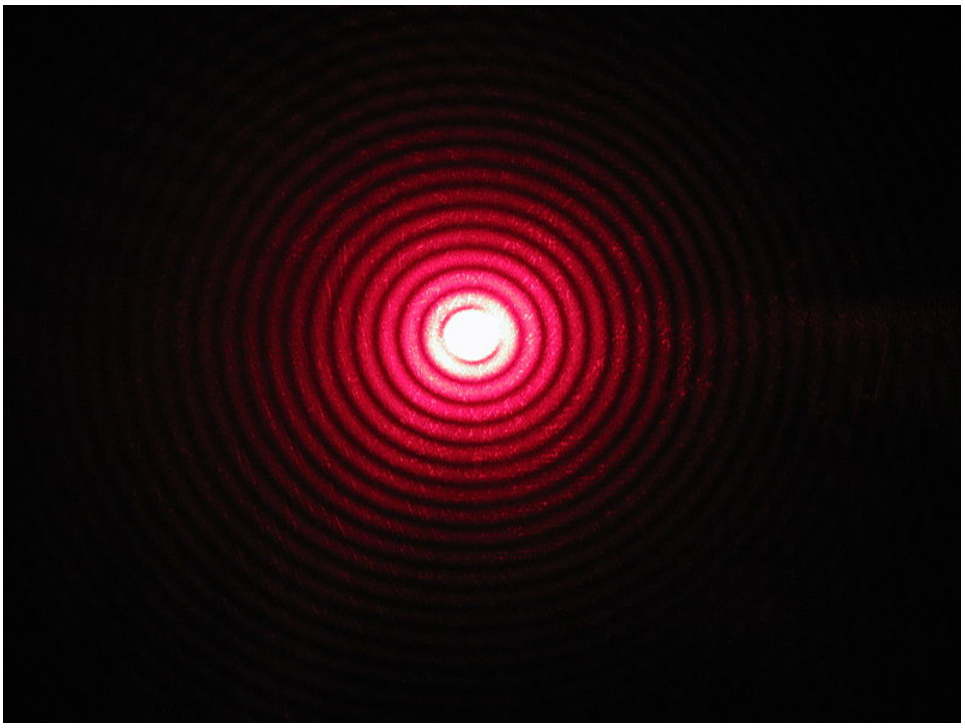


Optics

PY3101

M.P. Vaughan

University College Cork (2014)

Contents

Part I

Introduction to Optics

1. Overview

1.1 General remarks

In this overview, we present a short history of optics highlighting the central protagonists and concepts giving foundation to our modern understanding of the subject. Some of the important applications are then given briefly, although this is in no way intended to be a comprehensive list. Finally, we give a broad brush-strokes overview of the course material, indicating the scope and the intended development of those ideas. We also suggest, in the bibliography, additional texts that the student may find useful.

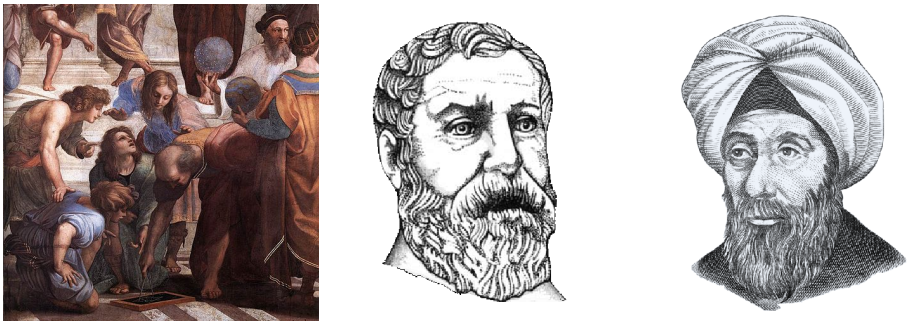
1.2 Learning objectives

The learning objectives of this chapter are to

- Gain a basic appreciation of the history of optics and the development of the principal concepts.
- Have a appreciation for some of the real-world applications of optics
- Gain an overview of how these ideas are developed in this course

1.3 A short history of optics

Traditionally, the subject of *optics* has dealt with the nature of visible light. Applications of optical principles may be traced back into antiquity, with references to the use of lenses (or water filled vessels) for magnification and burning glasses and the polishing of metal to obtain a mirrored surface. From a modern perspective, we would recognise these as applications of *geometrical optics*, although it was not until the considerations of the Ancient Greeks that any kind of theoretical understanding of these phenomena was attempted.



(a) (b) (c)

Figure 1.1: (a) Detail from Raphael's *The School of Athens* depicting Euclid (with dividers bending over) [*image in public domain*]. (b) Illustration of Hero of Alexandria from a 1688 German translation of Hero's *Pneumatics* [*image in public domain*]. (c) Image of Alhazen as shown on the obverse of the 1982 Iraqi 10 dinar note.

1.3.1 The Ancient Greeks

One of the earliest surviving theoretical text is Euclid's *Optics* (circa 300 BC). It is perhaps not too surprising for the Father of Geometry, that Euclid gave a geometrical account of light, in which he viewed light as a cone of rays with the eye at the vertex. His account then explains the phenomenon of perspective and unseen objects. However, the idea that light travels in straight lines remained an unexplained assumption.

Later, around 40 AD in the work *Catoptrica*, Hero (or Heron) of Alexandria was able to show geometrically that the path of a ray reflected from a plane to an observation point is the shortest possible path the light could have taken, subject to the constraint that the ray must touch the plane. In retrospect, we might call this 'the principle of least distance' and bears close resemblance to a modern explanation. Hero, however, did not appreciate that light travels at different speeds in different media and it turns out that the correct principle would be one of *least time*. Nowadays, this would often be described in terms of the *optical path length*, which has dimensions of distance but remains directly proportional to the time taken.

1.3.2 The Golden Age of Islam

With the demise of the Greek civilisation, the torch of philosophical enquiry was taken up in the Islamic world. The mathematical advances of in number bases, algebra and the development of algorithmic methods are well known, if not always attributed to Islamic thinkers.

One important early contributor to the field of optics was Ibn Sahl (c. 940

- 1000) who, in his treatise *On Burning Mirrors and Lenses* (984) explains the focussing of light by curved mirrors and lenses. Significantly, he also gives the first detailed formulation of what is now known as *Snell's Law of refraction*

Another major protagonist was the polymath Alhazen (c. 965 - c. 1040). Between 1011 to 1021 Alhazen wrote a seven volume treatise on optics *Kitab al-Manazir* (Book of Optics). In the course of this work, Alhazen conducted many experiments on the rectilinear (straight line) propagation of light, reflection and refraction. This early adoption of the Scientific Method makes Alhazen's work significant in the more general history of Science. However, in terms of optics, his main contribution is often held to be his detailed description of the human eye.

1.3.3 The Enlightenment

(a) (b) (c) (d)

Figure 1.2: (a) Portrait of Galileo Galilei by Giusto Sustermans [*image in public domain*]. (b) Portrait of René Descartes after Frans Hals, 1648 [*image in public domain*]. (c) Portrait of Isaac Newton by Godfrey Kneller 1689 [*image in public domain*]. (d) Portait of Pierre de Fermat [*image in public domain*].

The Enlightenment or *Age of Reason* in the 17th and 18th centuries saw the rise of many great thinkers contributing to our understanding of the world. Most notable of these for their work on the fundamentals of mathematical physics and general science were Galileo, Descartes and, of course, Newton.

Although Galileo (1564 - 1642) did not invent the telescope - that achievement is attributed to the German-Dutch spectacle-maker Hans Lippershey (1570 - 1619) - he did improve on Lippershey's design. Grinding his own lenses to optimise their power of magnification, he achieved a nine-fold magnification. Using his instrument, he went on to verify the phases of Venus, observe sunspots and discover the four largest Jovian moons (now known as the Galilean moons).

In his works on optics, René Descartes (1596 - 1650) independently derived Snell's Law, named after Willebrord Snellius (1580 - 1626) and first formulated by Ibn Sahl, in terms of sine functions. Using this result, he was also able to show that the angular radius of a rainbow from an observer is 42° (as we find in Chapter 6). Although incorrectly predicting that light would travel faster in a 'denser' medium (i.e. medium with higher refractive index) he consistently interpreted light as a form of *wave propagation*, in line with the modern view.

Another crucial concept for wave optics was provided by the Dutch physicist Christiaan Huygens (1629 - 1695). Huygens asserted his now well-known principle that each point on a wavefront acted as a source of secondary wavelets that, at some later time, added up to give a new wavefront. *Huygens' Principle* remains a powerful tool in the analysis of wave motion, although on its own cannot explain why the wavefront moves in a given direction rather than backwards as well. The explanation for this requires the concept of interference, which we shall shortly encounter.

Despite Descartes' adoption of the wave interpretation of light, Isaac Newton (1642 - 1727) came to reject the idea, favouring a particle description of light. This view is known as the *corpuscular theory of light*. Newton's own contributions to the field of optics were of great significance. In 1666, Newton demonstrated the decomposition of white light into the different colours of the rainbow via refraction through a *prism*. This represented an early observation of the *dispersion* of light, in which light of different colours travel at different speeds in a transparent medium, although Newton did not interpret his results in this way.

A similar phenomenon occurs with *chromatic aberration* in lenses. Newton was familiar with this problem and erroneously believing it to be irreparable, set about designing and building the first *reflecting telescope*, using curved mirrors to achieve magnification of the distant image.

Perhaps the most important contribution to geometrical optics was from the French mathematician Pierre de Fermat (1607 - 1665) (he of the famous 'last theorem'). In 1657 he enunciated his Principle that

- *the path taken between two points by a ray of light is the path that can be traversed in the least time*

This is closely related to Hero's Principle of least distance and, like that earlier concept, is a *variational principle* that finds the optimum path as an extremum of some observable. A modern version of Fermat's Principle multiplies the time of travel by the wave speed at a given time to obtain the *optical path length* in dimensions of length.

(a) (b) (c)

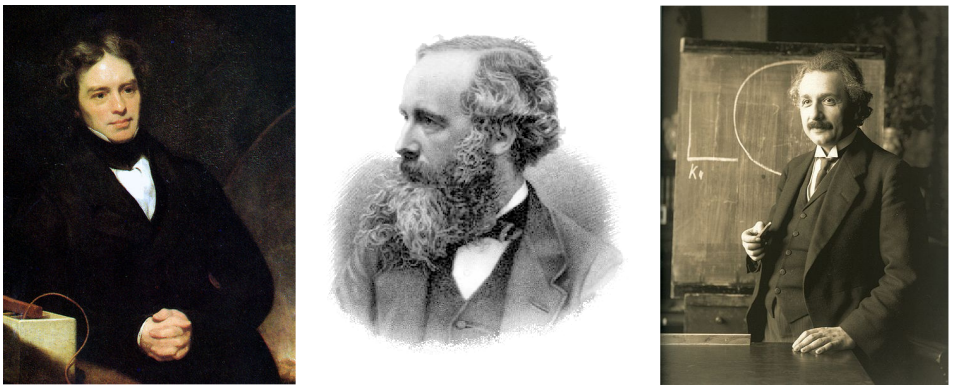
Figure 1.3: (a) Portrait of Christiaan Huygens by Bernard Vaillant [*image in public domain*]. (b) Engraving of Thomas Young [*image in public domain*]. (c) Portrait of Augustin-Jean Fresnel [*image in public domain*].

1.3.4 Wave optics

Such was Newton's intellectual influence at the height of his powers that the corpuscular theory of light tended to hold sway simply because Newton said it was so. Indeed, it is with some irony that when Augustin-Jean Fresnel (1788 - 1827) submitted a paper to the French Academy of Sciences in 1818 expounding the wave view of light, Poisson, a member of the judging committee and supporter of the corpuscular theory, attempted to disprove Fresnel's work by arguing that this would imply the appearance of a bright spot in the shadow of an illuminated sphere. In fact, such a bright spot does indeed exist, now known as the *Fresnel bright spot*.

Although a major figure in the development of wave optics, Fresnel's work was slightly pre-dated by that of Thomas Young (1773 - 1829). In fact it was Young who provided the first conclusive evidence for the wave nature of light in his now famous *double slit experiment* in 1803 (see Chapter 5). The double slit experiment showed that light clear exhibited the wave phenomenon of *diffraction*, an effect totally inexplicable in terms of corpuscles (particles).

Fresnel independently discovered Young's work, adding substantially to the theory of diffraction (Chapter 5) as well being one of the first people to appreciate the importance of the *polarisation* of light. He also derived the equations that now bear his name for the reflectance and transmission ratios of light crossing between different media (Chapter 8).



(a) (b) (c)

Figure 1.4: (a) Portrait of Micheal Faraday by Thomas Phillips, 1842 [*image in public domain*]. (b) Engraving of James Clerk Maxwell by G. J. Stodart from a photograph by Fergus of Greenack [*image in public domain*]. (c) Photograph of Albert Einstein during a lecture in Vienna in 1921 [*image in public domain*].

1.3.5 Electromagnetism

Although Young had convincingly demonstrated the wave nature of light, the underlying physics required the theory of *electromagnetism* for its elucidation. The theory of electromagnetism builds on the work of many mathematicians and scientists, Gauss, Coulomb and Ampère to name just a few. Here, we shall just introduce the reader to those who's work most greatly contributes to our understanding of optics.

One of the major contributors to the field is, of course, Michael Faraday (1791 - 1867). In particular, Faraday discovered the *Law of Electromagnetic Induction* (Chapter 6), in which a changing magnetic field induces a changing electric field.

