginary.
- The Fourier transform of the Fourier transform of f is the mirror image of f. This means $\mathcal{F}\{f(z)\} = 2\pi\,\mathcal{F}^{-1}\{f(-z)\}$.
- $\mathcal{F}$ is a linear operator so $\mathcal{F}\{af(z) + bg(z)\} = a\,\mathcal{F}\{f(z)\} + b\,\mathcal{F}\{g(z)\}$.
- If f is wide then $\mathcal{F}\{f\}$ is narrow. In particular if f has width L then $\mathcal{F}\{f\}$ has width proportional to $1/L$.

[16] Here 1_X is the indicator function defined by $1_X(x) = 1$ if $x \in X$ and $1_X(x) = 0$ if $x \notin X$.

- Convolution in real space becomes multiplication in Fourier space:

$$h(z) = (f * g)(z) = \int_{-\infty}^{\infty} f(x - x')g(x')\,\mathrm{d}x' \iff h(z) = \mathcal{F}^{-1}\{\mathcal{F}[f]\,\mathcal{F}[g]\}(z).$$

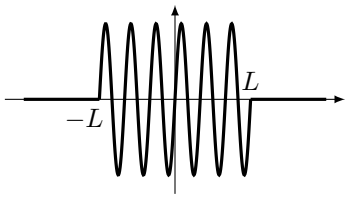
The linearity property and the inverse Fourier transform property allow us to identify

$$\mathcal{F}\{\cos(k_0 z)\} = \delta(k - k_0) + \delta(k + k_0).$$

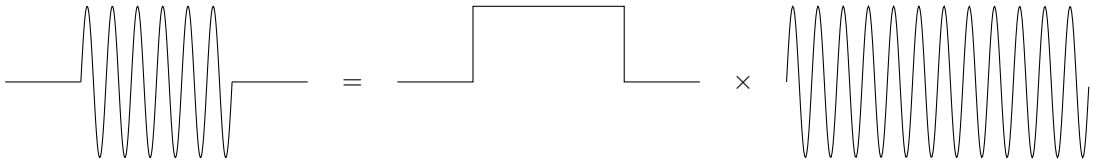
We will use this later.

32.1 Wave Packets

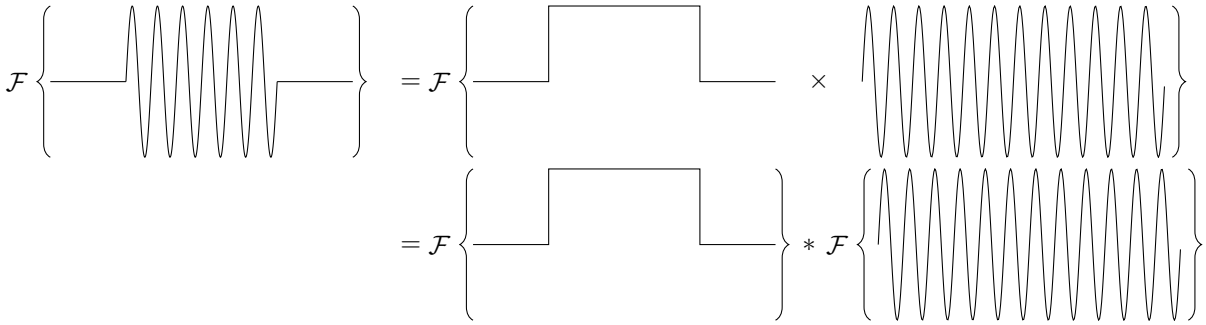
Until now we have mostly considered harmonic waves that extend forever in time and space. This is unrealistic but good enough in some cases. A more realistic case would be a wave that starts and finishes at some fixed time. A snapshot of this wave may look like



We can consider this to be the product of a top hat and a harmonic wave:



The Fourier transform of this can then be computed using the convolution theorem:



We can define units such that the wave starts at $-L$ and ends at L. We see that in the time it exists there are 6 full cycles of the wave. The wave length is $\lambda_0 = 2L/6$. Hence the wave vector is $k_0 = 2\pi/\lambda_0 = 6\pi/L$. Working in units of L we have

$$\begin{aligned}\hat{f}(k) &= \mathrm{sinc}(k) * [\delta(k - k_0) + \delta(k + k_0)] \\ &= \int_{-\infty}^{\infty} \mathrm{sinc}(k - k')\delta(k' - k_0)\,\mathrm{d}k' + \int_{-\infty}^{\infty} \mathrm{sinc}(k - k')\delta(k' + k_0)\,\mathrm{d}k' \\ &= \mathrm{sinc}(k - k_0) + \mathrm{sinc}(k + k_0).\end{aligned}$$

We can similarly show that for a Gaussian wave packet, that is a wave of the form

$$\cos(k_0 z)\exp\left[-z^2/(2\sigma^2)\right]$$

the Fourier transform is, up to a constant factor,

$$\exp\left[-k^2\sigma^2/2\right].$$

The important thing about both of these wave packets is that the Fourier transforms are *not* delta distributions. This means that there is more than one frequency present. We conclude that a wave of finite extent cannot be monochromatic. Instead it has many wave vectors with the most important being those near $k = k_0$ since this is where sinc/the Gaussian peaks.

32.2 Band Width

The following result follows from the reciprocal relationship:

$$\Delta x \Delta k \propto L\frac{1}{L} \propto 1.$$

The more general relationship is

$$\Delta x \Delta k \geq 1$$

with equality only for Gaussians. This relationship appears all over the place in physics. For example the uncertainty principle is essentially the same with momentum, $p = \hbar k$ replacing k and applied to the probability density, $|\psi|^2$. This is a consequence of the wavelike nature of matter.

In the last section we considered a wave with finite extent in space. The exact same analysis applies to a wave with finite extent in time replacing x with t and k with ω so we have

$$\Delta t \Delta \omega \geq 1.$$

The value Δt is known as the **coherence time** and $\Delta x = c\Delta t$ is the **coherence length**. These quantities give the time scales and length scales over which the phase relationship between two waves is approximately constant. For larger time periods/lengths the phase relationship is essentially random.

32.3 Group Velocity Again

Consider a wave packet, $\psi(z,t)$, with Fourier transform at time $t = 0$ given by Ψ so

$$\psi(z,0) = \overset{-1}{\mathcal{F}}\{\Psi(k)\} = \int_{-\infty}^{\infty} \Psi(k) e^{-ikz}\,\mathrm{d}k\,.$$

At some later time all components of the wave will have propagated at the phase velocity $\omega(k)/k$ and so

$$\psi(z,t) = \int_{-\infty}^{\infty} \Psi(k) e^{-i(kz-\omega t)}\,\mathrm{d}k\,.$$

As long as the wave packets all have wave vectors close to k_0 we can approximate $\omega(k)$ to first order by

$$\omega(k) \approx \omega(k_0) + (k-k_0)\left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_{k_0} = \omega_0 + (k-k_0)v_g.$$

Hence we have

$$\begin{aligned}\psi(z,t) &= e^{-i(v_g k_0 - \omega_0)t}\int_{-\infty}^{\infty} \Psi(k) e^{-i(kz - k v_g t)}\,\mathrm{d}k \\ &= e^{-i(v_g k_0 - \omega_0)t}\psi(z - v_g t, 0).\end{aligned}$$

So we see that wave packets with wave vectors centred on k_0 travel at the group velocity, v_g.

Part VII

Polarisation

33 Polarisation

So far we have considered superposition of waves travelling along the z-axis with $\boldsymbol{E}$ aligned and $\boldsymbol{B}$ aligned an perpendicular to $\boldsymbol{E}$. This isn't always the case and to generalise we need to look at polarised light.

33.1 Linear Polarisation

Consider the two following waves

$$\boldsymbol{E_1} = \boldsymbol{e_x} E_{0x} \cos(kz - \omega t - \Phi_1), \qquad \text{and} \qquad \boldsymbol{E_2} = \boldsymbol{e_y} E_{0y} \cos(kz - \omega t - \Phi_2).$$

The first of these is x-polarised as the electric field oscillates only along the x-axis, the second is y-polarised. We can write these as complex exponentials discarding the imaginary parts for the final values:

$$\boldsymbol{E_1} = \boldsymbol{e_x} E_{0x} \exp[i(kz - \omega t - \Phi_1)], \qquad \text{and} \qquad \boldsymbol{E_2} = \boldsymbol{e_y} E_{0y} \exp[i(kz - \omega t - \Phi_2)].$$

The superposition is then given by the sum of these two vectors:

$$\boldsymbol{E} = \boldsymbol{E_1} + \boldsymbol{E_2} = \boldsymbol{e_x} E_{0x} \exp[i(kz - \omega t - \Phi_1)] + \boldsymbol{e_y} E_{0y} \exp[i(kz - \omega t - \Phi_2)].$$

We can write these in terms of column vectors with the first component giving the x component and the second the y component:

$$\boldsymbol{E} = \begin{pmatrix} E_{0x} \exp[i(kz - \omega t - \Phi_1)] \\ E_{0y} \exp[i(kz - \omega t - \Phi_2)] \end{pmatrix} = \begin{pmatrix} E_{0x} e^{-i\Phi_1} \\ E_{0y} e^{-i\Phi_2} \end{pmatrix} \exp[i(kz - \omega t)].$$

Notice that we separate the oscillatory term into the exponential outside the vector and the vector is constant in time. This vector is called the **Jones vector** for $\boldsymbol{E}$. The Jones vector is simply the complex amplitudes of the components of the wave in the x and y directions.

33.1.1 No Phase Shift

Suppose there is no relative phase shift between the two components. That is $\Phi_1 = \Phi_2 = \Phi$, then

$$\boldsymbol{E} = \begin{pmatrix} E_{0x} \\ E_{0y} \end{pmatrix} e^{-i\Phi} \exp[i(kz - \omega t)].$$

We can then identify $\boldsymbol{E}$ as another linearly polarised wave with amplitude

$$E_0 = \sqrt{E_{0x}^2 + E_{0y}^2}$$

which is polarised at angle

$$\vartheta = \arctan\left(\frac{E_{0y}}{E_{0x}}\right)$$

to the x-axis.

33.1.2 Circularly Polarised Light

Suppose the two waves have a constant phase difference of $\pi/2$ so $\Phi_1 - \Phi_2 = \pi/2$. Suppose also that $E_{0x} = E_{0y} = E_{01}$. Then

$$\boldsymbol{E} = \begin{pmatrix} E_{01} e^{-i(\Phi_2 + \pi/2)} \\ E_{01} e^{-i\Phi_2} \end{pmatrix} \exp[i(kz - \omega t)] = \begin{pmatrix} -i \\ 1 \end{pmatrix} e^{-i\Phi_2} \exp[i(kz - \omega t)].$$

It is fairly easy to show that as z increases $(E_x(z), E_y(z))$ maps out a clockwise circle completing one circle each time z increases by one wavelength. Similarly $(E_x(t), E_y(t))$ maps out an anticlockwise circle with angular frequency ω. This is a circularly polarised state. We call this **left hand circularly polarised**, or $\mathcal{L}$. Similarly if $\Phi_1 - \Phi_2 = -\pi/2$ then we have Jones vector

$$\begin{pmatrix} i \\ 1 \end{pmatrix}$$

and the circles are in the opposite direction. We call this **right hand circularly polarised**, or $\mathcal{R}$.

33.1.3 Elliptically Polarised

The most general case has no relationship between amplitudes or phases. Then $\boldsymbol{E}$ both rotates and changes magnitude in the (x, y)-plane tracing out an ellipse as it does so competing one rotation every wave length with angular frequency ω. We call this **elliptically polarised light**.

33.2 Normalisation

As is often the case with vectors we wish to work with orthonormal vectors as a basis. For this we need a well defined inner product. Since the polarisation vectors are two dimensional and contain complex values we can identify them with $\mathbb{C}^2$ and the correct inner product is then

$$\boldsymbol{a}\cdot\boldsymbol{b}=\boldsymbol{a}^\dagger\boldsymbol{b}=\begin{pmatrix}a_1\\a_2\end{pmatrix}^\dagger\begin{pmatrix}b_1\\b_2\end{pmatrix}=\begin{pmatrix}a_1^* & a_2^*\end{pmatrix}\begin{pmatrix}b_1\\b_2\end{pmatrix}=a_1^*b_1+a_2^*b_2.$$

In particular we say that $\{\boldsymbol{e_i}\}$ are orthonormal if $\boldsymbol{e_i}\cdot\boldsymbol{e_j}=\delta_{ij}$.

Clearly the vectors denoting x and y polarised light,

$$\boldsymbol{e_x}=\begin{pmatrix}1\\0\end{pmatrix},\qquad\text{and}\qquad\boldsymbol{e_y}=\begin{pmatrix}0\\1\end{pmatrix}$$

are orthonormal. The vectors that we found for describing circularly polarised light are orthogonal and can be normalised by a factor of $1/\sqrt{2}$:

$$\boldsymbol{e_\mathcal{R}}=\frac{\sqrt{2}}{2}\begin{pmatrix}i\\1\end{pmatrix},\qquad\text{and}\qquad\boldsymbol{e_\mathcal{L}}=\frac{\sqrt{2}}{2}\begin{pmatrix}i\\-1\end{pmatrix}.\tag{33.1}$$

In order to be a proper basis a set must span the space while being linearly independent. This is known to be true for $\{\boldsymbol{e_x},\boldsymbol{e_y}\}$ and is easy to show for $\{\boldsymbol{e_\mathcal{R}},\boldsymbol{e_\mathcal{L}}\}$ since $\boldsymbol{e_\mathcal{R}}+\boldsymbol{e_\mathcal{L}}=i\sqrt{2}\boldsymbol{e_x}$ and $\boldsymbol{e_\mathcal{R}}-\boldsymbol{e_\mathcal{L}}=i\sqrt{2}\boldsymbol{e_y}$ so as long as $\{\boldsymbol{e_x},\boldsymbol{e_y}\}$ spans the space $\{\boldsymbol{e_\mathcal{R}},\boldsymbol{e_\mathcal{L}}\}$ spans the space too meaning both are orthonormal bases. Which basis we use depends on the types of polarisation present.

33.3 Intensity

The electric field can be written as

$$\boldsymbol{E}=\begin{pmatrix}E_{0x}e^{-i\Phi_1}\\E_{0y}e^{-i\Phi_2}\end{pmatrix}\exp[i(kz-\omega t)]=\boldsymbol{E_0}\exp[i(kz-\omega t)].$$

The intensity is given by the square of the magnitude of the electric field. Since we deal with the complex amplitude here this is

$$I=\frac{1}{2}c\varepsilon_0E_0|E_0|^2=\frac{1}{2}c\varepsilon_0\boldsymbol{E_0}\cdot\boldsymbol{E_0}=\frac{1}{2}c\varepsilon_0E_0^*E_0=\frac{1}{2}c\varepsilon_0(E_{0x}^2+E_{0y}^2).$$

34 Polarisers

Any individual wave has a polarisation however most light is formed of many different waves which are incoherent and, on average, have no net polarisation. For this reason we need a way to produce polarised light.

34.1 Linear Polarisers

A **linear polariser** (often just called a polariser) will only allow light polarised along a fixed axis to pass through. Natural (unpolarised) light has the form

$$\begin{pmatrix}E_x(z,t)\\E_y(z,t)\end{pmatrix}=\begin{pmatrix}E_{0x}(t)\cos[kz-\omega t-\Phi_x(t)]\\E_{0y}(t)\cos[kz-\omega t-\Phi_y(t)]\end{pmatrix}.$$

In general the two components are only coherent over short time scales, less than the coherence time, Δt. Over time scales larger than this the ratio of the amplitudes, E_{0x}/E_{0y}, and the relative phase, $\Phi_y-\Phi_x$, will be unpredictable. If this light passes through a polariser then this forces a correlation between the amplitudes and phases and therefore the resulting light will be coherent and the output from the polarised behaves like an ideal harmonic wave. For this reason it is common to use a polariser as the first step in an optics experiment.

34.2 Malus's Law

Consider light from one polariser passing through a second polariser. If the axes of polarisation are aligned then we expect that all light from the first polariser will pass through the second. If the axes of polarisation are perpendicular then we expect the second polariser to block all of the light. What happens if the axes are not parallel or perpendicular?

Suppose that the first polariser produces light, $\boldsymbol{E}_{\mathbf{in}}$, and that the axis of the second polariser, $\hat{\boldsymbol{p}}$, is at angle ϑ to the polarisation of the incoming light. The light transmitted through the second polariser, $\boldsymbol{E}_{\mathbf{out}}$, is the component of $\boldsymbol{E}_{\mathbf{in}}$ in the direction of $\hat{\boldsymbol{p}}$, and is polarised in the direction $\hat{\boldsymbol{p}}$. That is

$$\boldsymbol{E}_{\mathbf{out}} = (\boldsymbol{E}_{\mathbf{in}} \cdot \hat{\boldsymbol{p}})\hat{\boldsymbol{p}} = E_{\text{in}} \cos(\vartheta)\hat{\boldsymbol{p}}.$$

The intensity of the light coming from the second polariser is then

$$I_{\text{out}}(\vartheta) = \frac{1}{2} c\varepsilon_0 E_{\text{in}}^2 \cos^2\vartheta = I_{\text{in}} \cos^2\vartheta.$$

So as the second polariser is rotated we expect the intensity of the output to go as $\cos^2\vartheta$. This is **Malus's law**. Notice that in a whole rotation there are two maxima, $\vartheta = 0, \pi$, which correspond to the polarisers being aligned and anti-aligned, and two minima, $\vartheta = \pi/2, 3\pi/2$, which correspond to the two times the polarisers are perpendicular.

34.3 Jones Matrices

We managed to find the intensity in the last section by a logical argument. However there is a much more powerful mathematical framework that allows us to work with arbitrary combinations of polarisers. This framework, called Jones matrices, assigns to each polariser a matrix which then acts on the Jones vector of the incoming light. A linear polariser in the x direction has the Jones matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

since

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} E_x \\ 0 \end{pmatrix},$$

so we are left with only the x component. Similarly a linear polariser in the y direction has Jones matrix

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

From here we can construct the matrix for a polariser at arbitrary angle ϑ by realising that this is really just a rotated version of an x aligned polariser. This is best demonstrated by considering the same example of incoming polarised light, which without loss of generality we assume to be x polarised, and a polariser at angle ϑ. Then

$$\begin{aligned} \boldsymbol{E}_{\mathbf{out}} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} R(\vartheta) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} R(\vartheta) \begin{pmatrix} \cos\vartheta \\ \sin\vartheta \end{pmatrix} \\ &= \begin{pmatrix} \cos\vartheta \\ 0 \end{pmatrix}. \end{aligned}$$

From this we easily recover Malus's law for the intensity.

This worked well for this simple case. For a more complicated case it would be better to have each polariser have a single matrix and not have to worry about rotating as we go. To do this we define a frame of reference for each individual polariser in which the x direction aligns with the polarisation axis. In the case of incoming polarised light have the un-primed frame have x be aligned with the initial

polarisation. In this frame the input light has the Jones vector $\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$. We can be more general and assume that the incoming light has Jones vector $\boldsymbol{u}$ and the output is light with Jones vector $\boldsymbol{v}$. Let J be the Jones matrix of the polariser. Then we have $\boldsymbol{v} = J\boldsymbol{u}$. This must also hold if we express all terms in the frame of the polariser, which is simply a rotated frame. Denote by a prime the same vectors and matrices but in this rotated frame. Then we have $\boldsymbol{v'} = J'\boldsymbol{u'}$. We can convert between the two frames with a rotation matrix so $\boldsymbol{u} = R(\vartheta)\boldsymbol{u'}$ and $\boldsymbol{v} = R(\vartheta)\boldsymbol{v'}$. Noting that the inverse of a rotation is simply a rotation by the same amount in the opposite direction we have

$$\boldsymbol{v'} = \mathbb{1}\boldsymbol{v'} = R(-\vartheta)R(\vartheta)\boldsymbol{v'} = R(-\vartheta)\boldsymbol{v} = R(-\vartheta)J\boldsymbol{u} = R(-\vartheta)JR(\vartheta)\boldsymbol{u'} = J'\boldsymbol{u'}$$

so we identify

$$R(-\vartheta)JR(\vartheta) = J'.$$

This is just the normal rule for transforming matrices between bases.

We can repeat our calculation of the output but now in the primed frame:

$$\begin{aligned}\boldsymbol{E'_{out}} &= R(-\vartheta)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}R(\vartheta)\begin{pmatrix}1\\0\end{pmatrix}\\ &= \begin{pmatrix}\cos^2\vartheta & -\cos\vartheta\sin\vartheta\\ -\cos\vartheta\sin\vartheta & \sin^2\vartheta\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ &= \begin{pmatrix}\cos^2\vartheta\\ -\cos\vartheta\sin\vartheta\end{pmatrix}.\end{aligned} \tag{34.1}$$

This gives the same $\cos^2\vartheta$ relationship for the output intensity.

The transformations can be easily remembered with a commutative diagram[17]:

$$\begin{array}{ccc}\boldsymbol{u} & \xrightarrow{J} & \boldsymbol{v}\\ {\scriptstyle R}\uparrow & & \uparrow{\scriptstyle R}\\ \boldsymbol{u'} & \xrightarrow[J']{} & \boldsymbol{v'}\end{array}$$

To read this diagram simply pick the starting point and end point and find a path between them. Then traversing this path every arrow gives multiplication by the corresponding matrix if going in the direction of the arrow or the inverse of that matrix if going against the arrow.

35 More Polarisers

In a vacuum there is nothing to distinguish between different polarisations. Therefore anything that causes polarisation must have some level of anisotropy, it must preferentially select for one direction over the others. We have seen that result of light incident at an oblique angle (i.e. non-normal incidence) depends on the polarisation state and this can be used to create polarised light. This happens even if the medium is isotropic, it is the geometry of the experiment that introduces the anisotropy. We have been considering LIH materials so far. In this section we relax this and consider materials with some anisotropy (LH materials).

35.1 Wire Grid Polariser

The simplest polariser is a collection of wires running parallel with only a small amount of spacing between wires (small as usual compared to the wavelength). The gap can be on a visible scale if we are interested in microwaves which have wavelengths of a few centimetres. Oscillations of the electric field along the wires will lead to electrons moving along the wire. In a real metal this movement is damped by the resistance reducing the intensity of the field oscillating along that direction. Thus the field oscillating perpendicular to the wires is the field that passes through. This is somewhat counter-intuitive if we consider the slots between wires to be like holes one may expect the field oscillating across the wires to be blocked but this isn't the case. The property of strongly absorbing one polarisation and allowing another to pass is called **dichroism** and the effect in this case is to produce linearly polarised light so these are a form of **dichroic linear polaiser**.

[17] the 'commutative' part of this diagram is simply that any path between the same two points gives the same result. Matrix multiplication is non-commutative so the order arrows are traversed is important.

35.2 Polaroid

Polaroid works similarly to a wire grid polariser but on an atomic scale. They are made by creating sheets of polymer and then stretching in one direction introducing anisotropy as the long chained molecules are aligned. The material is then treated with iodine which binds to the molecules and fixes them in place. Polaroids can be made to work well across the entire visible spectrum.

The quality of a polariser is given by its **extinction ratio**. This is defined as the ratio $R = I_0/I_L$ where I_0 is the intensity of light incident on polariser orthogonally polarised to the transmission axis and I_L is the intensity of light that it lets through which should have been blocked. For polaroid a typical value is $R = 100$. For research a much higher value is needed.

35.3 Polarisation by Reflection

If a glass block is illuminated at Brewster's angle with unpolarised light then there is no reflection of P-polarised light and so the reflected light will be entirely S-polarised. For technical reasons it is often more convenient to work with the transmitted beam which is not entirely polarised but has far more P-polarised light than the incident beam.

35.4 Polarisation by Scattering

Recall that when a dipole absorbs light it re-emits it at the same frequency and the intensity is not spherically symmetrical. No light is emitted along the dipole vector and the highest intensity is in a plane normal to the dipole axis. Thus light which is scattered in this plane at a right angle to the incoming light will be scattered with oscillation only in the plane as oscillation perpendicular to the plane and along the dipole is not allowed.

The intensity of such scattered light is

$$I = \frac{\omega^4 p_0^2 \sin^2 \vartheta}{32\pi^2 \varepsilon_0 r^2 c^3}.$$

The factor of ω^4 means that blue light is scattered far more than red light. This is the reason that the sky is blue. Light from the sun, which is approximately white, is scattered when it passes through the atmosphere with blue being scattered far from the sun (so into the rest of the sky) and red light passing straight through without much scattering making the sun appear slightly red/yellow. This is called **Rayleigh scattering**.

36 Birefringence

Birefringence is the property by which the refractive index depends on the polarisation. A birefringent material is necessarily anisotropic. In this case the permittivity is a tensor but some simple cases can still be considered without needing tensors which is what we will do here.

36.1 Uniaxial Birefringence

A material is uniaxially birefringent if there exists a basis in which the permittivity tensor, ε, is diagonal and $\varepsilon_x = \varepsilon_y \neq \varepsilon_z$. Thus a ray travelling along the optical axis (that is the z-axis) will behave the same for any polarisation. If n_x (resp. n_y) is the refractive index for light polarised in the x direction (resp. y direction) then for uniaxial birefringence we have $n_x = n_y$.

If we consider Huygen's principle then each wavelet that sets out travels in all directions. Since the refractive index depends on the direction of polarisation (for travel not along the z-axis) this means that instead of spherical wave fronts we get ellipsoids but the net wave front still progresses forwards as normal.

For propagation perpendicular to the optical axis we instead get two wave fronts, one for light that with polarisation along the optical axis and one for light with polarisation perpendicular to the optical axis. Let n_o be the refractive index for light polarised perpendicular the optical axis and n_e be the refractive index for light with polarisation parallel to the optical axis. The o stands for ordinary as this wave propagates as we would expect for any material and the e stands for extraordinary. Let $\Delta n = n_e - n_o$. If $\Delta n > 0$ then we say that the material is **positive uniaxial** and if $\Delta n < 0$ we say it is **negative**

uniaxial. The extraordinary thing about waves polarised parallel to the optical axis is that the wave vector, $\boldsymbol{k}$, which gives the direction of propagation of the wave, doesn't necessarily align with the ray direction which is given by the Poynting vector, $\boldsymbol{S}$.

36.2 Biaxial Birefringence

A biaxial crystal has a different refractive index along each axis and is therefore characterised by three separate refractive indices, n_α, n_β, and n_γ. The biaxial part here is due to the fact that such crystals actually have two optical axis along which the refractive index is independent of polarisation (as opposed to the single axis of a uniaxial crystal).

36.3 Birefringence from the Oscillator Model

The oscillator model assumes that only the magnitude of the displacement of the electron cloud is important for the response of the system. We can picture this as the electron cloud being held in place by six springs (two along each axis) all of which are the same strength. If instead we take some of the springs to be of different strengths then the direction of displacement matters. This means that the natural frequency of oscillations will depend on the direction of displacement and therefore the polarisation of the electric field. This also explains dichroism as the different springs sill absorb energy at different rates and therefore the energy absorbed also depends on the polarisation.

36.4 Retarders

Consider linearly polarised light incident on a uniaxial material but such that the optical axis is not necessarily parallel or perpendicular to the incident polarisation. Suppose also that this light is monochromatic with frequency ω and wavelength in air, λ_{air}. Further assume that the material is transparent in the relevant region of the EM spectrum and we can ignore the small fraction of light that is reflected.

When the light enters the medium the component parallel to the optical axis experiences refractive index n_e and travels at speed c/n_e whereas the component perpendicular to the optical axis experiences refractive index n_o and travels at speed c/n_o. One component will therefore travel faster than the other. Which component depends on the material since $\Delta n = n_e - n_o$ can be positive or negative so from now on we will refer to a fast axis and a slow axis.