The great physicist James Clerk Maxwell (1831 - 1879) incorporated Faraday's and other laws, along with his own contributions, into a set of equations known as *Maxwell's Equations*. Maxwell was then able to show that these equations yielded a *wave equation* for self-propagating electromagnetic (EM) waves. Moreover, the speed of these waves could be calculated from fundamental universal constants: the permittivity and permeability of free space. This meant that the speed of EM waves c *must also be a fundamental constant*. When Maxwell actually calculated this value, he found that this speed was (allowing for experimental error) *the same as the measured speed of light*. This must have been a tremendously exciting moment for Maxwell, who then concluded that light is, in fact, a form of *electromagnetic radiation* [1].

A few years later in 1886, Heinrich Hertz (1857 - 1894) successfully

demonstrated the generation of radio frequency EM waves via the changing electric field of oscillating charge in a dipole antenna [2]. The SI unit of frequency is now named the *hertz* (Hz) in commemoration of this achievement.

At the time, it was generally believed that all wave propagation must take place in some kind of physical medium. The proposed medium for light was called the 'luminous ether'. However, since the speed of light c was found to be a constant, it was assumed that this must be relative to the ether. Moreover, since the Earth is moving in an orbit around the Sun, it would seem that the Earth itself must sometimes have a relative speed to the reference frame of the ether.

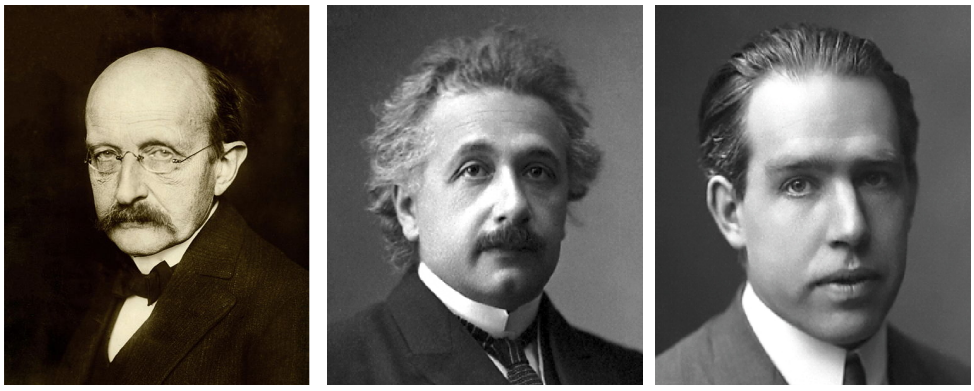
As a test of these ideas, in 1887 Michaelson and Morley attempted to measure the relative motion of the Earth against the luminous ether by measuring the speed of light at different points in its orbit using an *interferometer*. To their surprise, Michaelson and Morley discovered that the speed of light was *always the same*, which seemed a very counter-intuitive result. Hendrik Antoon Lorentz (1853 - 1928) modelled this using by developing his *Lorentz transformations* and attempted to explain the phenomenon in terms of 'contraction' of the ether. However, his transformations also implied a bizarre dilation of the temporal dimension that he could not explain.

It was left to Albert Einstein (1879 - 1955) to complete the electromagnetic explanation of light. In 1905, he published his paper on *Special Relativity* based on the mathematics of the Lorentz transformations [3]. He dispensed with the idea of 'absolute space and time', arguing that only relative motion was physically meaningful. In this way, he showed that the concept of the luminous ether was superfluous to requirements. Later, Einstein's former Mathematics teacher Hermann Minkowski (1864 - 1909) formalised Einstein theory, showing that his concepts of relativity could interpreted in terms of a four-dimensional continuum, now known simply as *spacetime*.

1.3.6 Quantum mechanics

At the turn of the century, it seemed that classical physics was about to mop up all the remaining problems of physics. Only a few loose threads hung from the tapestry, which were about to be pulled on.

One of these outstanding problems was the explanation of *blackbody radiation*. A blackbody is a body that both absorbs all the radiation incident on it and emits radiation with a spectral density characteristic of its temperature. This was the problem that Max Planck (1858 - 1947) was concerning himself with in 1894. At the time, the spectral density could be explained in both the high and low frequency limits but no single law for the spectrum of radiation over all ranges was known (Chapter 2).



(a) (b) (c)

Figure 1.5: (a) Photograph of Max Planck in 1933 [*image in public domain*]. (b) The official 1921 Nobel Prize in Physics photograph of Albert Einstein, which he won for his explanation of the photoelectric effect [*image in public domain*]. (c) Photograph of Niels Bohr around 1922 from the Nobel Prize Biography.

Planck eventually found that the only consistent explanation he could come up with required that the light must be emitted in discrete *quanta*. These quanta would then have an energy $\epsilon = hf$, where h is *Planck's constant* and f is the frequency of the light. This apparently went against all the conventional wisdom that light was in fact a form of electromagnetic radiation and seemed to be resurrecting the corpuscular theory of light again. Needless to say, Planck's paper [4] was not well-received.

One person who did take Planck's ideas seriously was Albert Einstein, who in 1905 realised that he could explain the *photoelectric effect* by extending Planck's ideas to the notion that the *electromagnetic field itself was quantised* [5]. In the photoelectric effect, electrons are liberated from an illuminated sample of metal. However, in an apparent contradiction with the classical theory of light, the energy of these electrons is *not* proportional to the intensity of the radiation. Rather, greater intensity just leads to a greater number of electrons. On the other hand, the energy of the electrons *is* dependent on the frequency of light. Einstein therefore postulated that these liberated electrons were ionised by the absorption of a *quantum* of light with an energy $\epsilon = hf$. The intensity of the radiation field is then just proportional to the *number of light quanta* in it. These quanta were later dubbed *photons*.

This dual explanation of light has given rise to the use of the expression *wave-particle duality*. However, this is possibly a misleading way of putting things. It suggests that light is 'sometimes' a wave and 'sometimes' a particle. In fact, all the particle-like effects associated with photons can be reproduced by the construction of a *wave packet*, which can be made

as localised as one likes. However, in doing so, one requires an increasing number of wavelength components to keep the packet localised. This leads, via Fourier's Theorem, to *Heisenberg's Uncertainty Principle* for knowledge of pairs of physical observables such as position and momentum. It should be remembered, however, that this is *not* inconsistent with Maxwell's explanation of light.

One of the most pressing problems left outstanding by classical physics was an explanation of atomic stability (see Chapter 2). According to electromagnetic theory, a negative electron orbiting a positive nucleus ought to radiate away its energy and spiral catastrophically into the nucleus. The first major insight into the resolution of this problem came when Niels Bohr (1885 - 1962) proposed that the *angular momentum of the electrons was quantised* [6]. This implied that there were only certain discrete orbits around the nucleus that the electron could occupy. Transitions between these orbits then occurred via the *emission* or *absorption* of a photon with the correct energy difference. This new theory also explained the discrete emission and absorption lines seen in the spectral signatures of different chemicals.

Whilst Bohr's model was not entirely correct, the elements of quantised angular momentum and the electron-photon interaction were. With the added ingredient of quantisation, the photon picture of light turns out to be consistent and complementary to the electromagnetic explanation.

The first formal theory of quantum mechanics was that of *matrix mechanics*, developed by Werner Heisenberg (1901 - 1976), Max Born (1882 - 1970) and Ernst Pascual Jordan (1902 - 1980). Although the original concepts were envisaged by Heisengberg, it seems that Born did most of the mathematical heavy-lifting - including the derivation of the famous *Uncertainty Principle* (if fact, it appears the Heisenberg did not even know what a matrix was when he handed his original ideas to Born).

Nonetheless, the Uncertainty Principle now bears Heisenbergs name and remains one of the most profound concepts of quantum theory. A more detailed discussion will be given in Chapter **??** but for now, it suffices us to recognise that, when rendered in terms of *wave mechanics* we can can resolve the apparent wave-particle duality paradox via the strategy of constructing *wave packets*, from which the Uncertainty Principle naturally emerges.

A further important result of wave mechanics is the *de Broglie wavelength* due to Louis-Victor-Pierre-Raymond, 7th duc de Broglie (1892 - 1987), giving the relationship between wavelength and momentum. Although this hypothesis was originally proposed to describe the wave nature of material particles, similar forms remain valid for photons.

1.4 Applications of optics

The applications of optical theory are many and varied. Here we list just a taster of possible applications.

1.4.1 Lenses and mirrors

The earliest practical applications of optics were based on the properties of lenses and mirrors. Most of the necessary theoretical understanding for such applications is based in *geometrical optics*, which we cover in Part IV.

Ophthalmetry

Preceding the invention of telescopes and microscopes was the use of lenses for correcting deficiencies in vision. This area is known as *ophthalmetry* and has traditionally made great use of geometrical optics. More recently, there has been growing use of lasers in corrective surgery.

Telescopes and microscopes

Optical telescopes may be categorised into *refracting* and *reflecting*. Refracting telescopes, as developed by Galileo, soon found use in seafaring and navigation. The area of optical astronomy was also opened up by both refracting and reflecting telescopes, the latter originally invented by Newton to overcome chromatic aberration.

The use of microscopes to image the very small has also opened up other realms of inquiry in turn, most notably in the biological sciences.

Photography

Another area where optimal imaging is a necessary requirement is in photography. Compound photographic lens of all types are designed for different purposes. Additionally, high quality lenses often incorporate anti-reflective (AR) coatings (Chapter **??**) and/or polarising filters (Chapter 7).

1.4.2 Spectroscopy

Spectroscopy concerns the spectral analysis of light. Many different techniques exist for achieving this, including diffraction gratings, prisms and resonant cavities.

Chemical analysis

Different chemical substances may often be identified by discrete peaks in their emission or absorption spectra. Typically, a substance may be heated to a gaseous phase and white light shone through it. Analysing the spectrum of the light in the direction of the light may reveal dark lines or *absorption resonances*, where the energy, and hence frequency, of the photons matched a difference in the electron energy levels of the chemical. In other directions, bright lines may be seen where these excited states drop down to lower energy levels, emitting a photon in the process. This absorption and emission lines are typically unique for a given chemical substance, so that they provide a chemical signature.

Astronomy

Although modern astronomy exploits all wavelengths of the electromagnetic spectrum, optical telescopes were of crucial importance in its development. Moreover, spectroscopic analysis of the light can provide information about the temperature of stars (see Chapter 2) and the chemical make-up of extra-terrestrial matter.

1.4.3 Photonics

Quantum mechanics spurred many of the technological revolutions of the 20th century. In particular, the development and fabrication of semiconductors via our understanding of solid state physics has lead directly to the inventions of the diode, transistor and integrated circuit. Whilst these are electronic devices, the optical properties of semiconductors (and other materials) have also been exploited to give us *photonic* devices.

A crucial aspect of semiconductors is that they have a *forbidden energy gap* between the energies of the electrons involved in molecular bonding (the *valence electrons*) and those free to take part in electronic or thermal conduction (the *conduction electrons*. No (real) electron states exist at energies between these two bands. This makes the optical properties of semiconductors highly useful.

Optical detectors and photovoltaic cells

The absorption of photons by electrons may be exploited in optical detectors or photovoltaic cells. The basic mechanism is that of *electron-hole pair creation*. When an electron in the valence band absorbs a photon of the right energy, it is excited into the conduction band, where it acts as a *negative charge carrier*. This leaves behind an ‘empty state’ or *hole*. Somewhat counter-intuitively, the hole then acts as a *positive charge carrier*.

Depending on the design of the device, this can either make the semiconductor conductive, allowing a *photocurrent* to flow or create a voltage that can be used to provide electrical power. The former case is called *photoconductivity*. The latter is the case for *photovoltaic cells* (i.e. *solar cells*).

Light emitting diodes

The emission of photons via electrons dropping down into lower energy states is utilised in both *light emitting diodes* (LEDs) and *semiconductor lasers*. The major difference between these two types of device is that for LEDs, the emission is *spontaneous*, whereas in a laser, we have *stimulated emission* (see Chapter 2).

The common feature for semiconductor devices is that the excited energy levels may be filled *electronically* (note in Chapter 2 we briefly discuss the *optical pumping* of the higher energy levels in a laser).

Unlike laser light, since the emission is spontaneous in an LED there is no fixed phase relation between the photons. The light is therefore *decoherent*.

Lasers

Lasers in general are discussed briefly in Chapter 2. Semiconductor lasers have found many applications, such as in fibre optic communications, reading CD and DVD drives and scanning technologies. A major advantage of such devices is their small size, cheapness and versatility. For instance, a block of semiconducting material naturally acts as a resonant cavity and may be readily adapted with Bragg reflectors (Chapter **??**) to provide the correct wavelength selectivity.

1.4.4 Fibre optics

Modern telecommunications exploits many regions of the electromagnetic spectrum. However, for long-haul communications networks, the backbone is provided by *fibre-optics*. With rising use of the internet for both download and uploading of data, these systems require a large communications bandwidth, and fast and efficient multiplexing and modulation of the carrier signals. This is achieved primarily through optical technologies. Note that we include the near infrared in the optical spectrum, with the main transmission window for silica fibre having a wavelength of around 1.5 μm.

Waveguiding

A fundamental optical application in fibre optic systems is that of *waveguiding*, the process by which the optical signal is contained within the fibre (or other optical waveguide). The basic physical principle involved here is that of *total internal reflection* (Chapters 4 and 8), although a rigorous treatment of the problem requires a solution of Maxwell's equations (Chapter 6).