Suppose that the medium is a slab with the incident face and back parallel. Then the light exits at normal incidence also. Since $\lambda = \lambda_{\text{air}}/n$ the number of wavelengths that the light will complete while in the material will be different. This means that if the two components are in phase when they enter the material they will be out of phase when they leave. This introduces a time delay to the slow ray which, when expressed in units of phase angle, is called the **retardation** of the slow wave, δ. As such uniaxial material arranged in slabs like this are called retarders or retardation plates.

The effect of a retarder is that the slow wave is retarded in time but advanced in space (in that the same point on the fast and slow waves will occur further away for the slow wave). The effect is that the the slow wave picks up a phase factor of $e^{i\delta}$. If we take the x-axis to be aligned with the fast axis then the Jones matrix of a retarder is

$$J = \begin{pmatrix} e^{i\delta_{\text{fast}}} & 0 \\ 0 & e^{i\delta_{\text{slow}}} \end{pmatrix}$$

since both waves are retarded somewhat. If the slab has physical thickness d then it has optical thickness nd and phase thickness $2\pi nd/\lambda_{\text{air}}$ and we have $\delta_{\text{fast}} = 2\pi n_{\text{fast}} d/\lambda_{\text{air}}$ and $\delta_{\text{slow}} = 2\pi n_{\text{slow}} d/\lambda_{\text{air}}$. It is often possible to use the freedom in deciding the global phase to put all of the effect in the slow direction and so

$$J = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\delta} \end{pmatrix}$$

where

$$\delta = \delta_{\text{slow}} - \delta_{\text{fast}} = \frac{2\pi d}{\lambda_{\text{air}}}(n_{\text{slow}} - n_{\text{fast}}) > 0.$$

So δ is the retardation of the slow wave relative to the fast wave.

36.4.1 Real Retarders

Suppose we want a retarder that results in a phase difference of π, that is half a cycle, then we need the physical thickness to be

$$d = \frac{\lambda_{\text{air}}/2}{n_{\text{slow}} - n_{\text{fast}}}.$$

Unfortunately this is often prohibitively thin. Fortunately if its only the phase difference we care about then there is no difference between a phase difference of π or 3π, or indeed $(2n+1)\pi$ for any $n \in \mathbb{Z}$. So we can simply add any integer number of wavelengths to the optical path difference and it results in the same phase difference.

Another problem is that in general refractive index depends on frequency, and this is still the case with retarders. In fact Δn depends on frequency which means that the phase difference achieved by a retarder depends on the frequency of the incident light. We will come back to this later.

36.5 Retarder Uses

36.5.1 Half-Wave Plate

A half-wave plate (HWP) is a retarder that results in a phase difference of π. If the fast axis is aligned with the x-axis then the Jones matrix is

$$J_{\lambda/2} = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\pi} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The result of this is to reverse the sign of the E_y component. For elliptically polarised light the HWP reverses its handedness and rotates the ellipse so that if its major axis was at angle φ before the HWP it will be at $-\varphi$ afterwards (so a rotation of -2φ).

36.5.2 Quarter-Wave Plate

Similarly a quarter-wave plate (QWP) results in a phase difference of $\pi/2$ which means it has the Jones matrix

$$J_{\lambda/4} = \begin{pmatrix} 1 & 0 \\ 0 & e^{-i\pi} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}.$$

For the special case where $E_x = \pm E_y$ if the beam is at an angle of $\pm\pi/4$ to the fast axis then the result is circularly polarised light.

36.5.3 Crystal Polarisers

Birefringent materials can be used to create a polariser. These all work off of the basic principle that different polarisations are refracted a different amount. This means that incident light is split into two beams when it enters the material and the two beams have orthogonal polarisations. To create a polarised source simply arrange the set up such that upon reaching the other side of the medium one of the beams is at an angle greater than the critical angle and so is reflected back into the material whereas the other beam passes out of the material for use in the experiment. Note that the critical angle also depends on the polarisation since it depends on the refractive index.

37 More Jones Algebra

One of the useful things about using Jones matrices to describe the effect of a polarisers and retarders is that the effect of multiple polarisers and retarders can be combined into one Jones matrix by simply multiplying all relevant Jones matrices and possibly changing basis if the polarisation/fast axes aren't aligned.

Example 37.1. Consider three polarisers aligned such that in the lab frame the first has its polarisation axis aligned along the x-axis, the second has its polarisation axis at an angle, ϑ, to the first, and the third is aligned along the y-axis. The Jones matrices for the first and third polarisers are simple:

$$J_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \text{and} \qquad J_3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Jones matrix for the second polariser is found by rotating the Jones matrix to align with the others which is done with the matrix in equation 34.1 but reversing the sign of ϑ since in that equation we are changing the basis and in this we wish to rotate the matrix (passive vs. active transform):

$$J_2 = \begin{pmatrix} \cos^2\vartheta & \cos\vartheta\sin\vartheta \\ \cos\vartheta\sin\vartheta & \sin^2\vartheta \end{pmatrix}$$

Therefore the matrix describing the effects of these three polarisers is

$$J = J_3J_2J_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos^2\vartheta & \cos\vartheta\sin\vartheta \\ \cos\vartheta\sin\vartheta & \sin^2\vartheta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos^2\vartheta & 0 \\ \cos\vartheta\sin\vartheta & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \cos\vartheta\sin\vartheta & 0 \end{pmatrix}.$$

So for unpolarised incident light the polarisation after these three polarisers is

$$\begin{pmatrix} 0 & 0 \\ \cos\vartheta\sin\vartheta & 0 \end{pmatrix}\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} 0 \\ E_x\cos\vartheta\sin\vartheta \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 0 \\ E_x\sin(2\vartheta) \end{pmatrix}.$$

So the result is, y-polarised light, which it must be given the last polariser is aligned with the y-axis. The intensity of said light is

$$I(\vartheta) = \frac{I_0}{4}\sin^2(2\vartheta)$$

where I_0 is the intensity of the incident light. The minimum transmission is 0 which occurs when any two neighbouring polarisers are orthogonal. The maximum transmission is 25 % which occurs when the middle polariser is directly between the other two, so $\vartheta = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$.

Example 37.2. Consider the same set up as the previous example but replace the middle polariser with a retarder with its fast axis at angle ϑ. The Jones matrix for this retarder is

$$\begin{aligned} J_2 &= \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & e^{i\delta} \end{pmatrix}\begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \\ &= \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}\begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -e^{i\delta}\sin\vartheta & e^{i\delta}\cos\vartheta \end{pmatrix} \\ &= \begin{pmatrix} \cos^2\vartheta + e^{i\delta}\sin^2\vartheta & \cos\vartheta\sin\vartheta(1-e^{i\delta}) \\ \cos\vartheta\sin\vartheta(1-e^{i\delta}) & e^{i\delta}\cos^2\sin\vartheta + \sin^2\vartheta \end{pmatrix} \end{aligned}$$

We could now find the Jones matrix representing the three different polarisers/retarders but instead we will make use of the fact that the first is an x aligned polariser so that the Jones vector after the first polariser is

$$\boldsymbol{E} = \begin{pmatrix} \sqrt{I_0/2} \\ 0 \end{pmatrix}$$

where we have used the fact that the light is unpolarised so $E_x = E_y$ and $E^2 = E_x^2 + E_y^2 = 2E_x^2 = I_0$. Hence the electric field that we get out the other side of the set up is

$$\begin{aligned} J_3J_2\begin{pmatrix} 1 \\ 0 \end{pmatrix}\sqrt{I_0/2} &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos^2\vartheta + e^{i\delta}\sin^2\vartheta & \cos\vartheta\sin\vartheta(1-e^{i\delta}) \\ \cos\vartheta\sin\vartheta(1-e^{i\delta}) & e^{i\delta}\cos^2\sin\vartheta + \sin^2\vartheta \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix}\sqrt{I_0/2} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos^2\vartheta + e^{i\delta}\sin^2\vartheta \\ \cos\vartheta\sin\vartheta(1-e^{i\delta}) \end{pmatrix}\sqrt{I_0/2} \\ &= \begin{pmatrix} 0 \\ \cos\vartheta\sin\vartheta(1-e^{i\delta}) \end{pmatrix}\sqrt{I_0/2} \\ &= \frac{\sqrt{2}}{4}I_0\begin{pmatrix} 0 \\ \sin(2\vartheta)(1-e^{i\delta}) \end{pmatrix}. \end{aligned}$$

So this time the result is y-polarised light with intensity

$$\begin{aligned} I(\vartheta) &= \frac{1}{8}I_0\sin^2(2\vartheta)(1-e^{i\delta})(1-e^{-i\delta}) \\ &= \frac{1}{8}I_0\sin^2(2\vartheta)(1-e^{i\delta}-e^{-i\delta}+1) \end{aligned}$$

$$= \frac{1}{8} I_0 \sin^2(2\vartheta)(2 - 2\cos\delta)$$
$$= \frac{1}{4} I_0 \sin^2(2\vartheta)(1 - \cos\delta).$$

To see the effect of changing the retarder lets fix $\vartheta = \pi/4$ and consider what happens as δ changes. Suppose that $d = 1\,\text{mm}$ and $\Delta n = 1.25 \times 10^{-3}$. If the system is illuminated with white light then the light that we get out will be blue-green. This is because for $\lambda \approx 500\,\text{nm}$ we have $\delta \approx \pi$ and so for blue-green light the retarder acts as a HWP. The effect of this is to rotate the polarisation by $\pi/2$ which means that this light passes through the final polariser unimpeded.

The effect described in the previous example by which crossed polarisers can result in colour can be used to analyse transparent media for stresses since internal stresses can rotate polarisations (as stresses are anisotropic).

37.1 Eigenpolarisations

As the name suggests the **eigenpolarisations** of a device are the polarisations which are not rotated by the polariser. That is if the Jones matrix of the polariser is J and $\boldsymbol{u}$ is an eigenpolarisation then

$$J\boldsymbol{u} = \lambda\boldsymbol{u}$$

where λ is some constant[18]. Finding eigenpolarisations is as simple as finding the eigenvectors of a 2×2 matrix.

Example 37.3. *Find the eigenpolarisations of a polariser aligned with the x-axis.*

The Jones matrix for this polariser is

$$J = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

First we need to find the eigenvalues by solving for the roots of the characteristic polynomial:

$$0 = \begin{vmatrix} 1-\lambda & 0 \\ 0 & -\lambda \end{vmatrix} = (1-\lambda)(-\lambda) \implies \lambda_{1,2} = 0, 1.$$

From here its a simple case of substituting these into the eigenvalue equation and solving for $\boldsymbol{u}$:

$$\boldsymbol{u_1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{u_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Notice that these are simply the two orthogonal linear polarisations that we use as a basis.

Example 37.4. *Find the eigenpolarisations of a retarder.*

The Jones matrix for a retarder is

$$J = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\delta} \end{pmatrix}.$$

First we need to find the eigenvalues by solving for the roots of the characteristic polynomial:

$$0 = \begin{vmatrix} 1-\lambda & 0 \\ 0 & e^{i\delta}-\lambda \end{vmatrix} = (1-\lambda)(e^{i\delta}-\lambda) \implies \lambda_{1,2} = e^{i\delta}, 1.$$

From here its a simple case of substituting these into the eigenvalue equation and solving for $\boldsymbol{u}$:

$$\boldsymbol{u_1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{u_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

So the eigenpolarisations are the same as for the polariser but with different eigenvalues.

---

[18] λ is not the wavelength here.

37.1.1 Diagonalisation

We can diagonalise the Jones matrix once we have found its eigenvectors and eigenvalues. To do this we use a matrix, T, which has the eigenvectors of J as its columns and the result is that $J' = T^{-1}JT$ is a diagonal matrix with the eigenvalues of J along its diagonal in the same order that the corresponding eigenvectors appear in T.

37.2 Circular Systems

Chiral molecules are ones which have exactly the same constituents but are mirror images of each other. A solution of these molecules treats all linear polarisations the same as a solution is necessarily isotropic. However the handedness of these molecules means that they don't treat circular polarisations equally.

We can consider a circular polariser which has eigenvectors $\boldsymbol{e_\mathcal{L}}$ and $\boldsymbol{e_\mathcal{R}}$. Using the circular basis, which we consider to be the primed basis, the Jones matrix of this device is of the form

$$J' = \begin{pmatrix} t_\mathcal{L} & 0 \\ 0 & t_\mathcal{R} \end{pmatrix}.$$

In order to combine the effects of this circular polariser with linear polarisers and retarders we need to express this in the linear polariser basis. To do this we need to find the transformation matrix T which has as its columns the vectors representing $\boldsymbol{e_\mathcal{L}}$ and $\boldsymbol{e_\mathcal{R}}$ in the linear polariser basis. These vectors are given in equation 33.1 and so we have

$$T = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}.$$

It is then trivial to show that

$$T^{-1} = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

Hence in the linear polariser basis

$$\begin{aligned} J &= TJ'T^{-1} \\ &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}\begin{pmatrix} t_\mathcal{L} & 0 \\ 0 & t_\mathcal{R} \end{pmatrix}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} \\ &= \begin{pmatrix} \bar{t} & i\Delta/2 \\ -i\Delta/2 & \bar{t} \end{pmatrix} \end{aligned}$$

where $\bar{t} = (t_\mathcal{L} + t_\mathcal{R})/2$ and $\Delta = t_\mathcal{R} - t_\mathcal{L}$. Notice here that we use $J = TJ'T^{-1}$ since we are inverting the transform $J' = T^{-1}JT$.

38 Reflections and Other Polarisation Effects

So far we have considered optical devices which transmit normally incident light. We can also consider non-normal incidence and reflective devices with the same Jones algebra. One complication is that we then need two different coordinate systems, one for the incident beam and one for the reflected beam. We use the convention that for any beam we define the x direction as the direction of oscillation of P-polarised light and the y-direction of oscillation of S-polarised light.

The simplest case of a reflection occurs at normal incidence. At normal incidence the polarisation is not important and $|r_\mathrm{S}| = |r_\mathrm{P}| = |r|$. The Jones matrix describing a reflection at normal incidence is then

$$J = |r|\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Notice the factor of -1 for y-polarised light (S-polarised light). This is because of the π phase shift that S-polarised light undergoes upon reflection. Ignoring the change in direction and assuming $|r| \approx 1$ we see that the effect of a reflection is essentially the same as the effect of a HWP. Notice that both of these reverse the handedness of the incident light, which is what we would expect from a mirror.

One interesting use of this is to construct an *anti*-reflection device. First polarise the light at $\pi/4$ to the x-axis. Then pass the light through a QWP which results in $\mathcal{L}$-circularly polarised light. Then reflect this light which results in $\mathcal{R}$-circularly polarised light. This then passes back through the QWP resulting in light polarised at $-\pi/4$ to the x axis which is then completely blocked by the polariser.

38.1 Polarimetry

Polarimetry is the general term describing experiments to determine the polarisation of a beam of light. For simplicity here we will assume that this light is indeed polarised and discus a few ways to find out what this polarisation is. It is possible to extend our work so far to partially polarised light but we then need a four dimensional analogue of Jones algebra known as Stokes–Mueller calculus.

Suppose that we have a polariser, a QWP, and an intensity detector and we wish to determine the polarisation of a beam of monochromatic polarised light. Passing the beam through the polariser and measuring the transmitted intensity as the polariser is rotated there are three possible outcomes:

- Two maxima are observed with the full intensity of the incident beam and two minima are observed with intensity 0. This is the behaviour predicted by Malus' law for linearly polarised light polarised such at the same angle at which the maxima occur.
- The intensity is half the incident intensity and doesn't vary with the polariser angle. This is due to circularly polarised light. This can be confirmed by inserting the QWP before the polariser which will turn the circularly polarised light into linearly polarised light which we can then check using the first bullet point.
- Two maxima and two minima are observed but the intensity is never zero. This corresponds to elliptically polarised light and is simply the combination of the two previous cases. To find the size and orientation of the major axis of the ellipse pass the beam through the QWP and polariser. There will be a particular angle of the QWP at which the light produced is linearly polarised. We can use the polariser to check for this with Malus' law. From here we can then compute the orientation and ellipticity of the polarisation.

Part VIII

Interference

39 Superposition Again

Consider two beams of light from two point sources, S_1 and S_2, which are separated by some distance a, propagating in an LIH medium. We assume that the two sources have the same angular frequency, ω, and that $a \gg \lambda$. At some point P, far from the two sources we effectively have plane waves:

$$\boldsymbol{E_i} = \boldsymbol{E_{0i}} \cos(\boldsymbol{k_i} \cdot \boldsymbol{r} - \boldsymbol{\omega t} - \boldsymbol{\Phi_i}).$$

The net field at this point is

$$\boldsymbol{E} = \boldsymbol{E_1} + \boldsymbol{E_2}.$$

As usual E oscillates far too quickly to be a useful quantity to measure. Instead we are interested in the intensity,

$$I = \varepsilon v \langle E^2 \rangle.$$

Here $\varepsilon = \varepsilon_0 \varepsilon_r$ and v is the speed of light in this medium. If the intensity of each source at this point is I_i then we find that the intensity of the combined light is

$$\begin{aligned} I &= \varepsilon v \langle E_1^2 + E_2^2 + 2\boldsymbol{E_1} \cdot \boldsymbol{E_2} \rangle \\ &= I_1 + I_2 + 2\varepsilon v \langle \boldsymbol{E_1} \cdot \boldsymbol{E_2} \rangle \\ &= I_1 + I_2 + 2\varepsilon v \boldsymbol{E_{01}} \cdot \boldsymbol{E_{02}} \langle \cos(\boldsymbol{k_1} \cdot \boldsymbol{r} - \omega t - \Phi_1) \cos(\boldsymbol{k_2} \cdot \boldsymbol{r} - \omega t - \Phi_2) \rangle. \end{aligned}$$

So the resulting intensity is the sum of the two intensities plus an interference term, I_{12}. We can evaluate this term by noticing that $\langle \sin^2 \vartheta \rangle = \langle \cos^2 \vartheta \rangle = 1/2$ and $\langle \sin \vartheta \cos \vartheta \rangle = 0$ so

$$I_{12} = 2\varepsilon v \boldsymbol{E_{01}} \cdot \boldsymbol{E_{02}} \cos \delta$$

where

$$\delta = (\boldsymbol{k_1} - \boldsymbol{k_2}) \cdot \boldsymbol{r} - \Phi_1 + \Phi_2.$$

It follows from this that if $\boldsymbol{E_{01}}$ and $\boldsymbol{E_{02}}$ are orthogonal then the interference term vanishes and $I = I_1 + I_2$ whereas if $\boldsymbol{E_{01}}$ and $\boldsymbol{E_{02}}$ are parallel then we have

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta.$$

Then for the specific case of $\delta = 2n\pi$ for $n \in \mathbb{Z}$ we have **total constructive interference** and I is maximised:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}.$$

On the other hand if $\delta = (2n+1)\pi$ then we have **totally destructive interference** and the intensity is minimised:

$$I = I_1 + I_2 - 2\sqrt{I_1 I_2}.$$

Often we will see patterns where we have both constructive and destructive interference, called **fringes**. A useful quantity to define is the **fringe contrast**:

$$C = \frac{\max\{I\} - \min\{I\}}{\max\{I\} + \min\{I\}}.$$

This is easily measured. It can be shown that C is maximised by the case of $I_1 = I_2$.

It is only possible to view fringes if δ is stable. In practice this means that the phase difference, $\Phi_1 - \Phi_2$ has to be stable. This is not the case for everyday light sources which is why we need to set up special lab conditions with lasers and slits in order to start seeing fringes.

At the start of this section we specified that $a \gg \lambda$. This is required for the two sources to act as two separate sources. The result is that I_{12} averages to zero over all space and so the effect of interference is to increase the energy at some points and decrease it at others. If $a \sim \lambda$ then the two sources behave more like one source with amplitude $E_{01} + E_{02}$.

40 Films

When testing Fresnel's equations physicist Lord Rayleigh found that they only held for freshly prepared glass. The conclusion was that some oxidation process lead to a film on the glass which changed the optical properties. Counter to what one may expect this layer of tarnish actually increased the transmission of the glass.

If light travels from medium 1 to medium 3 at normal incidence then we have seen that the reflection coefficient is

$$R_{13} = \left| \frac{n_1 - n_3}{n_1 + n_3} \right|^2.$$

We will now consider what happens if we include a thin layer of medium 2 in between media 1 and 3. For simplicity we will take medium 1 to be air and medium 3 to be glass. We will also assume that the transmission coefficient of the film is high and the reflection coefficient is low. For enhanced transmission it we must have decreased reflection. In fact it is the reflection that is decreased by this scenario and the increased transmission is simply a side effect.

The mechanism by which this occurs is as follows:

- The beam of light enters the film. A small portion is reflected with intensity $I_0 R_{12}$ but most passes through. The intensity of the transmitted beam is $I_0 T_{12}$.
- The beam reaches the glass and a small portion is reflected again. The intensity of this reflection is $I_0 T_{12} R_{13}$.

- This beam reaches the film-air boundary and is transmitted with intensity $I_0 T_{12}^2 R_{13}$ (note that $T_{12} = T_{21}$).

In theory part of this beam reflects back to the glass, then is reflected back to the film-air boundary and part of this is transmitted and so on but for very clear materials the reflectivity is low enough that we can neglect secondary reflections. Yet another approximation we can make is $T_{12} \approx 1$ and so the net intensity of the reflected light is $I_0(R_{12} + R_{23})$. For the case of $n_1 < n_2 < n_3$ the net reflectivity, $R_{12} + R_{23}$, is lower than the reflectivity of the air-glass boundary. Notice that R scales quadratically with Δn so if the film has a refractive index somewhere between the two other mediums then Δn is decreased at each step and since it is squared this has the effect of reducing the overall reflectivity since if $(\Delta n_{13})^2 = (\Delta n_{12} + \Delta n_{23})^2 = \Delta n_{12}^2 + \Delta n_{23}^2 + 2\Delta n_{12}\Delta n_{23} < \Delta n_{12}^2 + \Delta n_{23}^2$ since all Δn have the same sign if $n_1 < n_2 < n_3$.

We can take this to the extreme and use many thin layers all only with slightly higher refractive index than the last and increase transmissivity by a lot. The limiting case is then a medium with smoothly changing refractive index which experiences no reflection. This is exploited in optical fibres to bend light away from the outside rather than having an abrupt reflection. This helps signals stay together as they follow more similar paths.

40.1 Thin Film Interference

In the previous section we assumed that the two reflected beams didn't interfere. This is a reasonable assumption if the thickness of the film is greater than the coherence length of the light. If this isn't the case then the thickness of the film determines the number of wavelengths that the two beams travel before coming together and this in turn determines whether there will be constructive or destructive interference.

Of particular interest is when we can use destructive interest to reduce the reflectivity (and hence increase transmissivity) of a material. If the film has physical thickness d the for light of wavelength λ we define the phase thickness to be $\beta = 2\pi n_2 d/\lambda$. The reflection coefficient for the system is

$$r_{123} \approx r_{12} + e^{i\beta} r_{23} e^{i\beta}.$$

Here r_{12} gives the initial reflection off of the air-film boundary and r_{23} gives the reflection off of the film-glass boundary. There are then two factors of $e^{i\beta}$ as the reflected wave must travel through the film twice (once in each direction). The reason this is only approximate is we are ignoring secondary reflections. To minimise reflection we use a film with refractive index, n_2, somewhere between that of air and glass. The question of interest is what physical thickness of film do we need? To answer this question it is important to note that both reflections occur when in a low refractive index material and reflecting off of a higher refractive index material. This means that both reflections cause a phase shift of π, the actual value of this phase shift is not important, what is important is it is the same for both so actually has no effect on the type of interference. If instead we had n_2 be higher then the refractive index of glass then we would need to consider more carefully the phase shift due to reflection.

Noticing that the equation can be written as

$$r_{123} \approx r_{12} + e^{2i\beta} r_{23}$$

we see that choosing $\beta = \pi/2$ results in destructive interference, so we choose $d = \lambda/(4n_2)$. We now just have to choose a value for n_2 which we can do by expanding the Fresnel coefficients:

$$r_{123} \approx r_{12} - r_{23} = \frac{n_1 - n_2}{n_1 + n_2} - \frac{n_2 - n_3}{n_2 + n_3}$$

which will be zero if $n_2 = \sqrt{n_1 n_3}$. So for an air-glass system taking $n = 1$ for air and $n = 1.5$ for glass we have $n_2 = 1.22$. In reality there is not a material with the required refractive index and other physical attributes needed to form a thin film. It is common to use MgF_2 which has $n = 1.38$. The reflectivity can then be reduced from 4 % to 1.5 %. It is possible to do better if we use multiple layers of thin films and exploit the phase difference that is accrued when reflecting off of a high refractive index material but not a low refractive index material.

40.2 Soap Film

Consider a soap film stretched over a wire frame as shown in figure 40.1. If this frame is vertical then gravity will stretch the soap downwards causing the film to be thicker at the bottom and thinner at the top. In a similar way to the last section we can consider only the first order reflections off of the soap-air and air-soap boundaries. The refractive index of the soap film will be greater than that of air and so the air-soap reflection will have a phase shift of π and the soap-air reflection will not. At the top of the film, the thinnest part if the thickness $d_1 < \lambda$ then the phase thickness is $\delta \approx \pi$ due to the phase shift upon reflection. This results in destructive interference and the top of the film appears black. Further down when the thickness is $d_2 > \lambda$ the phase thickness will be

$$\delta = \frac{4d_2 n}{\lambda_0} + \pi.$$

Here λ_0 is the wavelength of light we are interested in. We have constructive interference at this wavelength when $\delta = 2\pi m$ for $m \in \mathbb{Z}$. As we move down the film and the thickness increases all colours in the spectrum will eventually interfere constructively at some point and so we see the entire spectrum. We then see the entire spectrum again for the next value of m and so on until the thickness of the soap film is greater than the coherence length at which point we just see white.

40.3 Newton's Rings/Newton's Wedge

Newton's rings and Newton's wedge are related phenomena where a slowly increasing gap between two glass components causes a pattern of fringes, see figure 40.2. Newton's wedge uses two flat pieces of glass laid on top of each other with the upper piece being raised slightly on one side such that the gap between the two pieces is wedged shaped see figure 40.2a. Newton's rings uses a planar-convex of radius R with the lens placed convex side down on a flat glass surface so that moving away from the centre results in the space between the lens and the glass increases as you move away fro m the centre of the lens, see figure 40.2b. The result of both of these is that incident light is split, ignoring second order reflections and reflections off of a glass-air boundary there will be two beams that leave the setup.

For Newton's wedge the way that this works is the first beam reflects off of the first air-glass boundary and the second makes it through this boundary and into the small space between the two pieces of glass before reflecting off of the next air-glass boundary. Consider Newton's wedge and in particular pick two adjacent bright fringes along the wedge such that the size of the air gap is d_a at the first point and d_b at the second. The phase thickness is then $\beta_a = 2\pi d_a/\lambda$ at the first point and $\beta_b = 2\pi d_b/\lambda$ at the second. If the distance between the two points is Δx then simple geometry gives us

$$\tan\alpha = \frac{d_b - d_a}{\Delta x} = \frac{\lambda}{2\Delta x} \implies \Delta x = \frac{\lambda}{2\tan\alpha} \approx \frac{\lambda}{2\alpha}.$$

Consider now Newton's rings. In this case we have normal incidence on the top plane of the lens. The two beams that interact this time are the beam that reflects off of the convex face of the lens and the beam that reflects off the glass block below. At a distance r from the centre of the lens the thickness of the air gap is d. Taking the convex side of the lens to be part of a sphere we have that

$$(R-d)^2 + r^2 = R^2 \implies r^2 = 2Rd - d^2 \implies d \approx \frac{r^2}{2R}$$

where the approximation assumes that $d \ll R$. This equates to approximating the spherical lens face as parabolic. The phase shift between the two beams is twice the thickness of this air gap minus π due to the phase change upon reflection at the glass-air boundary.

$$\delta = 2 \cdot 2\pi \frac{d}{\lambda} - \pi = \frac{2\pi r^2}{R\lambda} - \pi.$$

We get bright fringes when $\delta = 2m\pi$ for $\pi \in \mathbb{Z}$ so bright fringes occur at distances

$$r_m = \sqrt{\left(m + \frac{1}{2}\right)\lambda R}.$$

Notice that at the centre, $r = 0$, we get a dark fringe due to the fact that there is no boundary at the centre where the two glass components touch and therefore no reflection.