Optical modulation

Another important aspect of fibre optic communications is the question of how to modulate the light. In practice, one of the best ways of achieving this is to modulate the light *after* it has been generated by exploiting the electrooptic effect. This is an effect in which the electrical susceptibility of a material is changed by the application of an electric field (taken to be static on the time scale of the optical oscillations). Both linear and nonlinear effects exist and are used.

Optical amplifiers

When transmitting light over many miles of fibre optic cable, optical attenuation becomes a problem. In these circumstances, some form of amplification must be used to boost the signal. Such amplifiers are essentially lasers operated under the conditions that they only emit when a signal is passed through them. One possible technology is the *semiconductor optical amplifier* (SOA). However, the preferred technology is at present the *erbium doped fibre amplifier* (EDFA). This is a special fibre doped with erbium providing energy levels for laser operation.

Wavelength converters

One use of SOAs is in there use as wavelength converters for *wavelength division multiplexing* (WDM) systems. In such devices, a signal of one wavelength may create a population inversion for the amplification of a signal of another.

Nonlinear effects

Due to the high intensity of laser light confined to the small cross-sectional area of the fibre optic core, nonlinear effects in fibre optics often become significant. One of the most important of these is the third order effect known as *four wave mixing*, in which three frequency components of the electrical polarisation combine to produce a fourth. Although we mention

nonlinear optics at several points in the text, a proper study of these effects is beyond the scope of this course.

1.4.5 Computer graphics

In some senses, the technology of *computer graphics* or *computer generated imagery* (CGI) may be thought of as bringing the entirety of optical understanding together, since the purpose is often to simulate reality as accurately as possible. Since the subject also brings in a hefty dose of computer science and computational physics, it is a huge topic in its own right. Here we mention a couple of aspects of the technology.

Ray tracing

In optical texts, *ray tracing* usually refers to a technique of approximating the wave nature of light by following an optical ray in the propagation direction of the wavefront. It is often used to analyse the focussing of light through lenses or reflected from mirrors in *geometrical optics*.

In computer graphics, the term has a closely related meaning but usually refers to the rendering of a 2D screen by following rays into a 3D virtual space. This involves various aspects including

- the perspective projection of the virtual 3D image on to the 2D screen
- the reflection or refraction of a ray
- the diffuse scattering of a ray

All of these areas require a firm understanding of optics.

Scattering theory (rendering algorithms)

Scattering theory is an area that becomes increasingly complicated as one delves into it. Fundamentally, the way in which light is scattered from an object is straight forward. The incoming light induces electric dipole oscillations in the material medium which are then re-radiated. The details of these oscillations for a given material, however, may become very complicated. Nevertheless, modern CGI attracts a high level of financial investment and, as a result, techniques have become very sophisticated.

1.5 Course overview

The course is divided into five parts.

I **Introduction to Optics**

A general introduction to optics, including a brief history of its development and discussion of the complementary understandings provided by electromagnetism on the one one hand and quantum mechanics on the other. Since the course material focusses on the electromagnetic picture, a brief introduction to the physics of waves is also given, with the emphasis and the universality of the mathematical description of waves. Thus, the same form of wave equation that is used for electromagnetism can also be used to model the elastic vibrations of a continuous solid. A first categorisation of media is also given:

- Linear / nonlinear
- Isotropic / anisotropic
- Homogeneous / inhomogeneous

In this section, we describe linearity mathematically in terms of linear and nonlinear differential equations (wave equations in our case). Since we also introduce the general concepts of *polarisation* and refractive index, we also explain isotropy as the independence of the refractive index on polarisation direction. In crystal optics, this is often not the case and we have an *anisotropic* medium.

II **Wave Optics**

In the section of *wave optics* we introduce the powerful concept of the Huygens-Fresnel Principle, based on Huygens' Principle with the additional concepts of interference worked out in such detail by Fresnel. Using this principle, we obtain the Laws of Optical Propagation

- The Law of Rectilinear Propagation (light travels in straight lines)
- The Law of Reflection
- The Law of Refraction (Snell's Law)

This is applied to the important topic of *diffraction*, a characteristically wave-like phenomenon and providing (via Young's double slit experiment) the most convincing evidence for the wave description of light. Of practical importance for spectroscopic analysis is the *diffraction grating*, which is also introduced.

III **Electromagnetic Waves**

The central exposition of optics given in this text revolves around Maxwell's electromagnetic equations. We begin this topic with a discussion of EM wave propagation in a both a vacuum and a material medium. Here we meet the crucially important topic of the *electric*

susceptibility tensor, which gives the *frequency response* of a medium to an applied electric field. This response manifests in the form of the *electrical polarisation* of the medium. We spend quite some time discussing the susceptibility tensor and its frequency dependence (which gives rise to a frequency dependent refractive index). Much of this discussion is to provide a firm basis for our later description of *crystal optics*.

We then move on to discuss the *optical polarisation* (not to be confused with the *electrical polarisation* of a medium). Here, we introduce the formal apparatus of the *Jones vector* to describe the polarisation and the *Jones matrix* to describe the effect on the polarisation of an optical element. Initially, these concepts are dealt with formally in the main, awaiting the section on *crystal optics* for a full explanation of the physics.

Light propagation between different media is also covered at some length. This is begun with coverage of *Fresnel's equations*. Using the understanding we glean from this, we then look at the technologically important area of *thin-film interference*.

IV **Crystal Optics**

In this section, we consider wave propagation in *anisotropic* materials. We begin with a short account of group theory and crystal symmetry to provide both a means of categorisation and a powerful analytic tool for determining the optical properties of particular crystal systems. This is achieved by applying symmetry arguments to the form of the susceptibility tensor.

Finding the solutions for wave propagation in an anisotropic medium, we encounter the highly useful concept of the *index ellipsoid*. Using this, we may analyse simple cases of wave propagation. When we do, we encounter the phenomenon of *birefringence*, in which orthogonal components of the optical field can move out of phase with each other. Birefringence can be exploited to change the polarisation state of light from, say, linearly polarised to being elliptically polarised. Technologically, this allows the construction of optical elements known as *wave plates* (considered formally earlier in the chapter on polarisation). Wave plates are frequently used in photonic applications, such as optical modulators and in 3D glasses, to name just two possibilities.

Closely related to birefringence is the concept of *pleochroism* (or more often the limiting case of *dichroism*. This is the selective *absorption* of light depending on polarisation direction. One of the main uses of dichroism is in *linear polarisers*. In more aesthetic applica-

tions, dichroic glass is often used in jewellery due to the deep and varied colours it produces.

A further class of crystal optics concerns *optically active* materials in which a linear state of polarisation undergoes a rotation about its propagation axis. A well-known case of this is *Faraday rotation*, in which a magnetic field induces the rotation. This effect has technological applications in the fabrication of optical isolators for fibre optic systems as well. It is also occurs naturally in, for instance, light from astronomical objects passing through strong magnetic fields.

V **Geometrical Optics**

In the final section, we turn to *geometrical optics*. In some ways, this may be considered as far more elementary than the other topics covered in this course but it is as a result of that understanding that we can see how the assumptions of geometrical optics may be justified (and its limitations).

Here, our theoretical underpinning is that of *Fermat's Principle*. This is a remarkably powerful tool and has a firm theoretical justification. The main limitation of geometrical optics is that it employs ray-tracing throughout, which, although a very useful concept, remains an idealism due to the effects of diffraction. Nevertheless, we are again able to derive the Laws of Optical Propagation on the basis of Fermat's Principle, which is expressed in terms of the *optical path length*.

With the theoretical foundations established, we then move on to analyse both *perfect* imaging (in which light is perfectly imaged onto a point or plane and imaging via spherical mirrors and lenses. The *paraxial* approximation for small angles is introduced, which makes the analysis of thin lenses mathematically tractable. We also see how lens may be combined and how images are constructed.

In addition, we also consider the types of optical aberrations encountered and strategies for alleviating them.

1.6 References

[1] James Clerk Maxwell, *A Dynamical Theory of the Electromagnetic Field*, Philosophical Transactions of the Royal Society of London **155**, 459-512 (1865)

[2] Heinrich Hertz, *Untersuchungen über die Ausbreitung der elektrischen Kraft*, Johann Ambrosius Barth, Leipzig (1892)

[3] Albert Einstein, *Zur Elektrodynamik bewegter Körper, Annalen der Physik **17**, 891 (1905)*

[4] Max Planck, Annalen der Physik **309**, 564566 (1901)

[5] Albert Einstein, *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt*, Annalen der Physik **17**, 132148 (1905)

[6] Niels Bohr, *On the Constitution of Atoms and Molecules*, Philosophical Magazine **26**, 1-25 (1913)

Additional reading

[7] Grant R. Fowles, *Introduction to Modern Optics*, Dover Publications (1989)

[8] Eugene Hecht, *Optics*, Addison-Wesley (2001)

[9] Max Born and Emil Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, Cambridge University Press (1999)

2. Waves and Photons

2.1 General remarks

Throughout this work we shall focus on the classical description of light provided by Maxwell's equations. A complimentary picture of the subject that we shall often connect with is provided by quantum mechanics. These two explanations of the same phenomena may seem, at first, contradictory. On the one hand we have a classical wave picture of light and on the other something that looks (at first) like a classical particle picture in terms of photons.

Photons, however, are *not* classical particles. They remain quantum mechanical entities that may sometimes, at our convenience, be visualised as being particle-like. It is important to remember, however, that this is just a conceptual handle used to get a grip of the nature of photons. In fact, the quantum mechanical picture remains consistent with and complementary to classical electromagnetics.

We shall, therefore, dip a toe in quantum mechanics to get a feel of the water. We may then draw on these insights as we need throughout the predominantly classical exposition. For our purposes, the most important of these will be *the energy of photon*, the categories of the electron-photon interaction and photon description of optical absorption. Along the way, we encounter *Planck's Law* for spectral energy distribution of blackbody radiation, as well as the related *Stefan-Boltzmann* law.

The emission and absorption of photons in terms of electronic transitions is described in terms of *Bohr's model of the atom*. Although this model is not strictly correct, the physics relevant for our purposes is and the model requires less formal apparatus than solutions of the *Schrödinger equation*.

This description of electromagnetic generation is also of importance for the purposes of categorising the optical spectrum. In particular, we define optical radiation to be that produced via electronic transitions. Beyond the limits of this spectrum, the physical origin of very short wavelength radiation may be taken to be, for instance, *nuclear transitions* in the case of γ rays.

2.2 Learning objectives

- *The electromagnetic spectrum* and the classical description of light

- Definition of the *optical spectrum*
- The quantum mechanical description of light
  - Blackbody radiation
  - The energy of a photon
  - The Bohr model of the atom
  - The de Broglie wavelength
  - The electron-photon interaction
    - spontaneous absorption
    - spontaneous emission
    - stimulated emission
  - Heisenberg Uncertainty Principle
    - vacuum fluctuations
- Optical sources of radiation
  - incandescent sources
  - light emitting diodes
  - lasers
- Optical absorption

2.3 The electromagnetic spectrum

2.3.1 Maxwell's Rainbow

Although Thomas Young's double slit experiment of 1803 offered what appeared to be incontrovertible evidence for the wave nature of light, an explanation of the underlying physics was not forthcoming until Maxwell's proof [1] that the modified equations of electromagnetism yielded a *wave equation* for the propagation of electric and magnetic fields. Moreover, Maxwell showed that these waves would travel with the speed of light, which in turn could be calculated from fundamental constants. This prediction was decisively confirmed by Hertz [2], who generated *radio waves* from oscillating electric charge, having all the required characteristics of Maxwell's electromagnetic waves (EM waves).

In Fig. 2.1, we illustrate the spectrum of electromagnetic radiation (sometimes referred to as *Maxwell's rainbow*), from long wavelength radio waves

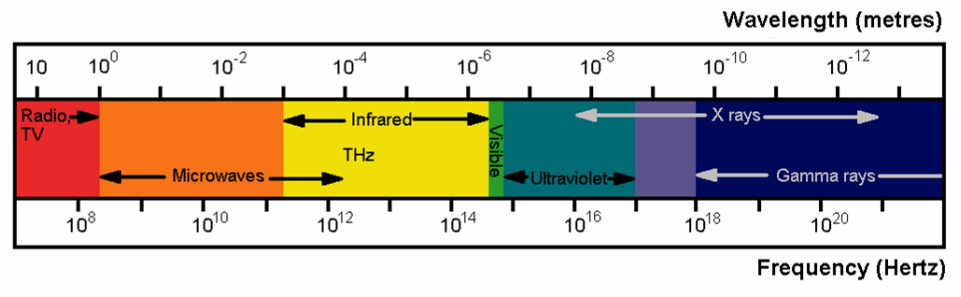


Figure 2.1: The electromagnetic spectrum in terms of wavelength and frequency, ranging from radio frequency (RF) to gamma rays (γ-rays).

(RF - standing for *radio frequency*) to extremely short wavelength *gamma rays* (γ-rays). The region of this spectrum typically viewed as the *optical spectrum* centres on the narrow band of *visible light*. We shall shortly introduce a criterion for the optical spectrum which would define the range to be from the near infra-red (NIR) to the near ultraviolet (UV).

2.3.2 RF radiation

At the low frequency end of radio frequency (RF) radiation, we have the ranges of *extremely low frequency* (ELF), from about 3 Hz to 3 kHz, and *very low frequency* (VLF) between about 3 kHz to 30 kHz. Such radiation is often bounded within the Earth's atmosphere as a *standing wave* with the surface of the Earth and the *ionosphere* forming a *resonant cavity* (the ionosphere is the region of the atmosphere between 85 and 600 km above sea level in which the atmospheric molecules are largely ionized by bombardment by solar radiation). Natural sources of this frequency of radiation include the movement of charge associated with lightening strikes. Extraterrestrial sources of ELF also include astronomical objects such as *neutron stars* with very high magnetic fields.