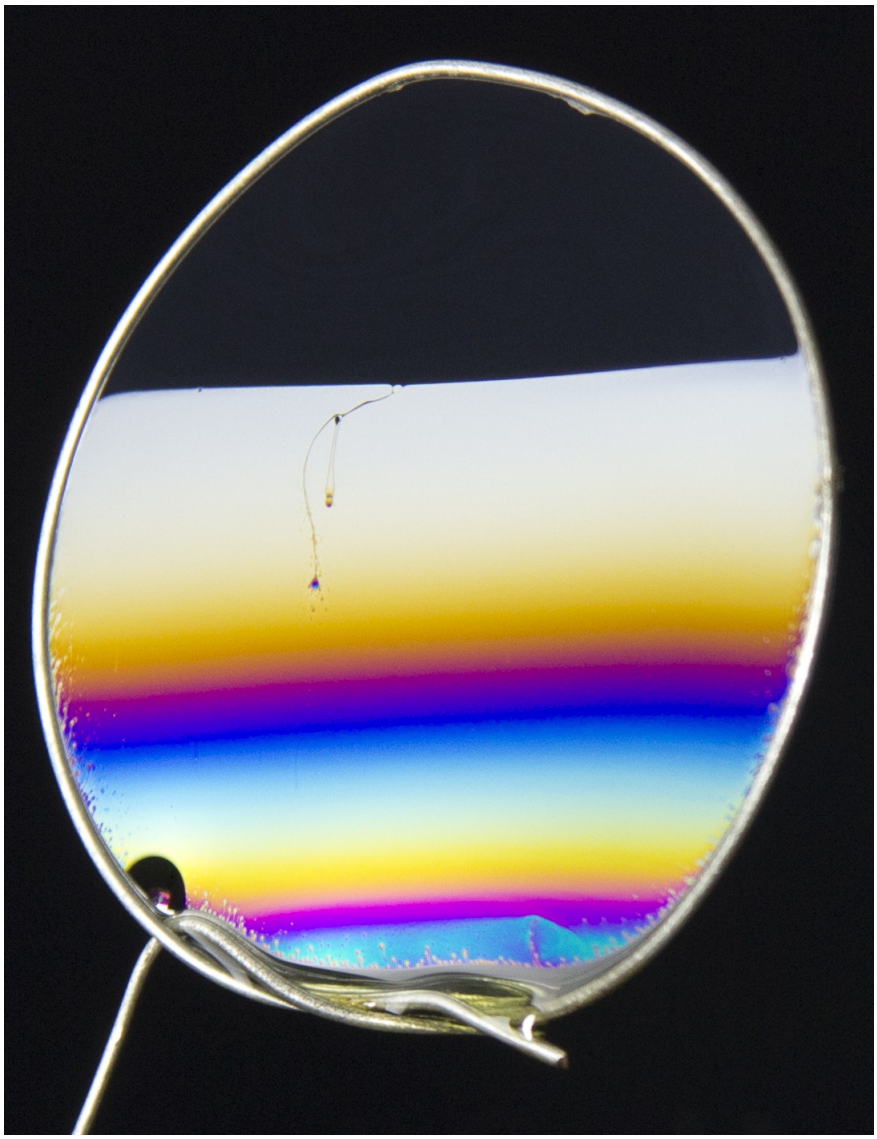
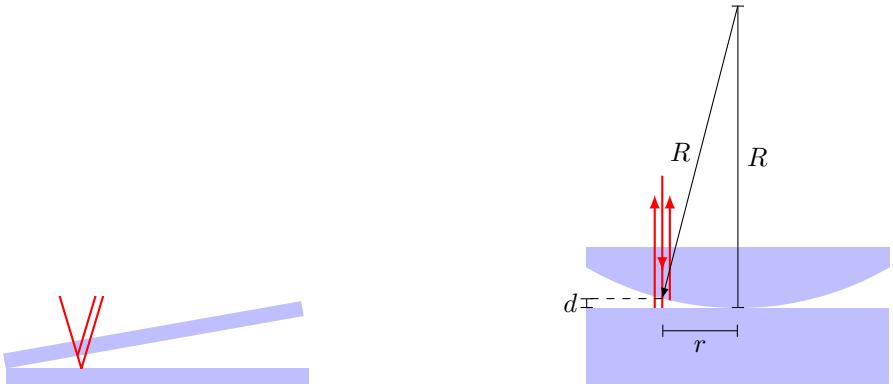


Figure 40.1: A soap film showing destructive interference at the top due to a phase shift upon reflection and a spectrum lower down due to varying thickness. Image credit: `https://www.animations.physics.unsw.edu.au/jw/light/soap-bubbles.htm` accessed on 27/04/2021.



(a) Newton's Wedge (b) Newton's Rings

Figure 40.2: Newton's Wedge and Newton's Rings create a series of bright and dark fringes through a similar mechanism of a gradually increasing air gap.

40.4 Films to Increase/Decrease Reflectivity

By cleverly layering films it is possible to dramatically change the reflectivity of a material. The simplest example involves two layers of film onto glass, first a low refractive index layer and then a high refractive index layer. This way light has to pass from air to low, to high, to glass. Each time the light passes through one of these boundaries the refractive index is relatively higher and therefore the light that is reflected picks up a phase shift of π. By constructing these layers to be the correct thickness it is possible to have destructive interference for the light that is reflected off of the different layers decreasing reflectivity and increasing transmissivity.

Similarly a system of alternating low and high refractive index films can encourage many reflections which, if the thickness of the layers is correct, interfere constructively resulting in high reflectivity which allows us to make a mirror out of a transparent medium.

41 More Thin Film

41.1 Non-Normal Incidence

Consider a bubble. When viewed at the right angle the bubble will appear multicoloured. This is due to thin film effects. The colours also change as the thickness of the bubble is not constant. More interestingly the colours also change based on the angle of viewing which is an effect known as **iridescence** which we will study in this section.

Consider the setup in figure 41.1, a thin transparent film of thickness d and refractive index n_f separating two media of refractive indices n_1 and n_2. Over a small region we can approximate the sides of the film as parallel. Illuminate this film with a monochromatic point source. Since the film is transparent we will ignore secondary reflections. Since we are no longer restricting ourselves to normal incidence the optical path difference between a ray that reflects of of the medium one-film boundary and a ray that reflects of of the film-medium two boundary now depends on the angle at which the ray travels through the film, which is the transition angle. In fact the optical path difference is $2n_f d/\cos\vartheta_t - n_1 AD$ which is the optical path length that the second ray travels through the film while the first ray travels the optical

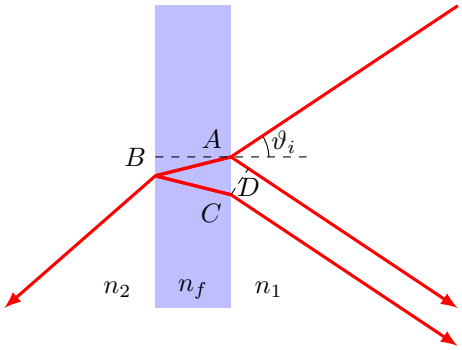


Figure 41.1: Thin film viewed at non-normal incidence.

length AD. For simplicity we take $n_1 = n_2$ and we find that the phase difference between the two rays is

$$\begin{aligned}\delta &= \frac{2\pi}{\lambda_{\text{vac}}}\left[\frac{2n_f}{cos\vartheta_t} - n_1 AD\right] - \pi \\ &= \frac{2\pi}{\lambda_{\text{vac}}}\left[\frac{2n_f d}{\cos\vartheta_t} - n_1 AC\sin\vartheta_1\right] - \pi \\ &= \frac{2\pi}{\lambda_{\text{vac}}}\left[\frac{2n_f d}{\cos\vartheta_t} - n_1 AC\sin\vartheta_t\frac{n_f}{n_i}\right] - \pi \\ &= \frac{2\pi}{\lambda_{\text{vac}}}\left[\frac{2n_f d}{\cos\vartheta_t} - n_f AC\sin\vartheta_t\right] - \pi.\end{aligned}$$

Here we've used Snell's law to get all angles in terms of ϑ_t and we've included a phase shift of π since exactly one of the reflections is off of a higher refractive index medium. A final bit of geometry tells us that $AC = 2d\tan\vartheta_t$ and so

$$\delta = \frac{4\pi n_f}{\lambda_{\text{vac}}} d\cos\vartheta_t - \pi.$$

So there is constructive interference if

$$d\cos\vartheta_t = (2m-1)\frac{\lambda_f}{4}$$

for some $m \in \mathbb{Z}$. Similarly there is destructive interference if

$$d\cos\vartheta_t = 2m\frac{\lambda_f}{4}.$$

We can see that in the case of normal incidence we have $\vartheta_i = \vartheta_t = 0$ and so this reduces to the same relationship that we have for film thickness at normal incidence.

41.1.1 White Light Illumination

We can use Snell's law to write the results of the last section in terms of ϑ_i:

$$d_{\text{constructive}} = \frac{(2m-1)\lambda_{\text{vac}}}{4\sqrt{n_f^2 - n_i^2\sin^2\vartheta_i}}, \qquad \text{and} \qquad d_{\text{destructive}} = \frac{2m\lambda_{\text{vac}}}{4\sqrt{n_f^2 - n_i^2\sin^2\vartheta_i}}.$$

This makes the connection between viewing angle and interference more explicit. Since this also depends on wavelength we see that if we illuminate the film with a white source then at different viewing angles we will see different colours.

This effect is common in nature. It is seen in bird's feathers and seashells, which are formed of many thin layers, as well as in bismuth crystals which quickly builds up a film of oxidised bismuth.

41.2 Multiple Reflections

Up to now we have assumed that all secondary reflections are sufficiently weak that they can be ignored. This isn't the case for a non-transparent material and is only an approximation for a transparent material.

(a) Bismuth. Image credit: ©Micha L. Rieser https://commons.wikimedia.org/wiki/File:Bismuth-crystal.jpg.

(b) Shell. Image credit: https://pixabay.com/photos/pearl-fire-oud-shells-nautilus-sea-1602541/.

(c) Bird. Image credit: https://pxhere.com/en/photo/653981.

Figure 41.2: Some natural examples of thin films causing iridescence.

Consider an air-film-glass setup. The first reflection off of the air-film boundary has reflection coefficient $r_1 = r_{af}$. The second reflection is off of the film-glass boundary and has reflection coefficient

$$r_2 = t_{af} e^{i\beta} r_{fg} e^{i\beta} t_{fa}.$$

The factor of t_{af} accounts for the fact that this ray is first transmitted at the air-film boundary. It then travels through the medium picking up a phase factor of $e^{i\beta}$ ($\beta = 2\pi n_f d \cos\vartheta_t/\lambda$). It is then reflected at the film-glass boundary picking up a factor of r_{fg}, travels back through the film picking up another $e^{i\beta}$ and finally is transmitted at the film-air boundary picking up a factor of t_{fa}. The third reflection follows almost the same path but instead of being transmitted at the film-air boundary it is reflected (r_{fa}), travels back through the film ($e^{i\beta}$), is reflected at the film-glass boundary (r_{fg}), travels back through the film, ($e^{i\beta}$), and is finally transmitted at the film-air boundary (t_{fa}). So the third ray to leave the film into the air has reflection coefficient

$$\begin{aligned} r_3 &= t_{af} e^{i\beta} r_{fg} e^{i\beta} (r_{fa} e^{i\beta} r_{fg} e^{i\beta}) t_{fa} \\ &= r_2 (r_{fa} e^{2i\beta} r_{fg}). \end{aligned}$$

By the same logic we can write the reflection coefficient for the fourth reflected ray as

$$r_4 = r_3 x = r_2 x^2, \qquad \text{where} \qquad x = f_{fa} e^{2i\beta} r_{fg}.$$

Continuing *ad infinitum* we see that the total reflection coefficient is

$$r_{afg} = r_1 + r_2(1 + x + x^2 + \cdots) = r_1 + \frac{r_2}{1-x}.$$

In the last step we have identified the geometric series which converges if and only if $|x| < 1$. It can be shown then that

$$f_{afg} = \frac{r_{af} + r_{fg} e^{2i\beta}}{1 + r_{af}} r_{fg} e^{2i\beta}.$$

Suppose we have an unsupported thin film, that is the film has the same medium on either side (as opposed to being layered onto glass). For simplicity we will also assume that this medium is air. Then

$$f_{afa} = \frac{r_{af} + r_{fa} e^{2i\beta}}{1 + r_{af} r_{fa} e^{2i\beta}} = \frac{r_{af} - r_{af} e^{2i\beta}}{1 - r_{af} r_{af} e^{2i\beta}} = \frac{r[1 - e^{2i\beta}]}{1 - r^2 e^{2-\beta}}$$

where we right $r = r_{af}$ for brevity. The reflectivity of the film is then

$$\begin{aligned} R &= |r_{afa}|^2 \\ &= \frac{r[1 - e^{2i\beta}]}{1 - r^2 e^{2i\beta}} \frac{r[1 - e^{-2i\beta}]}{1 - r^2 e^{-2i\beta}} \\ &= \frac{2r^2[1 - \cos(2\beta)]}{1 - 2r^2\cos(2\beta) + r^4}. \end{aligned}$$

We assumed here that $r \in \mathbb{R}$ which means that $n \in \mathbb{R}$ which means that we are considering only transparent, non-absorbing, materials. We then have that the transmissivity is

$$T = 1 - R = \frac{(1-r)^2}{1 - 2r^2\cos(2\beta) + r^4}.$$

This is minimised when $\cos(2\beta) = -1$ so $4\pi n_f d \cos\vartheta_t/\lambda = (2m-1)\pi$ for $m \in \mathbb{Z}$. We find that

$$\min\{T\} = \frac{(1-r^2)^2}{(1+r^2)^2}$$

and

$$\max\{R\} = 1 - \min\{T\} = \frac{4r^2}{(1+r^2)^2}.$$

We can introduce the **Finesse coefficient**, F, here defined as

$$F = \left(\frac{2r}{1-r^2}\right)^2$$

which allows us to write

$$R = \frac{F\sin^2\beta}{1 + F\sin^2\beta}, \qquad \text{and} \qquad T = \frac{1}{1 + F\sin^2\beta}.$$

For small r we have an approximately $\sin^2$ fringe pattern, just as we observed for two beam interference. For larger values of r however we have strong multiple reflections and a more complex fringe pattern. As $r \to 1$ we have $T \to 0$ and $R \to 1$ unless $\beta = m\pi$ for some $m \in \mathbb{Z}$ in which case we have narrow bands which appear bright viewing with the source on the other side of the film or as dark viewing on the same side as the source. R and T plotted for various values of r, F, and β can be seen in figure **??**.

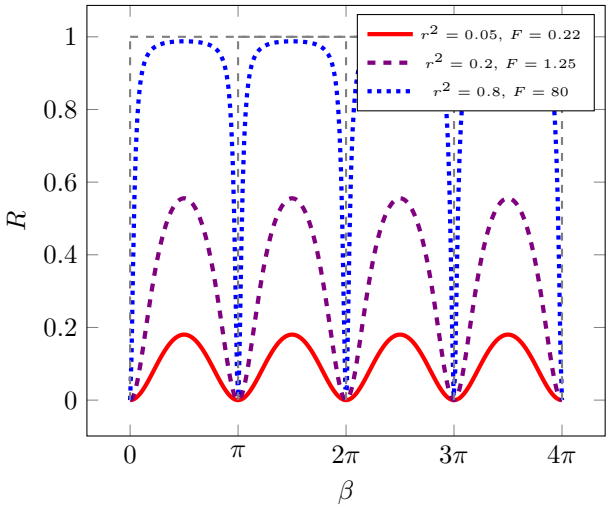
For high values of r or F the transmissivity is very selective for specific values of β, and hence for specific wavelengths. This is exploited in **Fabry-Perot** systems which are comprised of two parallel highly reflective surfaces, usually dielectric mirrors made with multiple layers of thin films. The two mirrors are separated by a distance d and filled with a medium of refractive index n. If the cavity is then illuminated by a collimated beam (that is light where all rays are parallel) of amplitude E_0 which enters at angle ϑ_0 to the normals of the mirrors. If we choose $\vartheta_0 = 0$ then we will get transmission of wavelengths close to

$$\lambda_m = \frac{2nd}{m}$$

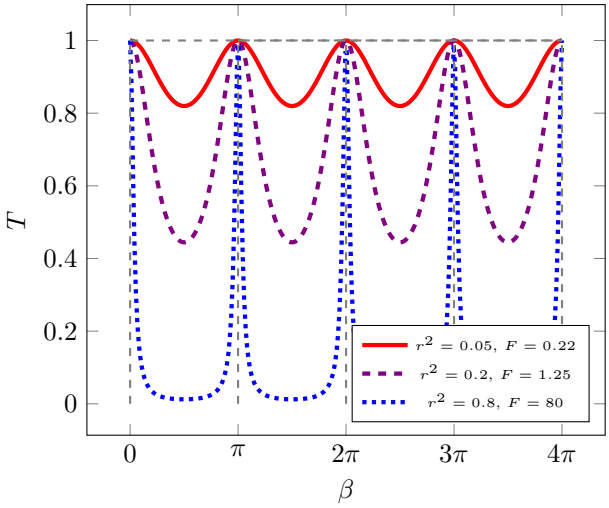
where $m \in \mathbb{Z}$. This corresponds to a half-integer number of wavelengths fitting in the cavity. The result is that only light within a very narrow range around these key values of λ_m can pass out of the cavity so it acts as a very narrow bandpass filter.

Part IX

Diffraction



(a) The reflectivity, R.



(b) The transmissivity, T.

Figure 41.3: The reflectivity and transmissivity, R and T, for an unsupported thin film in air.

42 Basic Diffraction

So far we have considered only homogenous media and abrupt boundaries. If we need to study inhomogeneous media then we need the idea of diffraction. This is the process by which light is partially blocked and how the light propagates past the blockage. We will consider a scalar wave model in this section, which means we will ignore polarisation effects and consider only the amplitude of the wave. We are mostly interested in the intensity, $I = cn\varepsilon_0 E_0^2/2$. However we only really care about relative intensity, which points will be dark and which will be light, so we will often drop the prefactor and take $I = E_0^2$.

42.1 The Huygens–Fresnel Principle

Huygens' principle was discussed in section 26.1. This same principle can be adapted to explain diffraction. One of the problems with Huygens' principle is that it treats all waves the same, regardless of wavelength, which doesn't match reality. For example it is possible to hear something that is happening around the corner but not see it. For some reason the sound waves are able to bend around the corner in a way that light cannot. We also ignored the backwards propagating wavelets when we first considered Huygens' principle and we only drew the wavelets necessary to create the next wavefront.

Fresnel proposed a correction to Huygens' principle that interference of secondary wavelets was what lead to only the forward propagation being important. The amended **Huygen–Fresnel principle** can be stated as:

> Every unobstructed point of a wavefront serves as a source of spherical secondary wavelets having the same frequency and speed as the primary wave. The resulting optical disturbance is given by the superposition of all the secondary waves.

This seems like a fairly superficial change initially, we've simply included the fact that the secondary wavelets are waves and therefore interfere with each other. However, this manages to fix a lot of the problems with Huygens' principle.

42.1.1 Single Slit Diffraction

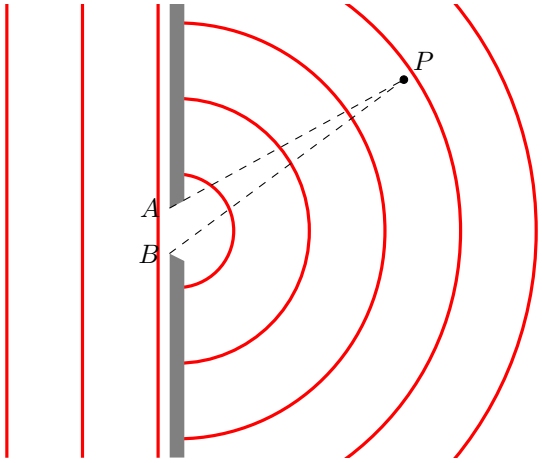


Figure 42.1: Single slit diffraction.

Consider a single slit of width $a = AB$, as shown in figure 42.1. At a given point, P, the maximum possible optical path difference (assuming the wave propagates through air so $n = 1$ for simplicity) is $\ell = |BP - AP|$. This is at most equal to the slit width, $a = AB$. If the wavelength of the light, $\lambda \gg a$ then we also have $\lambda \gg \ell$. Since all secondary wavelets at the slit are in phase we will then have constructive interference at P. This makes no assumption about where P is so for a narrow slit (narrow being defined as $\lambda \gg a$) we have constructive interference everywhere.

If instead $\lambda \ll a$ then we only have $\lambda \gg \ell$ if ℓ happens to be particularly small. This only occurs just in front of the slit, and slightly off to the side. Above a certain angle we will have $\ell \approx \lambda$ and we will have destructive interference. What we have discovered here is we can only have geometrically perfect shadows, with sharp edges, in the limit $\lambda/a \to 0$.

42.2 Near and Far Field Diffraction

Consider a point source of wavelength λ placed at the origin. If we ignore the time varying part of the waves then the scalar amplitude at a point $\boldsymbol{r}$

$$E(\boldsymbol{r}) = \frac{\mathcal{A}}{r}\cos(kr)$$

where $k = 2\pi/\lambda$. Close to the point source we have to use this full equation and treat the waves as spherical.

Notice that we can write r as

$$r = \sqrt{x^2+y^2+z^2} = z\sqrt{\frac{x^2+y^2}{z^2}+1} = z\left(1+\frac{x^2+y^2}{z^2}\right).$$

If we consider the waves at some point with $z \gg x, y$ then using the binomial expansion we have

$$r \approx z + \frac{x^2+y^2}{2z}.$$

This corresponds to assuming that the wave fronts are locally parabolic. We call this the Fresnel or near field diffraction regime.

Even further from the origin we can approximate the waves as locally plane waves. At a distance D from the point source at an angle ϑ_0 to the z-axis considering the point with $y = y_0$ we have

$$r \approx \frac{D}{\cos\vartheta_0} + (y-y_0)\sin\vartheta_0$$

where $y_0 = D\tan\vartheta_0$. See figure 42.2. We call this the Fraunhofer or far field regime. This is the regime in which we will work.

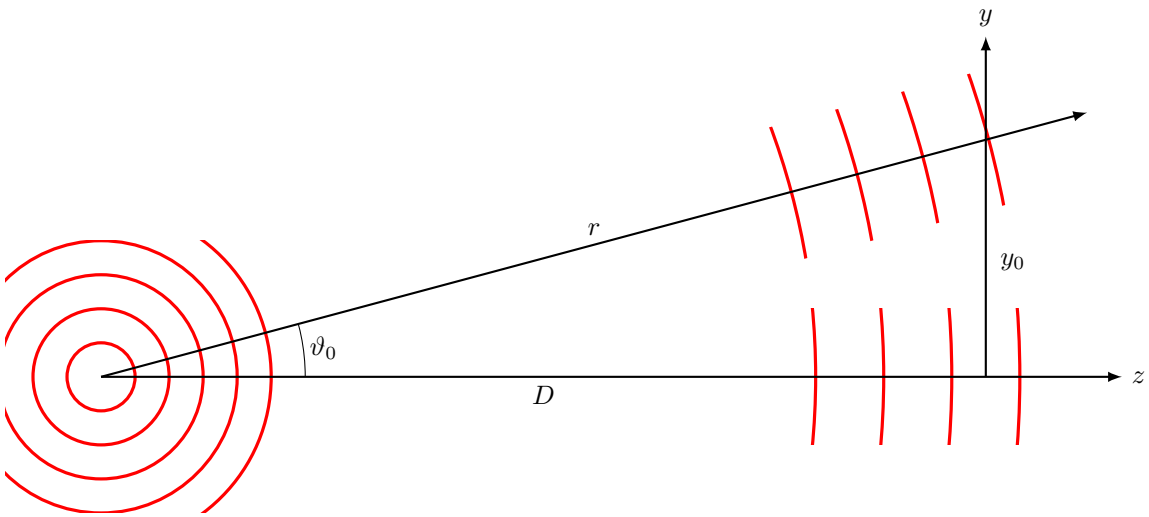


Figure 42.2: Far from the origin we can approximate spherical waves as plane waves.

43 Single Slit Diffraction

43.1 Far Field Single Slit Diffraction

Consider a single slit, a, in the plane P_0 being illuminated by a wave of wavelength λ. We assume the slit is infinite in the other direction and take this direction to be the x-axis. We take the z axis to be normal to the plane. Consider a point s on the plane P_1 which is at a distance D from the split and parallel to P_0. The line from the middle of the slit to s makes an angle ϑ to the z axis. We place the origin in the middle of the slit. As before we consider the two rays coming from A and B, the edges of the slit, and in the far field approximation we consider these two rays to be parallel.

Before we attempt a full solution we will consider what we expect. Suppose we split the slit into $2n$ sections (here $n \in \mathbb{N}$ is *not* the refractive index), each of width $a/2n$. Consider the path difference, ℓ, for a secondary wavelet reaching s from the upper edge of the slit and from a secondary wavelet reaching s from a point $a/2n$ below that. Simple geometry gives us $\ell = (a/2n)\sin\vartheta$. If s is such that $\ell = \lambda/2$ then we will have destructive interference. In fact any point in the first segment will undergo destructive interference, not just the end point. The same logic applies for the next segment over and so we have the condition for minima:

$$a\sin\vartheta = n\lambda, \qquad \text{for} \qquad n = 0, 1, 2, \ldots .$$

By symmetry the same applies for negative ϑ.

If instead s lies on the z-axis then $\vartheta = 0$ and all secondary wavelets arrive in phase causing a central maximum. This maximum is between two minima at $\sin\vartheta = \pm\lambda/a$ and so the width is $2\lambda/a$. We will also find maxima between the other neighbouring minima but the exact angle at which we find them is not simple to derive, however we do know that their widths are λ/a. Notice the reciprocal relationship between slit width and peak width. We will see this again later.

43.1.1 Analytical Solution

We considered special cases of rays in the previous argument. Here we aim to develop a theory that works for all rays coming from the slit. To do this we have to account for the phase shifts of all secondary wavelets that arrive at the point s. The optical path length of a ray leaving the slit at y is $\ell = \ell_0 + y\sin\vartheta$ where ℓ_0 is the optical path length of travelled by light leaving the centre of the slit, however, for our purposes we can simply treat ℓ_0 as a constant as it just depends upon how far away we place the screen. The phase shift due to the point at which the light leaves the slit is then

$$\delta(y) = 2\pi\frac{\ell(y)}{\lambda} = 2\pi\frac{y}{\lambda}\sin\vartheta.$$

We can approximate the distance of s from the slit to be $R = D/\cos\vartheta$ if D is large. We find that the total amplitude, ignoring oscillations, is then given by the superposition of the amplitudes of all rays arriving from anywhere on the slit:

$$E(\vartheta) = \frac{\mathcal{A}}{R} = \int_{-a/2}^{a/2}\cos\delta\,\mathrm{d}y = \frac{\mathcal{A}}{R}\int_{-a/2}^{a/2}\cos\left(2\pi\frac{y}{\lambda}\sin\vartheta\right)\mathrm{d}y = \frac{a\mathcal{A}}{R}\frac{\sin\beta}{\beta} = \frac{a\mathcal{A}}{R}\operatorname{sinc}\beta \tag{43.1}$$

where $\beta = (\pi a/\lambda)\sin\vartheta$. The intensity is then

$$I(\vartheta) = \varepsilon v\langle E^2\rangle = \frac{\varepsilon v}{2}\left(\frac{a\mathcal{A}}{R}\right)^2\operatorname{sinc}^2\beta = I(0)\operatorname{sinc}^2\beta.$$

This vanishes for $\beta = n\pi$ where $n \in \mathbb{Z}\setminus\{0\}$. This condition can also be written as $a\sin\vartheta = n\lambda$, which agrees with our first attempt. See figure 43.1 for a plot of the intensity.