Moving towards higher frequencies, we have *long wave* and *medium wave* radio. These frequencies my be *amplitude modulated* for carry acoustic signals (i.e. *AM radio*). Above this in the very high and ultra high frequencies (VHF and UHF), signals may be *frequency modulated* (i.e. *FM*) with the higher frequencies traditionally used from television broadcasting.

Radio waves are typically generated via oscillating charge in an *antenna* of some kind (the most well known type being the *dipole antenna*. These oscillations are driven by electronic circuitry, typically an electronic oscillator coupled to a power amplifier.

2.3.3 Microwave radiation

Above radio waves, the next major band of the electromagnetic spectrum is dubbed *microwaves*, lying between about 0.3 GHz and 300 GHz. This range of radiation is often generated by very fast charge oscillations in devices such as the magnetron (often used to power microwave ovens) or, in solid state devices via the *Gunn diode*. Microwaves are also widely used in free-space telecommunications.

2.3.4 The need for quantum mechanics

Engineered RF and microwave generation relies on electronic circuitry of some kind to produce the charge oscillations leading to EM radiation. However, despite the high speed of modern electronics, all such circuits are limited by what is known as the *RC time constant*. This is a time constant describing how fast the circuitry can actually respond and is due to both the resistance R and capacitance C of the electronics. Since all such circuits will have at least parasitic resistance and capacitance, the RC constant is always finite and acts to limit the high frequency response of such technology.

In the near infra-red (NIR) and visible (i.e. the *optical* frequencies we are concerned with in this work), the limitations of electronics may be overcome by exploiting the *optical* properties of materials, i.e. the interaction of light with matter. This technology is known as *photonics* and incorporates, for instance, the physics of lasers, solar cells, light emitting diodes and so on. As we shall see, the extremely fast charge redistributions leading to optical generation are due (mainly) to electronic transitions between electron levels in materials.

Understanding these electronic levels necessitates the application of quantum mechanics. Moreover, it is often far easier to visualise the optical interactions occurring at this fundamental level using the *photon* picture, in which a photon is imagined to have close similarity to a classical particle.

In fact, light (as well as matter) retains wavelike properties even at this scale and Maxwell's account of electromagnetic generation and propagation remains valid. Photons are not quite the 'corpuscles' that Newton envisaged light to be composed from but quantum mechanical entities entailing a wave-like nature via *Heisenberg's Uncertainty Principle*. A modern view of a 'particle', then, is that of a *localised wave packet*. Such a wave packet can be constructed, via Fourier's Theorem, from a superposition of waves of well-defined wavelength.

For this reason, we may justifiably focus on the classical treatment of light in an exposition of modern optics, pausing here and there to indicate the connection to the underlying quantum theory. Hence, we review briefly the rise of the quantum picture, which first emerged with Max Planck's ef-

forts to understand *blackbody radiation* [3]. Armed with this additional understanding, we shall be in a better position to offer insight into the nature and generation of the thermal infra-red, optical and higher frequency radiation.

2.4 Blackbody radiation

A blackbody is defined as an ideal body that absorbs all incident radiation incident on it and radiates at all spectral frequencies with an energy distribution dependent on the absolute temperature T of the body. Despite this abstract definition, there are in fact physical systems that very closely approximate a blackbody radiator. A fairly common example would be an aperture into an oven or kiln of some kind. For instance, an experienced potter might judge the temperature inside a clay firing kiln on the basis of the colour of the light from it. This is due to the spectral energy density of the radiation having a maximum at a characteristic wavelength (or frequency) characteristic of the temperature. As the temperature increased, the colour would vary from a dark red through orange and on to white (actually a mixture of light that we perceive as white). Such temperatures are therefore referred to as 'red hot' and 'white hot'. If the temperature could be allowed to increase further, the colour would start becoming blue - i.e. the kiln was then 'blue hot'. Thus, even though we typically think of blue as a 'cool' colour, in fact it is physically associated with very high temperatures.

2.4.1 Stars as blackbodies

Another example of a blackbody radiator is a *star*, such as our Sun. The spectrum of stars then allows us to determine the temperature of the star's 'photosphere' - i.e. the outer limit of what can be visually observed of the star before it becomes opaque to radiation. The Harvard spectral classification of stars associates with different temperature ranges the classes 'O,B,A,F,G,K,M' ('Oh Boy An F-Grade Kills Me'), with 'O' representing the hottest and 'M' the coolest. Table 2.1 then shows the correspondence between temperature and colour.

Thus, our Sun is a G-type star with a photospheric temperature between 4500 and 6000 K. On the other hand, Rigel, the brightest star in the constellation of Orion, is a B-type star with a temperature of between 11000 and 12000 K.

2.4.2 Planck's Law

The problem that Planck set out to resolve was that the spectral profile of blackbody radiation could not be modelled by existing laws. According to

Table 2.1: . Harvard spectral classification (L,T and Y classes omitted).

Class	T (K)	Colour
O	≥ 33000	blue
B	$10000 - 33000$	blue-white
A	$7500 - 10000$	white
F	$6000 - 7500$	yellow-white
G	$5200 - 6000$	yellow
K	$3700 - 5200$	orange
M	$2400 - 3700$	red

the *equipartition theorem* of statistical mechanics, each vibrational mode of the system should take up an average energy $k_B T$, where k_B is *Boltzmann's constant*. At the same time, however, the cavity would support an increasing number of modes of smaller wavelength (and hence larger frequency), each of which would take on the average thermal energy. This would then imply a runaway effect in which the radiated power becomes unlimited as the wavelength tends to zero - a conclusion known as the *ultraviolet catastrophe*.

In fact, Planck was able to infer an *empirical* distribution that matched the observed spectral density. However, he ran into problems when he attempted to justify his formula theoretically. Planck employed the mathematical trick of first assuming that the modes of the cavity were emitted in quanta. He then intended to take the continuum limit via calculus. In the event, this proved to be an impossible task and he was left to conclude that the radiation was indeed *emitted in quanta*. The energy of these quanta is proportional to the frequency and given by

$$\boxed{\epsilon = hf,} \tag{2.1}$$

where f is the frequency and h is Planck's constant. This can be written in terms of the angular frequency ω using the notion introduced by Dirac $\hbar = h/(2\omega)$, so

$$\boxed{\epsilon = \hbar\omega.} \tag{2.2}$$

This meant that the spectral density of a blackbody could now be given in terms of *Planck's Law*. As a function of frequency and temperature, the intensity I is given by

$$\boxed{I\left(f,T\right)=\frac{2hf^{3}}{c^{2}}\frac{1}{e^{hf/k_{B}T}-1},} \tag{2.3}$$

where c is the speed of light in a vacuum.

We can derive the relation between the frequency at the maximum of Eq. (2.3) and temperature T by putting

$$x=\frac{hf}{k_{B}T} \tag{2.4}$$

and finding

$$\frac{dI\left(x\right)}{dx}=0. \tag{2.5}$$

This yields

$$x\left(e^{x}-1\right)-xe^{x}=0, \tag{2.6}$$

which must be solved numerically (e.g. using the bisection method). Note the solution at $x=0$ is *not* the maximum and should be avoided. To 6 DP, we then find a value of $x_{\max}=2.821439$. From Eq. (2.4) we then find

$$\frac{f_{\max}}{T}=58.8\ \mathrm{GHz\ K^{-1}}. \tag{2.7}$$

In terms of wavelength, we have $f\lambda=c$, so

$$\lambda_{\max}=\frac{c}{f_{\max}}=\frac{5.1\times10^{-3}\mathrm{m\cdot K}}{T}. \tag{2.8}$$

Note that this result is obtained by considering the *frequency* to be linear. A slightly different result emerges if we re-formulate the problem in terms of *linear wavelength*. In this case we get

$$\boxed{\lambda_{peak}=\frac{2.898\times10^{-3}\mathrm{m\cdot K}}{T},} \tag{2.9}$$

which is known as *Wien's Displacement Law*.

The *total radiative power* or *luminosity* L of a blackbody may be found by integrating Planck's Law for the intensity over both solid angle and frequency. The result is the *Stefan-Boltzmann Law*, which tells us that L is proportional to the fourth power of the absolute temperature.

The Stefan-Boltzmann Law may be written

$$\boxed{L=\varepsilon A\sigma T^{4},} \tag{2.10}$$

where ε is the *emissivity* (a measure of how well the object radiates), A is the surface area of the blackbody and σ is the Stefan-Boltzmann constant.

Use of blackbodies in metrology

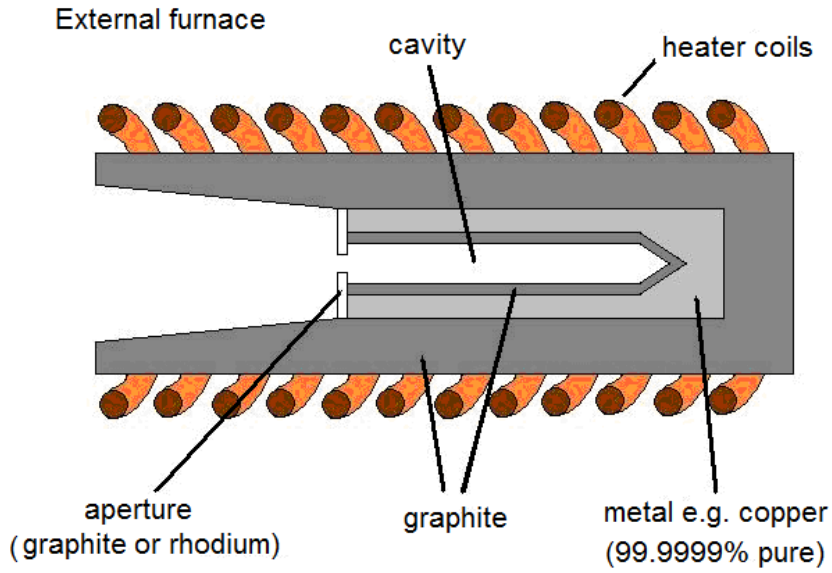


Figure 2.2: Generalised schematic of a fixed-point blackbody. The blackbody radiation is emitted from the central cavity, held at a constant temperature by the freezing of the very pure metal surrounding it.

Given the precise description of the spectral radiation from a blackbody in terms of fundamental constants (h and c), blackbodies are a prime candidate for the first link in the calibration chain for optical sources and detectors. In fact, at the National Physical Laboratory in the UK, where the SI unit of luminous intensity the *candela* is maintained, artificial blackbodies are used that can be held very precisely at a known temperature. A generalised sketch of a *fixed-point blackbody* (i.e. operating at a fixed reference temperature) is shown in Fig. 2.2.

As an example, the *copper point blackbody* (CuBB) operates by melting very pure copper surrounding the cavity itself and then cooling it below its freezing temperature. Due to the purity of the copper, it will under-cool without freezing as there are few seed sites to initialise the crystallisation. As it does freeze, the temperature rises back up to the freezing point temperature (1.358×10^3 K) where it remains stable for some time. By Wien's Law, the maximum emission at this temperature is at 2.134 μm.

2.4.3 The photoelectric effect

Planck's idea was later extended by Einstein in his explanation of the *photoelectric effect* [3] in 1905.

As discussed in Chapter 1, in the photoelectric effect, electrons are liberated from an illuminated sample of metal. However, whereas the classical theory would predict that the energy of these electrons should proportional to the intensity of the radiation, instead, the energy is related to the *frequency* of the light. Einstein explained this by assuming that the *radiation field itself was quantized*. He then formulated his explanation in terms of the maximum energy of the liberated electrons

$$\boxed{\epsilon_{\max} = hf - \phi,} \tag{2.11}$$

where ϕ is known as the *work function* of the metal. That is, the energy that the electron has to expend to be released from the metal. These 'light quanta' later became dubbed *photons*.

2.5 The Bohr model of the atom

In early quantum theory, the notion of quantisation introduced by Planck was adopted by Bohr in an attempt to explain the stability of the atom [5]. According to Rutherford's experiments firing alpha particles at gold film (an alpha or α particle is a Helium atom nucleus), most of the mass of an atom was concentrated in the positively charged nucleus. Negatively charged electrons were then held to this nucleus via Coulomb's electrostatic force. A naive interpretation of this would picture the electrons as particles orbiting the nucleus like planets around the Sun. However, because electrons are negatively charged, the centripetal acceleration required would cause the electron to *radiate*. Hence, the electron's energy would be rapidly radiated away and the particle would spiral catastrophically into the nucleus.

Bohr was able to resolve this state of affairs with a *partially correct* ad hoc theory, in which he postulated that the *orbital angular momentum* of the electrons was quantised in units of Planck's constant. This yielded the prediction of a series of discrete states which also had discrete energies. Transitions between orbits were then achieved via the emission or absorption of a photon of the correct energy difference (see next section for more detail).

Bohr's theory was also able to account for the discrete absorption and emission lines seen in atomic spectra. Such lines are routinely used as chemical identifiers to analyse chemical composition.

Despite these successes, Bohr's theory was only partially correct. The correct solution awaited Schrödinger's *wave mechanics*, the addition of *electron spin* and the *Pauli Exclusion Principle* (PEP). We shall leave detailed discussion of these topics to quantum mechanical texts.

2.6 The de Broglie wavelength

A great deal is made in the literature of the expression 'wave-particle' duality. This suggests a a paradoxical picture of things in which quantum entities are 'sometimes like waves and sometimes like particle'. This popularist interpretation is open to severe criticism, although there remains a sense in which this description contains meaningful physics.