43.1.2 Slit and a Lens

We are considering only the far field limit as this allows us to approximate rays leaving the slit and arriving at the screen at the same point as parallel. Another way to make the same assumption is to place a lens of focal length f just after the slit. This causes parallel rays to converge at a point on the back focal plane, P_{f_b}, which is a distance f from the lens. So a plane wave diffracted in the direction ϑ will be focused at the point $s = f\tan\vartheta \approx f\vartheta$. The intensity on the back focal plane will be

$$I(s) = I_0\operatorname{sinc}^2\left(\frac{\pi a s}{\lambda f}\right).$$

This has zeros at

$$s = n\frac{\lambda f}{a}, \qquad \text{for} \qquad n = \pm1, \pm2, \pm3, \ldots .$$

The central maxima has width $w = 2\lambda f/a$. Here we see again the reciprocal relationship between peak width and slit width.

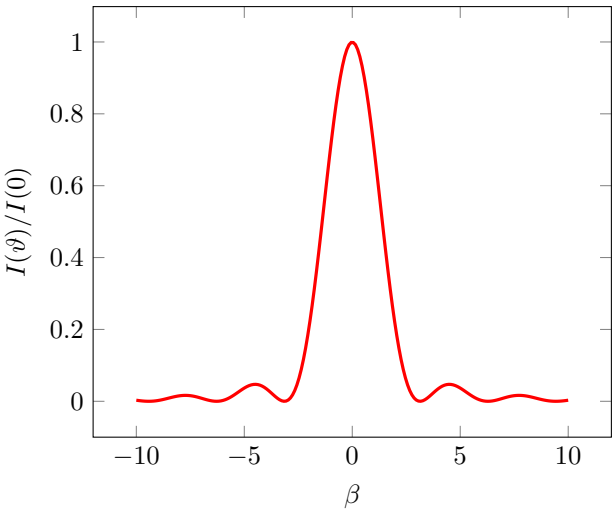


Figure 43.1: The ratio of intensities, $I(\vartheta)/I(0)$, as a function of $\beta = (\pi a/\lambda)\sin\vartheta$

43.2 Fourier Approach

Let $f\colon \mathbb{R} \to \mathbb{C}$ have a Fourier transform. Then the Fourier transform is given by

$$F(q) = \mathcal{F}\{f(x)\} = \int_{-\infty}^{\infty} f(x)e^{iqx}\,\mathrm{d}x\,.$$

If $f\colon \mathbb{R} \to \mathbb{R}$ then

$$F(q) = \int_{-\infty}^{\infty} f(x)\cos(qx)\,\mathrm{d}x + i\int_{-\infty}^{\infty} f(x)\sin(qx)\,\mathrm{d}x\,.$$

We can define a function, p, which represents the slit:

$$p(y) = \begin{cases} 1, & |y| < a/2, \\ 0, & |y| \geq a/2. \end{cases}$$

We can rewrite equation 43.1 as

$$\mathrm{Re}[E(\vartheta)] = \frac{\mathcal{A}}{R}\int_{-\infty}^{\infty} p(y)\cos\left(\frac{2\pi}{\lambda}y\sin\vartheta\right)\mathrm{d}y = \frac{\mathcal{A}}{R}\,\mathrm{Re}[P(q)]$$

where we take E to be the complex amplitude of the field and $P(q) = \mathcal{F}\{p(t)\}$ and $q = (2\pi/\lambda)\sin\vartheta$.

Similarly for the case of the slit and lens set up we have

$$\mathrm{Re}[E(s)] = \frac{\mathcal{A}}{f}\int_{-\infty}^{\infty} p(y)\cos\left(\frac{2\pi}{\lambda}y\frac{s}{f}\right)\mathrm{d}y = \frac{\mathcal{A}}{f}\,\mathrm{Re}[P(q)],$$

where this time $q = 2\pi s/(\lambda f)$. The intensity is then $I(s) = I_0|P(q)|^2$. So we see that the intensity in the back focal plane is proportional to the modulus squared of the Fourier transform of the function representing the slit. It can be shown that

$$\mathcal{F}\{p(x)\} = P(q) = a\,\mathrm{sinc}\left(\frac{aq}{2}\right).$$

We can extend this to a two-dimensional slit which is rectangular with side lengths a and b. This can be defined by the function

$$p(x,y) = \begin{cases} 1, & |x| < a/2, \text{ and } |y| < b/2, \\ 0, & \text{else.} \end{cases}$$

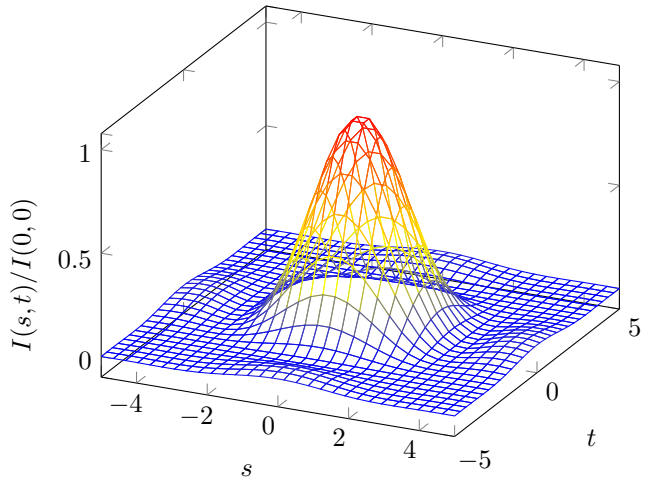
The two-dimensional Fourier transform is then

$$\mathcal{F}\{p(x,y)\} = P(x,y) = ab\,\mathrm{sinc}\left(\frac{aq}{2}\right)\mathrm{sinc}\left(\frac{bp}{2}\right).$$

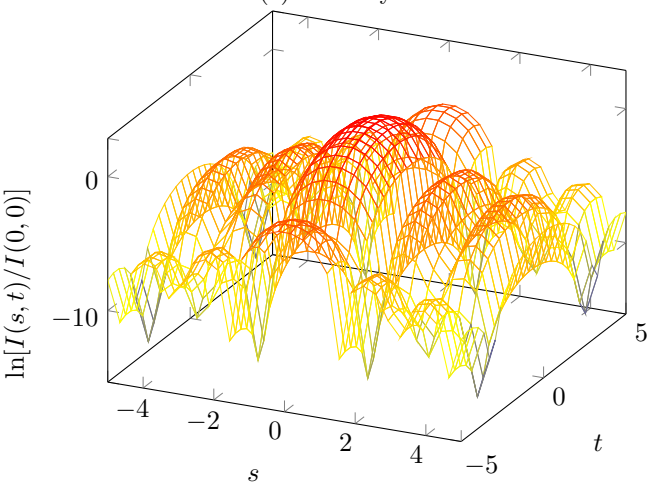
Notice that we can write P and p as products of one-dimensional functions. This means that the intensity can be derived in the same way as above and we find that

$$I(s,t) = I_0 \operatorname{sinc}^2\left(\frac{\pi as}{\lambda f}\right) \operatorname{sinc}^2\left(\frac{\pi bt}{\lambda f}\right).$$

See figure 43.2.



(a) Intensity



(b) Log intensity.

Figure 43.2: The intensity, $I(s,t)$, and log intensity, $\ln[I(s,t)]$ of a slit of width 2 and height 1.

44 Two Slit Diffraction

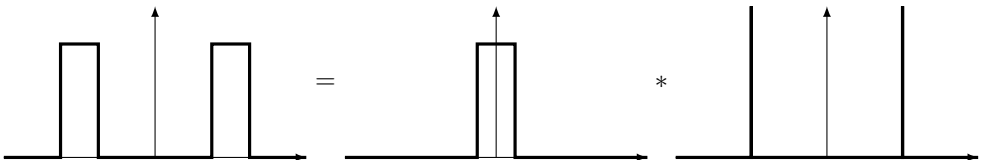
We have now seen that the Fourier approach is a powerful one. We will apply it in this section to the famous two slit experiment. Let p be a single slit transition function for a slit of width a centred at $x = 0$:

$$p(x) = \begin{cases} 1, & |x| < a/2, \\ 0, & \text{else.} \end{cases}$$

We can then model two slits using delta distributions and a convolution:

$$f(x) = p(x) * \left[\delta\left(x - \frac{d}{2}\right) + \delta\left(x + \frac{d}{2}\right)\right].$$

Graphically this corresponds to



Using the convolution theorem we can write the Fourier transform of f as the product of the Fourier transforms of p and the delta distributions:

$$F(q) = \operatorname{sinc}\left(\frac{aq}{2}\right)\cos\left(\frac{dq}{2}\right).$$

The intensity is then

$$I(\vartheta) = I_0|F(q)|^2 = I_0\operatorname{sinc}^2\left(\frac{\pi a}{\lambda}\sin\vartheta\right)\cos^2\left(\frac{\pi d}{\lambda}\sin\vartheta\right).$$

If we add in a lens of focal length f then

$$I(s) = I_0\operatorname{sinc}^2\left(\frac{\pi a}{\lambda f}s\right)\cos^2\left(\frac{\pi d}{\lambda f}s\right).$$

What we end up with is $\cos^2$ fringes modulated by sinc^2, as can be seen in figure 44.1.

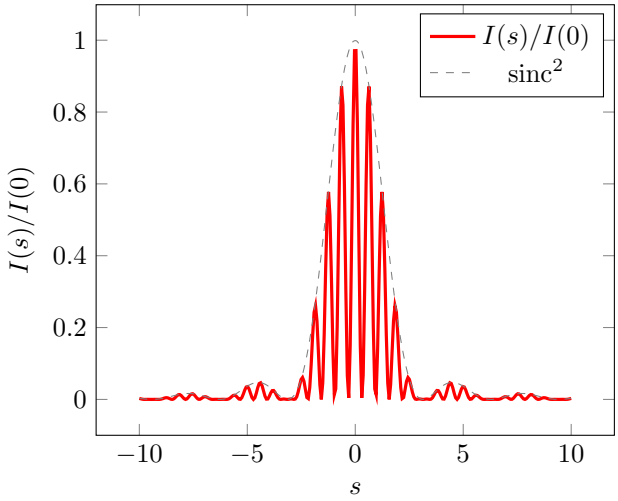


Figure 44.1: The intensity from a double slit is $\cos^2$ fringes modulated by sinc^2.

This derivation was much quicker than the geometric argument that one has to make if they lack Fourier analysis. We can see that the extrema occur whenever the argument of cos is $n\pi$ or $(n+1/2)\pi$ for $n \in \mathbb{Z}$, which corresponds to

$$d\sin\vartheta = n\lambda, \qquad \text{or} \qquad d\sin\vartheta = (n+1/2)\lambda$$

which are the conditions for constructive and destructive interference respectively.

44.1 Diffraction Gratings

A diffraction grating is a large number of equally spaced thin slits. The phase difference between light from adjacent slits is

$$\delta = \frac{2\pi}{\lambda}d\sin\vartheta$$

where d is the space between adjacent slits. We know that wavelets will add constructively when $\delta = 2\pi n$ for $n \in \mathbb{Z}$. Therefore we expect peaks when

$$d\sin\vartheta = n\lambda.$$

If we include a lens of focal length f then we get peaks at

$$s = n\frac{\lambda f}{d}.$$

This doesn't tell us anything about the shapes of the peaks, just their locations, to get more information we again turn to Fourier analysis.

If we approximate the number of slits as infinite and take the slits to have no width then we can then model the transmission function as a Dirac comb:

$$c(x) = \sum_{m\in\mathbb{Z}} \delta(x - md).$$

The Fourier transform of this is simple to compute and is

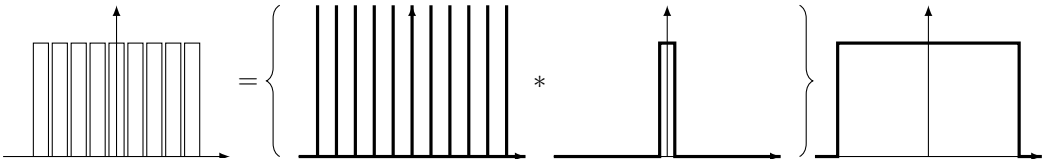
$$C(q) = \frac{1}{d}\sum_{m\in\mathbb{Z}} \delta(q - 2\pi m/d).$$

Notice that the spacing of the original comb, d, is reciprocal to the spacing of the Fourier transformed comb, $2\pi/d$. The intensity is then $I(\vartheta) \propto |C(q)|^2$. This is zero except when $q = 2\pi m/d$ for $m \in \mathbb{Z}$. Since $q = (2\pi/\lambda)\sin\vartheta$ this corresponds to $m\lambda = d\sin\vartheta$. This is the same condition we found before.