Despite the two apparently contradictory explanations of light emerging around the turn of the century, there was no formal link between them. This comes from *de Broglie's hypothesis* in 1924 that matter also has a wavelike nature. Although principally applied to so-called 'matter waves' the theory is rooted in the physics of waves and Special Relativity.

According the Special Relativity, the energy and momentum of a particle travelling with a speed v are

$$\epsilon = \gamma m_0 c^2 \tag{2.12}$$

and

$$p = \gamma m_0 v, \tag{2.13}$$

where γ is the *Lorentz factor*

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \tag{2.14}$$

and m_0 is the *rest mass*.

In the meantime, wavelength λ is related to *phase velocity* v_p and frequency f via

$$\lambda = \frac{v_p}{f}. \tag{2.15}$$

Using Planck's energy relation Eq. (2.1), we then have

$$f = \gamma \frac{m_0 c^2}{h} \tag{2.16}$$

and

$$\lambda = \frac{h v_p}{\gamma m_0 c^2}. \tag{2.17}$$

Letting the phase velocity be given by

$$v_p = \frac{\epsilon}{p} = \frac{c^2}{v}, \tag{2.18}$$

we therefore have

$$\boxed{\lambda = \frac{h}{p}.} \tag{2.19}$$

Using Dirac's notation, we can write this in vector form as

$$\boxed{\mathbf{p} = \hbar\mathbf{k},} \tag{2.20}$$

where $\mathbf{k}$ is the *wavevector*.

2.7 The electron-photon interaction

Elementary quantum mechanics then gives us a picture of the transition of electrons between discrete energy states via the emission or absorption of a quantum of light - the *photon*. Planck had determined the energy of a photon as given by Eq. (2.2). The *emission* of a photon via the transition of an electron from a state with energy ϵ_2 to a state with lower energy ϵ_1 may then be expressed in terms of the conservation of energy as

$$\epsilon_2 \rightarrow \epsilon_1 + \hbar\omega, \tag{2.21}$$

where

$$\epsilon_2 - \epsilon_1 = \hbar\omega. \tag{2.22}$$

Conversely, a photon may be *absorbed* via the interaction

$$\epsilon_1 + \hbar\omega \rightarrow \epsilon_2 \tag{2.23}$$

2.7.1 Categories of emission and absorption

The ways in which light may be emitted or absorbed via the electron-photon interaction can be described by one of three categories:

1. *absorption* by which the energy of an incoming photon is absorbed by an electron, which is then excited to a higher energy state. The photon may be thought of as being *destroyed* by this process. Note that since electrons are fermions and, as such, subject to the PEP, there must be an empty higher state of the right energy for this processes to occur.

2. *spontaneous emission* by which an excited electron *spontaneously decays* to a lower energy state with the emission of a photon with the energy difference. This process may be viewed as one of photon *creation*. Again, the conditions of the PEP apply.

3. *stimulated emission* by which an excited electron is *stimulated* to emit a photon and decay to a lower energy state by the proximity of another photon with the same energy. In this process, the emitted photon and incident photon are *coherent*, having the same energy, phase and direction. This is then another photon creation processes.

   Stimulated emission was first predicted by Einstein on the basis of thermodynamic considerations. Although not an intuitively obvious process, a quantitative understanding of stimulated emission may be gained from *quantum perturbation theory* (See, for example, Dirac [6]).

2.7.2 Momentum and angular momentum

So far our simple picture of the electron-photon interaction gives a good account of photon energies, which, via Planck's relation of Eq. (2.2), applies equally to frequency. Now, since light also carries momentum, we can see how the same picture accounts for this. It is simple to imagine the electron transitions taking place between states with different *momenta*. The difference in these momenta is then just the momentum of the photon. Using the de Broglie relation Eq. (2.20), we have for *emission*

$$\hbar\mathbf{k}_2 \rightarrow \hbar\mathbf{k}_1 + \hbar\Delta\mathbf{k}, \tag{2.24}$$

where the momentum of the photon is

$$\hbar\Delta\mathbf{k} = \hbar\mathbf{k}_2 - \hbar\mathbf{k}_1. \tag{2.25}$$

a similar relation holds for absorption.

In chapter 7 we will discuss the *polarisation* of light in terms of the wave picture furnished by Maxwell's electromagnetic equations. We shall see that the polarisation corresponds to the direction of the electric field vector and that this, in a general case, may rotate around the axis of propagation as *elliptically polarised light*. We defer detailed discussion of this until chapters 6 and 7. For now, it is enough to conjecture that such waves carry a non-zero angular momentum.

In fact, we find that elliptically polarised light occurs in the photon picture as a consequence of transitions between electronic states with opposite *spin*. Introductory texts on quantum mechanics tell us that the spin of an electron is its quantized intrinsic angular momentum. Thus, we see how a photon may gain angular momentum as well as linear momentum and energy.

2.8 Heisenberg Uncertainty Principle

Much of the seemingly paradoxical nature of quantum mechanics may be traced back to *Heisenberg's Uncertainty Principle* (HUP), originally introduced in the context of *matrix mechanics*. Formally, this is a principle that holds between *non-commuting* observables. By this we mean if, say, you were to measure the values of two non-commuting variables, the result would depend on the *order of measurement*. This has an analogy in matrix multiplication, where the product of two matrices depends on which multiplies which.

A more intuitive picture is obtained from *wave mechanics*, which offers a deep insight into the resolution of the wave-particle paradox. As an example, let us consider the Uncertainty Relation between position and momentum.

We have seen that the momentum of a particle is related to wavevector via Eq. (2.20). Now it is one of the fundamental postulates that the squared modulus of a wave gives the probability density for finding the particle at a given point in space. That is, if the probability density is $p(\mathbf{r})$ and the wavefunction is $\Psi(\mathbf{r})$, then

$$p(\mathbf{r})\, d^3\mathbf{r} = |\Psi(\mathbf{r})|^2\, d^3\mathbf{r}. \tag{2.26}$$

In order that this yields a probability, the integral of this expression over all space must be finite. That is, normalising the integral to unity, we must have

$$\int |\Psi(\mathbf{r})|^2\, d^3\mathbf{r} = 1. \tag{2.27}$$

Suppose, now, that $\Psi(\mathbf{r})$ has some *exact* wavelength $\mathbf{k}$. In this case, the squared modulus of $\Psi(\mathbf{r})$ will have the same value everywhere and the integral over all space will be infinite. Thus, such a wavefunction *cannot represent a particle*.

The resolution of this problem is the construct a *localised wave packet* from $\mathbf{k}$-states. The more localised we make the packet, the greater the number of $\mathbf{k}$-states we require. Thus, improving our knowledge of the position, means losing information about the exact momentum. Formally, the relation between the uncertainty in position Δx and that in momentum Δp_x is given by

$$\boxed{\Delta x \Delta p_x \geq \frac{\hbar}{2}.} \tag{2.28}$$

Note that the concept of the wave packet *retains our particle-like description* within the scope of *wave mechanics*.

Another important example is the *energy-time uncertainty principle*

$$\boxed{\Delta\epsilon\Delta t \geq \frac{\hbar}{2}.} \tag{2.29}$$

This may be interpreted as saying that the energy of system may vary by an arbitrary amount over a small enough time period. It is this uncertainty principle that lies at the heart of *quantum tunnelling* and makes possible the fusion reactions fuelling the Sun.

2.8.1 Vacuum fluctuations

Another fascinating consequence of energy-time uncertainty is that of *vacuum fluctuations*. If the vacuum had an exact energy (for instance 0), this would be a violation of the HUP, so the actual energy is therefore not defined. This means that pairs of particles and anti-particles may pop out of the vacuum with any energy provided they appear and disappear again within a short enough time. Now the photon is its own antiparticle, so in terms of radiation, vacuum fluctuations are the appearance and disappearance of photon pairs. Since such particles are transient, we often refer to them as *virtual particles*.

Vacuum or quantum fluctuations are not theoretically unproblematic. The possibility of the spontaneous appearance of particles of *any* energy leads to the ungainly notion that the energy of the vacuum is infinite. This means that ad hoc methods have to be employed to *renormalise* the field to make it finite. These methods essentially involve 'adding an infinite constant' to the energy - a mathematically invalid procedure!

However, strong evidence exists for the the existence of vacuum fluctuations. For our current purposes, the fluctuations suggest an interpretation of *spontaneous emission*. Earlier, we stated that there were three distinct processes for the electron-photon interaction: absorption, spontaneous emission and stimulated emission. In fact, we may think of this as reducing to just two with spontaneous emission being emission *stimulated* by the vacuum fluctuations. According to this picture of things, each possible photon state of the vacuum may be called a *mode* and the vacuum fluctuations as a *single* photon per mode.

Spontaneous emission is then the coupling of the process of emission with the rise from the vacuum of the single virtual photon state. Now, the process of emission can occur only in one way - from an excited state to a lower state having a difference in energy, momentum and spin equal to that of the virtual particle. On the other hand, the virtual particle may emerge as the end result of a very large number (approaching infinity) of different processes. Thinking of this in terms of probability, the entropy change involved with such emission will be extremely large. In other words, this is an *irreversible process*.

2.9 Thermal infra-red

The discussion of blackbody radiation did not broach the particular mechanisms by which the electromagnetic waves are generated. The *far*, *mid wavelength* and *short wavelength infra-red* (FIR, MWIR and SWIR) regions of the EM spectrum between about 300 GHz to 100 THz are often termed *thermal radiation*. This range of radiation is generally associated with the kinetic energies of material bodies, where the emitted EM radiation is due to energy loss in inelastic collisions. This may involve changes in the linear kinetic energy or in the vibrational energy of atomic oscillations.

The average energy exchanged in such interactions is the thermal energy $\frac{1}{2}k_BT$, giving a definition of the temperature. This then corresponds to the peak of the spectral energy distribution. However, there is also a large spread of kinetic and vibrational energies, giving a wide distribution of radiation.

2.10 Optical sources

2.10.1 Incandescent sources

The kind of emission just described is not, of course, restricted to thermal radiation. In fact, this mechanism of EM production is a principle source of optical frequency radiation. A typical example is the *incandescent source* such as the heated filament of a light bulb. In this case, an electric current heats the wire, which then emits in the infra-red and optical. By 'incandescent', we mean *decoherent* light. In other words, light for which there is no relative phase between the components from different sources.

2.10.2 Light emitting diodes

Another decoherent source of light is the *light emitting diodes* (LEDs). This is a semiconductor device in which the emission of photons is via electrons dropping down into lower energy states. However, unlike the incandescent source, this is not a thermal source of radiation, with the frequency of emitted light lying in a relatively narrow band. The LED is therefore far more efficient at converting the driving input energy (an electric current) into the required frequency range. Hence, visible light LEDs lack the hot infrared emission associated with incandescent bulbs and feel cool.

LEDs are have similarities to *semiconductor lasers*. The major difference between these two types of device is that for LEDs, the emission is *spontaneous*, whereas in a laser, we have *stimulated emission* (see Chapter 2).

The common feature for semiconductor devices is that the excited energy levels may be filled *electronically* (note in Chapter 2 we briefly discuss the *optical pumping* of the higher energy levels in a laser).

2.10.3 Lasers

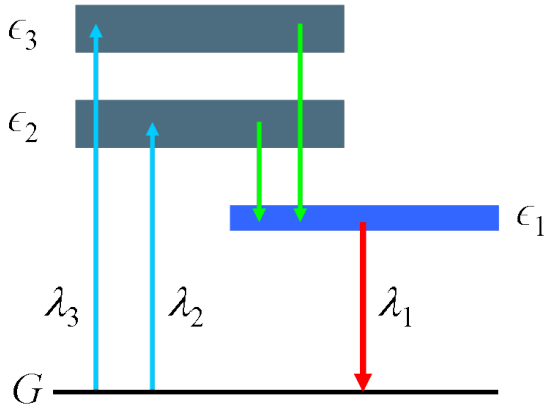


Figure 2.3: Energy level diagram for a ruby laser (a *three-level system*). The chromium ions are pumped optically from the ground state G to excited states ϵ_2 and ϵ_3 at wavelengths 5500 Å and 4000 Å via discharge from a xenon flash lamp. The excited states ϵ_2 and ϵ_3 are very short-lived, with lifetimes of $\tau \sim 10^{-8}\ \mathrm{s} - 10^{-9}\ \mathrm{s}$. These then decay very quickly to the metastable state ϵ_1 ($\tau \sim 3\ \mathrm{ms}$). The lasing transition then takes place from this level back down to ground, with a wavelength of 6943 Å.

Lasers, on the other hand, produce extremely coherent light. The physical mechanism underpinning such devices is that of *stimulated emission* (the word 'laser' is an acronym for *light amplification by stimulated emission of radiation*. Very briefly, a laser three conditions for its operation

1. A *population inversion* in which electrons are pumped up to an energy level ϵ_2, sitting above the final level ϵ_1 that the electrons are to drop down to. This means that a laser must have *at least three levels*: the highest level ϵ_3 that electrons are pumped to in addition to ϵ_1 and ϵ_2. Moreover, we require that the lifetime of the ϵ_3 should be very short so that ϵ_2 may be populated rapidly. In addition, the lifetime of ϵ_2 needs to be long so that it is not depopulated by spontaneous emission before the process of stimulated emission takes place (see Fig. 2.3 showing the energy level diagram of a ruby laser).

2. A *feedback mechanism* such as mirrors at the ends of an optical cavity that allow light (of energy $\epsilon = \epsilon_2 - \epsilon_1$ to travel back and forth, stimulating emission of the ϵ_2 level on each pass.

3. *seed noise*, always present due to vacuum fluctuations, initialising the lasing process.

2.11 Optical absorption

Later, in Chapter 6, we shall discuss the mechanism of optical loss in a medium via the classical electromagnetic theory. Here, we shall offer an alternative model yielding exactly the same results based on the photon picture.

Let us define the probability $\bar{p}$ that a photon will be absorbed over an infinitesimal distance dx within a material as

$$\bar{p} = \alpha dx. \tag{2.30}$$

The probability p that the photon is *not* absorbed at $x + dz$ is then

$$p\left(x + dx\right) = p\left(x\right)\left(1 - \alpha dx\right). \tag{2.31}$$

Rearranging and taking the limit as $dx \to 0$, we have

$$\frac{dp}{dx} = -\alpha p\left(x\right), \tag{2.32}$$

which, with the initial condition $p\left(0\right) = 1$, has the solution

$$p\left(x\right) = e^{-\alpha x}. \tag{2.33}$$

So the probability that the photon *is* absorbed at the end of this flight is given by

$$P\left(x\right) dx = e^{-\alpha x}\alpha dx. \tag{2.34}$$

($P\left(x\right)$ is a *probability density function*). Integrating over all distances,

$$\int_0^\infty P\left(x\right) dx = \int_0^\infty e^{-\alpha x}\alpha dx = 1, \tag{2.35}$$

as required. Multiplying $P\left(x\right)$ by x and integrating again, we may find the *average distance travelled*

$$\begin{aligned} \left\langle x\right\rangle &= \int_0^\infty xP\left(x\right) dx = \int_0^\infty xe^{-\alpha x}\alpha dx, \\ &= -\left[xe^{-\alpha x}\right]_0^\infty + \int_0^\infty e^{-\alpha x} dx = \frac{1}{\alpha}. \end{aligned} \tag{2.36}$$

Thus another interpretation of α is that $1/\alpha$ is the average path length of a photon. The quantity α is known as the *absorption coefficient* and we shall see that exactly the same exponential decay of the electromagnetic intensity is predicted by the classical theory.

Table 2.2: . Constants introduced in this chapter.

Name	Symbol	Value
Boltzmann	k_B	$1.3806488 \times 10^{-23}$ JK^{-1}
Planck	h	$6.62606957 \times 10^{-34}$ Js
Dirac	$\hbar$	$1.054571726 \times 10^{-34}$ Js
Stefan-Boltzmann	σ	5.67×10^{-8} Js^{-1}m^{-2}K^{-4}

2.12 X-rays and γ-rays

Here we briefly mention the very short wavelength end of the EM spectrum. Whilst ultraviolet (UV) is often taken to be part of the optical spectrum, X and γ-rays may be distinguished by the method of generation. Whilst radiation in the optical spectrum originates from charge transfer between *electronic* energy levels in materials, high energy γ rays typically originate from energy transitions within the *nucleus* of an atom. Some frequencies of X-rays may also be generated in this way, although usually we define X-rays to produced in other ways, such as *Bremsstrahlung* ('braking radiation') produced when electrons are rapidly decelerated by positively charged nuclei in a material.

Both X-rays and γ-rays have a plethora of applications, particularly in medical physics in radiography and radiotherapy.

2.13 Summary

- *The electromagnetic spectrum* and the classical description of light
- Definition of the *optical spectrum*
  Radiation arising from electronic transitions within a material
- The quantum mechanical description of light
  - Blackbody radiation
    * The energy of a photon

$$\boxed{\epsilon = hf.} \tag{2.37}$$

* Planck's Law

$$\boxed{I\left(f,T\right)=\frac{2hf^{3}}{c^{2}}\frac{1}{e^{hf/k_{B}T}-1}.} \tag{2.38}$$

* Wien's Displacement Law

$$\boxed{\lambda_{peak}=\frac{2.898\times10^{-3}\mathrm{m}\cdot\mathrm{K}}{T}.} \tag{2.39}$$

* Stefan-Boltzmann Law

$$\boxed{L=\varepsilon A\sigma T^{4}.} \tag{2.40}$$

– The Bohr model of the atom

Angular momentum of electronic states quantised, leading to discrete energy levels.

– The de Broglie wavelength

$$\boxed{\lambda=\frac{h}{p}.} \tag{2.41}$$

$$\boxed{\mathbf{p}=\hbar\mathbf{k}.} \tag{2.42}$$

– The electron-photon interaction

* absorption
  The energy of an incoming photon is absorbed by an electron, which is then excited to a higher energy state.
* spontaneous emission
  An excited electron *spontaneously decays* to a lower energy state with the emission of a photon with the energy difference.
* stimulated emission
  An excited electron is *stimulated* to emit a photon and decay to a lower energy state by the proximity of another photon with the same energy. In this process, the emitted photon and incident photon are *coherent*, having the same energy, phase and direction.

  - Heisenberg Uncertainty Principle

$$\boxed{\Delta x \Delta p_x \geq \frac{\hbar}{2}.} \tag{2.43}$$

$$\boxed{\Delta \epsilon \Delta t \geq \frac{\hbar}{2}.} \tag{2.44}$$

    * vacuum fluctuations

- Optical sources of radiation
  - incandescent sources
  - light emitting diodes
  - lasers

- Optical absorption

  Absorption coefficient α. $1/\alpha$ is the average path length of a photon through a material.

2.14 References

[1] James Clerk Maxwell, *A Dynamical Theory of the Electromagnetic Field*, Philosophical Transactions of the Royal Society of London **155**, 459-512 (1865)

[2] Heinrich Hertz, *Untersuchungen über die Ausbreitung der elektrischen Kraft*, Johann Ambrosius Barth, Leipzig (1892)

[3] Max Planck, Annalen der Physik **309**, 564566 (1901)

[4] Albert Einstein, *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt*, Annalen der Physik **17**, 132148 (1905)

[5] Niels Bohr, *On the Constitution of Atoms and Molecules*, Philosophical Magazine **26**, 1-25 (1913)

[6] P.A.M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, USA (1982)

3. The Physics of Waves

3.1 General remarks

Physics is about recognising patterns. The natural world presents itself to us in myriad, complex guises. Making sense of this world requires abstracting from it the structural forms underpinning reality. The fact that such forms exist is what makes physics possible.

Perhaps the most profound aspect of the underlying logic of nature is that the same forms permeate many diverse and, superficially, unrelated areas. There is maybe no better example of this than that of *wave phenomena*. The mechanics of waves is of paramount importance in practically every area of physics, spanning classical mechanics, electromagnetism quantum physics and general relativity.

In this work, we are concerned with *optics* - a subject built almost exclusively on the physics of waves. Even when we venture beyond classical optics into the realm of quantum optics, wave mechanics is still fundamental. In this chapter, we shall attempt to capitalise on the universality of wave phenomena to connect our study of optics with concepts of general applicability. In particular, we would be hard pressed to find a more general introduction to wave phenomena than the *simple harmonic oscillator*. This will not be a mere philosophical exercise - we shall find that the mathematical concepts such an analysis yields will be of direct use with little or no modification.

Conversely, our studies of optics will similarly equip us to tackle the wave mechanics of other areas of physics. Throughout this course, we will often draw on models of linear oscillators from classical mechanics. However, perhaps of greater relevance are the analogies with the wave mechanics of quantum physics. Very often, the mathematical formulation of a wave is the same even when the interpretation differs.

3.2 Learning objectives

A principal objective of this chapter is to convey the general applicability of wave concepts to many different areas of physics concentrating on those topics that transfer directly into optics. The particular topics focused on will be

- *The simple harmonic oscillator*
- The relation between energy and the amplitude of a wave
- The origin of the general form of the *wave equation* for wave propagation in a linear medium
- The meaning of *linearity* and *nonlinearity*
- The concepts of *phase velocity* and *wavevector*
- *Plane wave solutions* of the wave equation
- The general concept of *polarisaton*
- The general concept of *refractive index*
- The categories of a physical medium
  - *Linear*
  - *Isotropic*
  - *Homogeneous*
- The meaning of *dispersion*
- The general concept of *group velocity*

3.3 The simple harmonic oscillator

3.3.1 The ideal spring

A ubiquitous concept within physics is that of the *simple harmonic oscillator*. This type of oscillator occurs, for instance, whenever some physical displacement in a field or material medium is resisted by a *restoring force* that is proportional to the displacement. Another way of putting this, which is of general significance, is that this restoring force is *linear* in the displacement.

Although modern physics informs us that light is not a displacement in a material medium (it was originally supposed that this was the case, the medium being called the 'ether'), we may still view light as being a displacement of a *field*. For the time being, and through much of this book, we shall refer to this as the 'optical field'. We may then apply much of the same intuition, and indeed mathematics, that we glean the elastic displacement of a physical material in mechanics.

A good deal of insight may be initially gained from a mechanical example. An elementary but universal concept is that of the *ideal spring*. By definition, an ideal spring is a mathematical model of a real spring to which *Hooke's law* applies. This just says that the restoring force is proportional to the extension of the spring. The constant of proportionality is then known as the *spring constant*, which we shall denote by K. The force exerted by the spring is then proportional to the displacement u

$$F = -Ku. \tag{3.1}$$

Applying Newton's second law of motion, we have

$$m\frac{d^2u}{dt^2} = -Ku. \tag{3.2}$$

This has the general solution

$$u\left(t\right) = A\sin\left(\omega t + \phi\right), \tag{3.3}$$

where

$$\omega = \left(\frac{K}{m}\right)^{1/2} \tag{3.4}$$

is the *angular frequency* and ϕ is a phase factor determining the position at time $t = 0$. Note that the angular frequency of Eq. (3.4) is characteristic of the system, being given in terms of the spring constant K and the mass m. In fact, this is the *resonant angular frequency* of the system and will usually be denoted by $\omega = \omega_0$. This, of course, may be expressed in terms of frequency f_0 and the oscillation period T_0

$$\omega_0 = 2\pi f_0 = \frac{2\pi}{T_0}. \tag{3.5}$$

3.3.2 Optical sources as harmonic oscillators

So how does the concept of the simple harmonic oscillator relate to optics? As we shall see in Chapter 6, light may be understood as an electromagnetic oscillation propagating through space. A possible *source* of such a disturbance is an oscillating *electric dipole*, in which we have two charges of opposite sign separated by some distance z. Although this is not the place for a detailed treatment of electric dipoles, we wish to argue for an approximate model of a dipole that may be treated as a harmonic oscillator.

Classically, the dipole is visualized as two point charges displaced from one another. In reality, however, in the absence of any other restraints, two such classical charges, q and $-q$, would be drawn together with an increasing force, given by *Coulomb's Law*

$$F = -\frac{q^2}{4\pi\epsilon_0 z^2}, \tag{3.6}$$

where the constant ϵ_0 is known as the *permittivity of free space* and z is, as defined above, the distance between the charges (we assume, for simplicity, that the material medium has a *relative* permittivity of unity). This is symptomatic of a general problem of classical physics in that it fails to explain the stability of matter composed, as it is, of equal amounts of negative and positive charge.

This problem is solved very elegantly in quantum theory, where the concept of the point-like particle evaporates and is replaced with a *wave* description of particles. For our present purposes, this means that we may imagine a negatively charged electron, not as a point particle, but rather a 'cloud' of charge, with the highest charge densities spread out over surfaces in space. For instance, an 's-type' electronic state can be imagined approximately as a sphere of charge.

The positively charged protons will also be smeared out in space, although, due to there greater mass, the uncertainty principle will allow them to be far more localized. We will therefore continue to imagine a proton as a point-like particle.

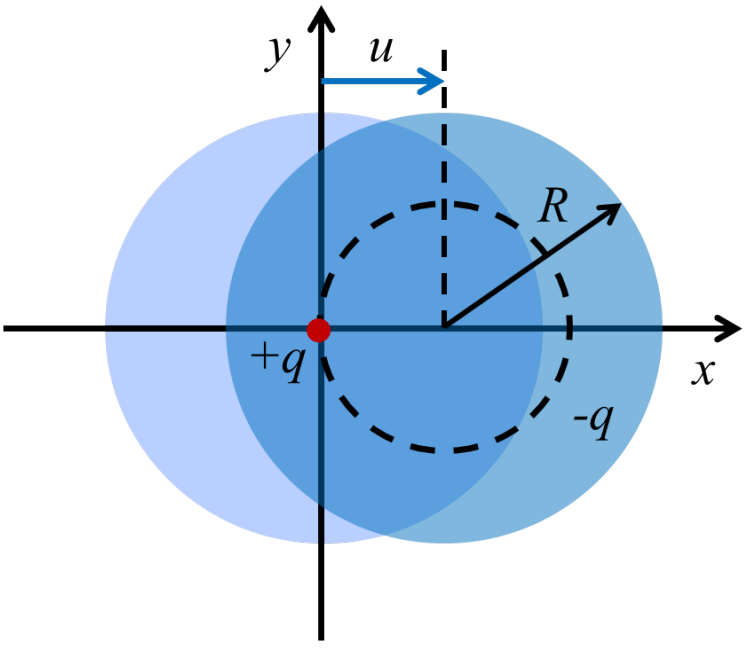


Figure 3.1: An electric dipole modelled as a sphere of radius R and constant negative charge density displaced from a positive point charge by u.

.