This result is unrealistic as in reality there aren't an infinite number of slits and the slits have finite width. We can modify for the finite width by convolving the Dirac comb with a single slit transmission function, p, and we can modify for the finite number of slits by multiplying by a top hat function, w, of width $D = Nd$ where N is the total number of slits. Thus the transmission function is

$$f(x) = [c(x) * p(x)]w(x).$$

Graphically this is



By the convolution theorem the Fourier transform of the transmission function is

$$F(x) = [W(q) * C(q)]P(q).$$

These functions are given by

$$\begin{aligned} C(q) &= \sum_{m\in\mathbb{Z}} \delta(q - 2\pi m/d), \\ P(q) &= \operatorname{sinc}(aq/2), \\ W(q) &= \operatorname{sinc}(Ndq/2). \end{aligned}$$

The intensity is

$$I(\vartheta) = I(0)|F(q)|^2 = I(0)\left|F\left(\frac{2\pi}{\lambda}\sin\vartheta\right)\right|^2.$$

44.1.1 Spectroscopy

The intensity of light from a diffraction grating depends on the wave length, and therefore colour, of the light. So if a white light source is used we will see the colours separate out and each colour will have a maxima at a slightly different point. This makes diffraction gratings useful in spectroscopy. At this point we note that most diffraction gratings used for spectroscopy aren't actually made of lots of slits in an opaque material but rather a transparent medium with lots of thin parallel scratches. It is also common to instead use a reflection grating, which is basically the same but with reflective coating, such as aluminium. This is useful because glass absorbs strongly in the UV part of the spectrum and so glass transmission gratings don't work for UV light.

A large amount of light from a diffraction grating ends up in the zeroth order, $\vartheta = 0$, maxima. This is a problem because the intensity of this maxima *doesn't* depend on λ making this fringe useless for spectroscopy. This is a problem if the light source we are using isn't very intense as the first order fringes, which can be used for spectroscopy, may be hard to see. One solution to this is to use a reflection grating with a surface that looks like this:

This is called a blazed grating. Light that enters normal to the macroscopic surface is reflected mostly at an angle 2γ where γ, called the blazing angle, is the angle of the microscopic surface to the macroscopic surface. The zeroth order maxima from the diffraction however remains in the same place but is now much weaker. Instead by carefully selecting the blazing angle we can focus the majority of light onto where we expect the first order maxima to be and this light will be separated by wavelength.

45 More Diffraction

45.1 Circular Aperture

It is very common for an optical system to have cylindrical symmetry about the beam axis, for example telescopes and microscopes. Therefore it is worth considering a circular aperture such as one might use for an eyepiece. The transmission function for a circular aperture of diameter a is

$$p(x,y) = \begin{cases} 1, & x^2 + y^2 \leq a^2/4, \\ 0, & \text{else.} \end{cases}$$

The Fourier transform of this is

$$P(\rho) \propto \frac{J_1(\rho)}{\rho}$$

where $\rho = (\pi a/\lambda)\sin\vartheta$ and J_1 is the first order Bessel function of the first kind, the exact details of which aren't important. The intensity at angle ϑ is then

$$I(\vartheta) = 4I(0)\left|\frac{J_1([\pi a/\lambda]\sin\vartheta)}{(\pi a/\lambda)\sin\vartheta}\right|^2.$$

If we include a lens of focal length f then the intensity at position (s,t) on the back focal plane is

$$I(s,t) = 4I(0,0) = \left|\frac{J_1(\pi ar/(\lambda f))}{\pi ar/(\lambda f)}\right|^2, \qquad \text{where} \qquad r^2 = s^2 + f^2.$$

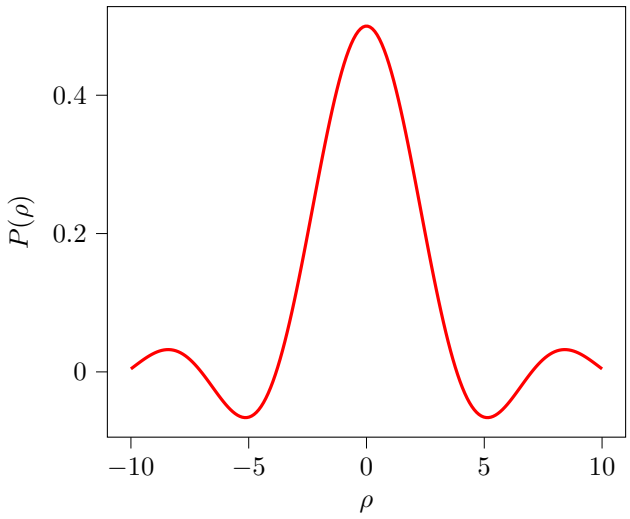
This result is very similar to the result of a single slit and has a similar profile at $y = 0$ as shown in figure 45.1. The first minimum occurs at $\rho = 1.22\pi$, which corresponds to $\sin\vartheta = 1.22\lambda/a$ (as opposed to $\sin\vartheta = \lambda/a$ which we saw for a single slit).

The function $J_1(\rho)/\rho$ is actually two-dimensional with ρ just being the distance from the origin. The is plotted in figure 45.2.

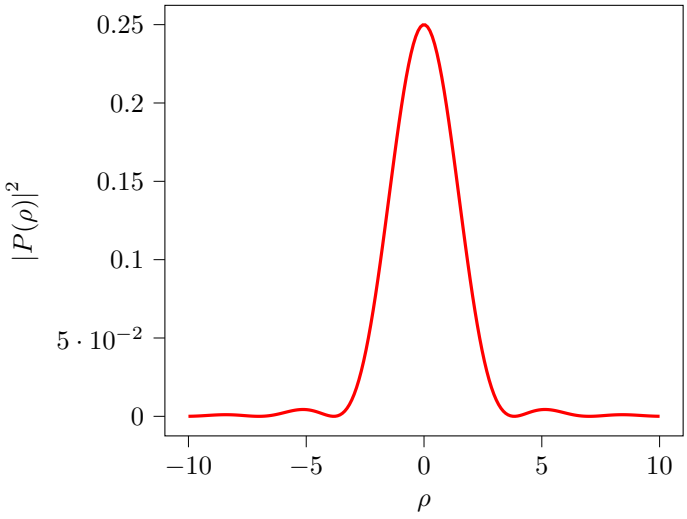
Suppose we view a distant point object through a lens of focal length f and diameter a. The lens is circular so in the back focal plane of the lens we will get a diffraction pattern corresponding to a circular aperture. This has the first zero at radius

$$r_0 = \frac{1.22\lambda f}{a} = 1.22\lambda F_{\text{No}}.$$

Here we have introduced the **f-number**, $F_{\text{no}} = f/a$ which is the ratio of the focal length and aperture. This number is important in photography, where it is usually written as f/F_{No}, with f left as f but F_{No} replaced with a number, for example $f/2.8$, if we substitute for the focal length as well then we would get the aperture of the lens. If F_{No} is large then a photograph will have a blurred background as the intensity of light from each point is very spread out. If F_{No} is small then the photograph will have a sharp background as the intensity drops to zero very quickly. round the image of the point object there will be rings, known as **Airy rings**, and the whole pattern that appears from a single point object is called the **point spread function**.

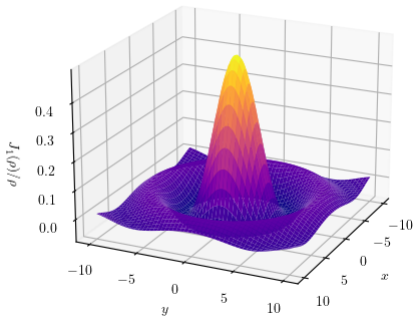


(a) $P(\rho) \propto J_1(\rho)/\rho$.

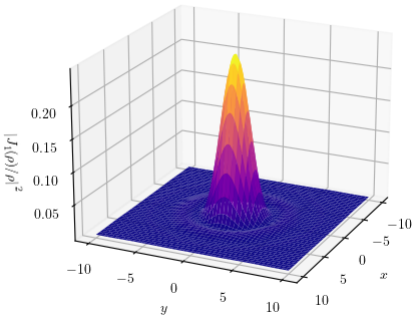


(b) $I(\rho) \propto |P(\rho)|^2 \propto |J_1(\rho)/\rho|^2$.

Figure 45.1: The Fourier transform, P, of the transmission function $p(x, y) = 1$ if $x^2 + y^2 \leq a^2/4$ and $p(x, y) =$ otherwise, and the intensity of the light that results from this circular aperture.



(a) $J_1(\sqrt{x^2+y^2})/\sqrt{x^2+y^2}$.



(b) $\left|J_1(\sqrt{x^2+y^2})/\sqrt{x^2+y^2}\right|^2$.

Figure 45.2: The functions $J_1(\sqrt{x^2+y^2})/\sqrt{x^2+y^2}$ and $\left|J_1(\sqrt{x^2+y^2})/\sqrt{x^2+y^2}\right|^2$.

The point spread function limits the spatial resolution of the system. If two point objects have angular separation $\Delta\vartheta$ then in the back focal plane of the lens we will have two point spread functions with their

peaks separated by $s = f\Delta\vartheta$ (assuming $\Delta\vartheta \ll 1$). There are three possibilities:

$s \gg r_0$ Two well separated point spread functions, it is easy to tell the two points apart.

$s \ll r_0$ The two point spread functions merge and it is not possible to tell the two points apart.

$s \approx r_0$ There will be some limit at which we say the points are resolved.

The **Rayleigh criterion** states that two points are first resolved when $s = r_0$. Under this condition the peak of one point spread function appears at the zero of the other resulting in a twin peak with a 20 % dip in the middle. This corresponds to the angular condition that

$$\Delta\vartheta = 122\frac{\lambda}{a}.$$

See figure 45.3.

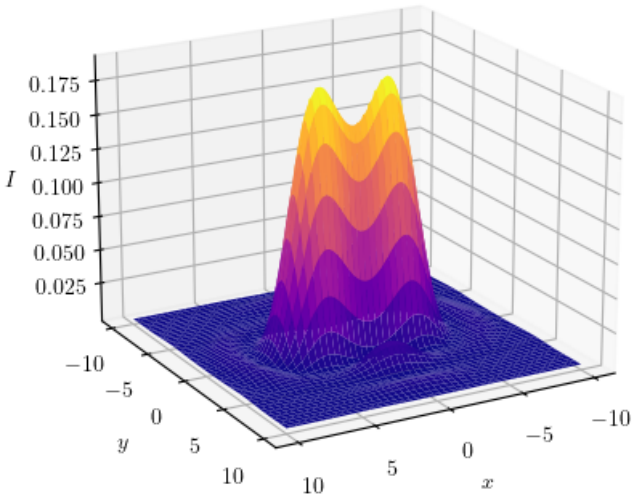


Figure 45.3: The intensity on the back focal plane for two points at separation $s = 5$.

46 Even More Diffraction

46.1 Diffraction in 1D

When we first considered diffraction gratings in section 44.1 we considered the slits to be narrow in one dimension and infinitely long in the orthogonal direction. The incident light was normal to both of these axes. If we truly restrict ourselves to one dimensional slits then we can consider non-normal incidence. If light is incident on the diffraction grating (either a transmission or reflection grating) at an angle ϑ_i to the surface normal and leaves the grating at an angle ϑ_m to the surface normal then the condition for constructive interference is

$$m\lambda = d(\sin\vartheta_m - \sin\vartheta_i)$$

for $m \in \mathbb{Z}$. As ϑ_i is increased from 0 to $\pi/2$ then we lose diffraction peaks with $m > 0$ and gain higher order peaks for negative m. For grazing incidence at $\vartheta_i = \pi/2$ the condition for constructive interference reduces to

$$\sin\vartheta_m = \frac{m\lambda}{d} + 1.$$

As well as this since the light simply passes straight past the diffraction grating we require that $\vartheta_m = \pm\pi/2$. The case of $\vartheta_m = \pi/2$ corresponds to $m = 0$ so the zeroth order peak is directly ahead. The case of $\vartheta_m = -\pi/2$ corresponds to $-m\lambda = 2d$ which corresponds to the beam heading back the way it came. This returning beam can only exist if d is equal to an integer number of half-wavelengths, or if $k = m\pi/d$.

46.2 X-ray Diffraction

Moving back to three dimensions consider an array of identical holes. We expect the diffraction pattern to be the superposition of the diffraction patterns from each hole. As with the initial diffraction grating we can describe the transmission function with a convolution of single hole transmission function and a two dimensional Dirac comb. We can then deal with the finite extent of the grating by multiplying by a two dimensional top hat function. Once we've done this we simply take the square of the Fourier transform for the intensity. The Fourier transform is easy to find if the scattering centres are evenly distributed, in particular they are located at $\boldsymbol{R} = l\boldsymbol{a} + m\boldsymbol{b}$ for $l, m \in \mathbb{Z}$, and basis vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

We can also consider a three dimensional diffraction grating. The most common example being a crystalline solid. The repeat distance in this case is the interatomic spacing, which is on the order of 10^{-10} m, which is about four orders of magnitude smaller than the wavelength of visible light. This means that visible light will display only the first order diffraction, which isn't very useful. Instead we use X-rays. Again we can decompose the system into scattering units with some given transmission function that we then replicate through space by convolution with a three-dimensional Dirac comb and multiplication by some three-dimensional top hat function. Taking $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ as basis vectors then we have a scattering unit at each $\boldsymbol{R} = l\boldsymbol{a} + m\boldsymbol{b} + n\boldsymbol{c}$ for $l, m, n \in \mathbb{Z}$. The set of all such $\boldsymbol{R}$ gives all of the locations and we can distinguish any one such point with three numbers, (l, m, n). Consider one particular, but arbitrary, scattering centre positioned at $\boldsymbol{R}$. It can be shown that there will be constructive interference if and only if

$$\boldsymbol{R} \cdot (\boldsymbol{k} - \boldsymbol{k'}) = 2\pi q$$

for some $q \in \mathbb{Z}$. Here $\boldsymbol{k}$ and $\boldsymbol{k'}$ are the wave vectors of the incident and scattered rays. All scattering sites which are an integer multiple of $\boldsymbol{R}$ away from this original scattering centre will also contribute constructively to the interference. There are a few observations to be made:

- The wave vectors, $\boldsymbol{K}$, satisfying $\boldsymbol{R} \cdot \boldsymbol{K} = 2\pi q$ for some $q \in \mathbb{Z}$ are special. The correspond to harmonic waves with the same periodicity as the lattice. These are the waves that appear in the Fourier series of the lattice.
- This allows us to state the constructive interference condition generally as $\boldsymbol{k} - \boldsymbol{k'} = \boldsymbol{K}$.
- We can also use Fourier methods to find the intensity as well as just the extrema, this requires us to take a Fourier transform of the electron density surrounding each scattering centre as it is the electrons that are responsible for scattering the X-rays.