Let us imagine a somewhat fictionalized (but mathematically tractable) form for an electron state in which the negative charge $-q$ is distributed uniformly over a sphere of radius R. In equilibrium, this cloud of charge is centred on a positive point charge $+q$ at the origin. The sphere then becomes displaced from the positive charge by u (see Fig. 3.1). Now the point charge only sees a field due to the charge within the sphere centred on u. Thus, if the negative charge density is ρ, the charge contributing to the Coulombic force is

$$Q = -\frac{4\pi\rho u^3}{3}. \tag{3.7}$$

The force between the positive charge and electron cloud is therefore

$$F = \frac{qQ}{4\pi\epsilon_0 u^2} = -\frac{q\rho}{3\epsilon_0}u. \tag{3.8}$$

Hence, *this model of an electric dipole yields the same force law as the ideal spring* with $K = q\rho/\left(3\epsilon_0\right)$.

3.3.3 Energy in a simple harmonic oscillator

In the case of the mass on an ideal spring, the mass will have its maximum kinetic energy as it passes through its minimum of potential energy, at which point there is no force acting on it. Thereafter, the restoring force acts to reduce the kinetic energy, converting it to the potential energy stored in the spring. For a displacement u, the potential energy U is given by

$$U = -\int_0^u F\, du' = K\int_0^u u'\, du' = \tfrac{1}{2}Ku^2 + U_0, \tag{3.9}$$

where U_0 is the potential energy at $u = 0$. Since this is arbitrary, we are free to set $U_0 = 0$. Note that this parabolic form for the potential energy is just another way of specifying a harmonic oscillator.

At the maximum displacement, all the energy of the oscillator is stored as potential energy. Since the total energy ϵ of the system is a constant, we must have

$$\boxed{\epsilon = \tfrac{1}{2}KA^2,} \tag{3.10}$$

where A is the *amplitude* of the oscillation. Thus, *the energy of the oscillator is proportional to square of the amplitude*. An analogous result holds quite generally for linear oscillations, i.e. oscillations in which force is proportional to a displacement. Whilst the particular physics of the optical field are quite different, we shall find that the energy of waves in this field is also proportional to the square of the amplitude.

3.4 The wave equation

3.4.1 The linear chain of harmonic oscillators

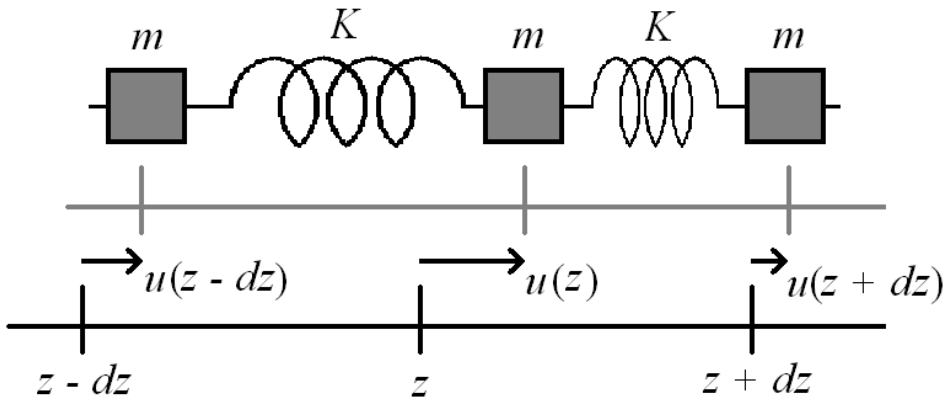


Figure 3.2: Diagram of a one-dimensional elastic medium modeled as a chain of harmonic oscillators. The displacements of the masses (dark blocks) are given in terms of a vector function $\mathbf{u}(z)$

.

So far, we have only considered the harmonic oscillator as a localised source. To see how such a disturbance may be propagated throughout space, we extend our treatment to consider many small masses coupled together by ideal springs. To simplify our analysis, we shall restrict ourselves to the one-dimensional case (and neglect external forces such as gravity). A diagram of such a system is shown in Fig. 3.2.

The mass of an infinitesimal section of the coupled system may be given as $m = \mu_z dz$, where μ_z is the mass per unit length. We may also write the spring constant K as the tension T_z per unit length, $K = T_z/dz$. Allowing the displacement at a point z to be given by the vector quantity $\mathbf{u}(z)$ and applying Newton's second law to the element at z , we then have

$$\mu_z dz \frac{\partial^2 \mathbf{u}(z)}{\partial t^2} = \frac{T_z}{dz}\left[\mathbf{u}(z+dz) - \mathbf{u}(z)\right] - \frac{T_z}{dz}\left[\mathbf{u}(z) - \mathbf{u}(z-dz)\right], \quad (3.11)$$

so

$$\frac{\partial^2 \mathbf{u}(z+dz)}{\partial t^2} = \left(\frac{T_z}{\mu_z}\right)\frac{\mathbf{u}(z+2dz) - 2\mathbf{u}(z+dz) + \mathbf{u}(z)}{dz^2}. \quad (3.12)$$

Taking the limit $dz \to 0$, we get

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} = \left(\frac{T_z}{\mu_z}\right) \frac{\partial^2 \mathbf{u}}{\partial z^2}. \tag{3.13}$$

This may equation may be generalised to three dimensions, yielding

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} = \left(\frac{T}{\mu}\right) \nabla^2 \mathbf{u}. \tag{3.14}$$

3.4.2 The phase velocity

Suppose that the function

$$\mathbf{u} = \mathbf{f}\left(z \pm vt\right) \tag{3.15}$$

is a solution of Eq. (3.14). If $\mathbf{f}$ maintains its shape as it propagates, then we see that v must be the velocity with which this wave profile propagates in the positive or negative z-direction. This is known as the *phase velocity*. Now let us put

$$\xi = z \pm vt, \tag{3.16}$$

so that, by the chain rule,

$$\frac{\partial}{\partial t} = \frac{\partial \xi}{\partial t}\frac{\partial}{\partial \xi} = \pm v \frac{\partial}{\partial \xi} \tag{3.17}$$

and

$$\frac{\partial^2}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial \xi}{\partial t}\frac{\partial}{\partial \xi}\right) = v^2 \frac{\partial^2}{\partial \xi^2}. \tag{3.18}$$

Similarly, we have

$$\frac{\partial}{\partial z} = \frac{\partial \xi}{\partial z}\frac{\partial}{\partial \xi} = \frac{\partial}{\partial \xi} \tag{3.19}$$

so

$$\frac{\partial^2}{\partial z^2} = \frac{\partial^2}{\partial \xi^2}, \tag{3.20}$$

whilst

$$\frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial y^2} = 0. \tag{3.21}$$

Hence, substituting Eq. (3.15) into Eq. (3.14), we have

$$v^2 \frac{\partial^2 \mathbf{f}}{\partial \xi^2} = \left(\frac{T}{\mu}\right) \frac{\partial^2 \mathbf{f}}{\partial \xi^2}, \tag{3.22}$$

which gives

$$v = \left(\frac{T}{\mu}\right)^{1/2}. \tag{3.23}$$

We may therefore write the wave equation in the more general form

$$\boxed{\nabla^2 \mathbf{u} = \frac{1}{v^2}\frac{\partial^2 \mathbf{u}}{\partial t^2}.} \tag{3.24}$$

This is the *general form* of the wave equation for a linear medium. Note that, although was derived for the oscillations of an elastic medium, the particular physics of the system have been washed out, leaving just the abstract mathematical form. This equation may then be used to to find the wave solutions for diverse physical systems, so long as the harmonic oscillator remains a valid model. In Chapter 6, we shall find that the same equation governs the electromagnetic oscillations of light in a linear medium.

3.4.3 The polarisation

The vectorial nature of the displacement implies that that $\mathbf{u}$ may be in the direction of the propagation (longitudinal) or transverse to it. This constitutes the *polarisation* of the medium. In general, there are various cases to consider, although we shall find later that not all applied to the optic field.

- *Unpolarised*

  means that the displacement is randomly orientated in space.

- *Longitudinal polarisation*

  In this case, the displacement is in the propagation direction . As an example, sound waves in fluids and solids are longitudinal. However, electromagnetic waves generally are not.

- *Transverse polarisation*

  In this case, the displacement is perpendicular to the propagation direction. For example, surface waves on a liquid are transverse. Electromagnetic waves (constituting light) are generally transverse. Moreover, the transverse polarisation may be

  - *linearly polarised*
    in which the displacement always lies in the same plane, or
  - *elliptically polarised*
    in which the polarisation rotates around the direction of propagation.

We shall defer detailed discussion of the polarisation of optical waves until Chapter 7. For the time being, we shall assume for simplicity that we have *linear polarisation* (although this is not the most general case). Taking the direction of propagation to be in the z axis, the polarisation may then be specified by a column vector

$$\mathbf{u}_0 = |\mathbf{u}_0| \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \tag{3.25}$$

where $|\mathbf{u}_0|\cos\theta$ and $|\mathbf{u}_0|\sin\theta$ are the x and y components of $\mathbf{E}_0$ respectively, and θ is the angle the polarisation vector makes with the x-axis.

3.4.4 Linearity of the wave equation

Linear differential equations

Equation (3.24) is an example of a *linear differential equation*. This categorization has a more general, abstract meaning than just that the forces depend only on linear displacements. Rather, it implies that the *principle of linear superposition* applies to the solutions. That is, if we can find two solutions $\mathbf{f}$ and $\mathbf{g}$ of a linear equation, then the function $\mathbf{h} = a\mathbf{f} + b\mathbf{g}$, where a and b are scalars, is also a solution.

It is straightforward to show that Eq. (3.24) observes this rule. Let us suppose that the solutions we have found are $\mathbf{f}(\mathbf{r},t)$ and $\mathbf{g}(\mathbf{r},t)$. Then, inserting $\mathbf{f}(\mathbf{r},t)$ into Eq. (3.24), we have

$$\nabla^2\mathbf{f} - \frac{1}{v^2}\frac{\partial^2\mathbf{f}}{\partial t^2} = 0 \tag{3.26}$$

and similarly for $\mathbf{g}(\mathbf{r},t)$. Let us now substitute

$$\mathbf{h}(\mathbf{r},t) = a\mathbf{f}(\mathbf{r},t) + b\mathbf{g}(\mathbf{r},t) \tag{3.27}$$

into Eq. (3.24). This gives

$$\nabla^2\mathbf{h} - \frac{1}{v^2}\frac{\partial^2\mathbf{h}}{\partial t^2} = a\left(\nabla^2\mathbf{f} - \frac{1}{v^2}\frac{\partial^2\mathbf{f}}{\partial t^2}\right) + b\left(\nabla^2\mathbf{g} - \frac{1}{v^2}\frac{\partial^2\mathbf{g}}{\partial t^2}\right) = 0, \tag{3.28}$$

which shows that $\mathbf{h}(\mathbf{r},t)$ is also a solution.

Linear operators

More generally, a linear differential equation does not involve products of the derivatives of the solutions. This includes the 'zero'-th derivative of a solution (i.e. the solution itself). For instance if $\mathbf{E}$ is a solution for a

dynamical system, then $\mathbf{E} \cdot \mathbf{E}$ is a (second order) non-linear term. Such terms would *not* appear in a linear differential equation.

In general, we may characterise a linear differential equation by saying that every term within it is a *linear operator*. Such an operator is characterised by the fact that it is *distributive over addition*. For instance, if $L(u)$ is a linear operator, then

$$L(f+g) = L(g) + L(f). \tag{3.29}$$

As an example, let

$$L(u) = \frac{du}{dx}. \tag{3.30}$$

Then

$$L(f+g) = \frac{d}{dx}(f+g) = \frac{df}{dx} + \frac{dg}{dx}. \tag{3.31}$$

On the other hand, suppose we had

$$L(u) = u^2. \tag{3.32}$$

Now

$$L(f+g) = (f+g)^2 = f^2 + 2fg + g^2 \neq L(f) + L(g). \tag{3.33}$$

Hence, in this case, $L(u)$ is *nonlinear*. As another example, consider

$$L(u) = u\frac{du}{dx}. \tag{3.34}$$

In this case, we have

$$L(f+g) = (f+g)\left(\frac{df}{dx} + \frac{dg}{dx}\right) \neq f\frac{df}{dx} + g\frac{dg}{dx}, \tag{3.35}$$

so again, $L(u)$ is nonlinear.

This discussion is of relevance to optics since the response of a material medium to a very intense optical field may include nonlinear terms. This is the subject matter of *nonlinear optics*, which is beyond the scope of this course. It is therefore an important classification of a given medium that it is *linear* to good approximation, by which we mean that the dynamics of light propagation within it may be modelled by linear operators.

3.5 The refractive index

3.5.1 Properties of a medium

In the derivation of the wave equation Eq. (3.24), we assumed that the forces on an element of the displaced field or medium depended only on the linear displacement. This yielded a constant phase velocity v. Although the details of different systems may be very different, this result will remain generally true where-ever this modelling assumption is valid.

In reality, of course, additional physical effects may emerge that will change this fundamental result. Invariably, this will lead to a phase velocity that depends in some way on these extra factors. Since the details are likely to be particular to the physical system, we make no attempt to quantify them here. Later, in Chapter 6, we shall see how the simple model of wave propagation becomes modified in optics.

For now, let us denote the constant phase velocity found in Eq. (3.24) by c. In optics, this will be correspond to *the speed of light in a vacuum*. As we shall see, this is a universal constant, emerging from the two other constants of nature the *permittivity* and *permeability* of free space. In analogy with the spring constant, we might think of these constants as describing the fundamental 'stiffness' of space.

In a material medium, other factors come into play. Rather than plunge into the details of these forces, for the time being we merely subsume their effect into a material parameter known as the *refractive index* n. The modified wave speed v is then

$$v = \frac{c}{n}. \tag{3.36}$$

This allows us to re-write the wave equation as

$$\boxed{\nabla^2 \mathbf{u} = \frac{n^2}{c^2} \frac{\partial^2 \mathbf{u}}{\partial t^2}.} \tag{3.37}$$

The refractive index then characterizes the properties of the material medium that determine optical propagation through it. These properties may grouped into the following important categories:

- *Linear medium*

  This category has already been discussed earlier. To reiterate, in a linear medium, wave propagation in the medium may be accurately modelled by a *linear differential equation*, which means that we may construct linear superpositions of the solutions. In practice, the linearity of a medium usually means that only terms linear in the displacement are considered.

- *Isotropic medium*

  In an isotropic medium, the material properties are the same in every direction. This means that the refractive index will not depend on direction of propagation or the orientation of the polarisation. Conversely, materials that do have a directional dependence are known as *anisotropic*.

- *Homogeneous medium*

  In a homogeneous medium, the material properties are the same everywhere. That is, they do not depend on spatial position. Conversely, a medium in which the properties *do* depend on spatial position is called *inhomogeneous*.

3.6 Plane waves

3.6.1 The wavefront

Locally, each part of the propagating wave solution to Eq. (3.24) will behave like a simple harmonic oscillator, each vibrating with the frequency of the source. Points in space having the same phase within their cycle of oscillation constitute a *wavefront*.

As the phase of each point in space changes, the wavefront must then propagate outwards from the source. The speed of the wave front is then the *phase velocity* v featured in the wave equation. The distance between successive wavefronts in the direction of propagation is, of course, the *wavelength* λ, related to the phase velocity by

$$v = \frac{\lambda}{T} = f\lambda = \frac{\omega}{2\pi}\lambda. \tag{3.38}$$

A slightly more useful description is given via the definition of the *wavevector* $\mathbf{k}$. The magnitude of the wavevector is related to the wavelength via

$$\boxed{|\mathbf{k}| = k = \frac{2\pi}{\lambda},} \tag{3.39}$$

so that Eq (3.38) may be rendered in the form

$$\boxed{v = \frac{\omega}{k}.} \tag{3.40}$$

However, since $\mathbf{k}$ is a vector, it also gives information about the direction of the propagation. If the phase velocity does not depend on direction in space (i.e. the medium is *isotropic*), then $\mathbf{k}$ will be perpendicular to the wavefront.

3.6.2 Plane wavefronts

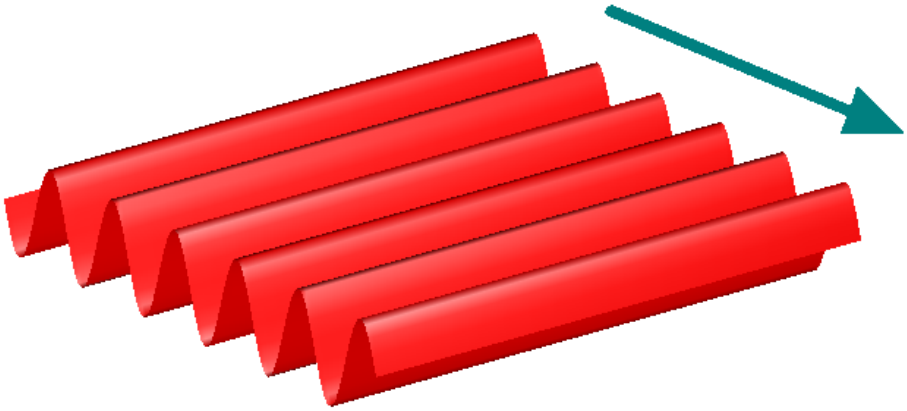


Figure 3.3: Illustration of a section of plane wave (shown in two dimensions). The arrow shows the direction of propagation. Note that the example shown here is for a *transversely polarised* wave - i.e. the displacement of the wave is at right angles (or 'transverse') to the direction of propagation. The wavefronts are defined by the regions of constant phase, which here are the planes perpendicular to the direction of motion (only one dimension of these planes is illustrated here but we see that the wavefronts are along straight lines perpendicular to the propagation direction).

A very useful mathematical description of a wave is given by the *plane wave* (see Fig. 3.3). Such planar wavefronts parallel to the direction of motion. Hence, if, for example, the wave is traveling in the z direction, all points in the $x-y$ plane at a given value of z have the same phase (an therefore define a wavefront). Using a general complex form, we may express a plane wave as

$$\mathbf{E}\left(\mathbf{r},t\right)=\mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \tag{3.41}$$

where, $\mathbf{E}_0$ is a vector containing information about the polarisation.

A particular utility of plane wave solutions is that we may immediately make the substitutions

$$\nabla = i\mathbf{k}, \tag{3.42}$$

$$\nabla^2 = -k^2, \tag{3.43}$$

$$\frac{\partial}{\partial t} = -i\omega, \tag{3.44}$$

and

$$\frac{\partial^2}{\partial t^2} = -\omega^2. \tag{3.45}$$

Note, however that these substitutions are only valid for *single* plane wave solutions. Assuming this is the case, the wave equation may now be written

$$k^2\mathbf{u} = \omega^2\frac{n^2}{c^2}\mathbf{u}, \tag{3.46}$$

from which we confirm

$$\frac{\omega}{k} = \frac{c}{n}. \tag{3.47}$$

3.6.3 Fourier's Theorem

The plane wave is a somewhat fictionalized entity since it takes as its source an infinite two-dimensional plane. In practice, such waves may be taken to approximate spherical waves radiating from a point source as the distance to the source approaches infinity. However, a particular utility of plane waves lies in their mathematical versatility by virtue of *Fourier's Theorem*.

Fourier's Theorem tells us that any periodic function (subject to certain mathematical constraints) may be represented by a sum of sinusoidal terms with the same periodicity. Moreover, as we take the limit of the period $T \to \infty$, we find that we may represent a given function by an integral of sinusoids. However, this is precisely what the plane waves are. This means that, in principle, we can represent any solution of the wave function by an integral over plane waves.

Hence, labeling the plane waves by wavevector, the general solution of the wave equation may be written

$$\mathbf{E}\left(\mathbf{r}, t\right) = \int \mathbf{E}_{\mathbf{k}} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega_{\mathbf{k}} t)} d^3\mathbf{k}. \tag{3.48}$$

3.7 Group velocity

Very often, the refractive index will depend on the frequency of the oscillation. This means that, in turn, the phase velocity will also depend on frequency. Hence, the components of the general solution to the wave equation, Eq. (3.48), will tend to drift apart - a phenomenon known as *dispersion*.

With no unique phase velocity, the question of the how fast a wave signal actually propagates naturally arises. Consider the general solution wave expression given by Eq. (3.48). We may think of this as describing a

wave packet, which must travel with some average velocity. We call this the *group velocity* $\mathbf{v}_g(\mathbf{k}_0)$ associated with some average wavevector $\mathbf{k}_0$.

3.7.1 Derivation of the group velocity

If the domain of $\mathbf{k}$ does not exceed by too much some average wavevector $\mathbf{k}_0$, then to a good approximation we can write $\omega_{\mathbf{k}}$ as a Taylor series expanded to first order

$$\omega_{\mathbf{k}} \approx \omega_{\mathbf{k}_0} + \nabla_{\mathbf{k}}\omega_{\mathbf{k}_0} \cdot (\mathbf{k} - \mathbf{k}_0), \tag{3.49}$$

where $\nabla_{\mathbf{k}}$ is the gradient operator with respect to wavevector

$$\nabla_{\mathbf{k}} = \hat{\mathbf{x}}\frac{\partial}{\partial k_x} + \hat{\mathbf{y}}\frac{\partial}{\partial k_y} + \hat{\mathbf{z}}\frac{\partial}{\partial k_z} \tag{3.50}$$

and $\nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}$ is evaluated at $\mathbf{k} = \mathbf{k}_0$. Substituting Eq. (3.49) into Eq. (3.48) we obtain

$$\begin{aligned}\mathbf{E}(\mathbf{r},t) &= \exp\left[i\left(\nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}\cdot\mathbf{k}_0 - \omega_{\mathbf{k}_0}\right)t\right]\int \mathbf{E}_{\mathbf{k}}\exp\left[i\mathbf{k}\cdot\left(\mathbf{r} - \nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}t\right)\right]\, d^3\mathbf{k} \\ &= e^{i\theta(t)}\Psi(\mathbf{z}(\mathbf{r},t),0),\end{aligned} \tag{3.51}$$

where

$$\theta(t) = \left(\nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}\cdot\mathbf{k}_0 - \omega_{\mathbf{k}_0}\right)t \tag{3.52}$$

and

$$\mathbf{z}(\mathbf{r},t) = \mathbf{r} - \nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}t. \tag{3.53}$$

Hence $\mathbf{E}(\mathbf{r},t)$ is modulated by a time-dependent phase $e^{i\theta(t)}$ whilst being translated with a velocity $\nabla_{\mathbf{k}}\omega_{\mathbf{k}_0}$. This is the *group velocity*. Dropping the '0' subscript on $\mathbf{k}$, this is

$$\boxed{\mathbf{v}_g(\mathbf{k}) = \nabla_{\mathbf{k}}\omega_{\mathbf{k}}.} \tag{3.54}$$

3.7.2 Significance of the group velocity

What, then, is the physical significance of the group velocity? The question is of particular importance in optics when, under certain conditions, the *phase* velocity may exceed the speed of light c. However, Einstein's Special Theory of Relativity tells us that nothing can travel faster than the speed of light. Does this not imply a contradiction?

In fact, the answer is *no*. To be precise, the Special Relativity tells us that *no physical signal* may travel faster than c. This means that anything 'traveling' faster than c cannot be carrying any meaningful information, which might be re-interpreted as saying that nothing is really traveling at this speed. Only physically observable quantities such as energy, momentum or charge can transmit information. The group velocity of a wave-packet is then the speed at which such observables may be transmitted. The phase velocities of the component parts of the wave-packet should then be viewed more in terms of a mathematical framework describing the physical phenomenon, rather than an observable part of the physics.

3.8 Summary

- **The simple harmonic oscillator**
  - *The ideal spring*
    The equation of motion for an ideal spring of stiffness constant K suspending a mass m is

$$m\frac{d^2u}{dt^2} = -Ku. \tag{3.55}$$

    This has the general solution

$$u(t) = A\sin(\omega t + \phi), \tag{3.56}$$

    where

$$\omega = \left(\frac{K}{m}\right)^{1/2} \tag{3.57}$$

    is the *angular frequency* and ϕ is a phase factor determining the position at time $t = 0$.
  - *Energy in a simple harmonic oscillator*
    The total energy of an simple harmonic oscillation with amplitude A is

$$\boxed{\epsilon = \tfrac{1}{2}KA^2.} \tag{3.58}$$

- **The wave equation**

  The general form of the wave equation is

$$\boxed{\nabla^2 \mathbf{u} = \frac{n^2}{c^2} \frac{\partial^2 \mathbf{u}}{\partial t^2},} \tag{3.59}$$

where n is the *refractive index* and c is the *phase velocity* of a simple harmonic wave.

- **Polarisation**
  - *Unpolarised*
    which means that the displacement is randomly orientated in space.
  - *Longitudinal polarisation*
    In this case, the displacement is in the propagation direction. As an example, sound waves in fluids are longitudinal. However, electromagnetic waves are generally not.
  - *Transverse polarisation*
    In this case, the displacement is perpendicular to the propagation direction. This is generally the case for electromagnetic waves. Moreover, the transverse polarisation may be
    - *linearly polarised*
      in which the displacement always lies in the same plane, or
    - *elliptically polarised*
      in which the polarisation rotates around the direction of propagation.

- **Properties of a medium**
  - *Linear medium*
    In a linear medium, wave propagation in the medium may be accurately modeled by a *linear differential equation*.
  - *Isotropic medium*
    In an isotropic medium, the material properties are the same in every direction. This means that the refractive index will not depend on direction of propagation or the orientation of the polarisation. Conversely, materials that do have a directional dependence are known as *anisotropic*.
  - *Homogeneous medium*
    In a homogeneous medium, the material properties the same everywhere. That is, they do not depend on spatial position. Conversely, a medium in which the properties *do* depend on spatial position is called *inhomogeneous*.

- **The wavefront**

  Points in space having the same phase within their cycle of oscillation constitute a *wavefront*.

  - The speed of the wave front is then the *phase velocity* v
  - The distance between successive wavefronts in the direction of propagation is the *wavelength* λ
  - The magnitude of the wavevector $|\mathbf{k}|$ is related to the wavelength via

$$\boxed{|\mathbf{k}| = k = \frac{2\pi}{\lambda}.} \tag{3.60}$$

    In an *isotropic* medium $\mathbf{k}$ is perpendicular to the wavefront.
  - The phase velocity is related to the wavevector via

$$\boxed{v = \frac{\omega}{k}.} \tag{3.61}$$

- **Plane waves**

  A plane wave solution of the wave equation may be written as

$$\mathbf{E}\left(\mathbf{r}, t\right) = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}, \tag{3.62}$$

  allowing us to write a general solution in the form

$$\mathbf{E}\left(\mathbf{r}, t\right) = \int \mathbf{E}_{\mathbf{k}} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega_{\mathbf{k}} t)}\, d^3\mathbf{k}. \tag{3.63}$$

- **Group velocity**

  The group velocity is given by

$$\boxed{\mathbf{v}_g\left(\mathbf{k}\right) = \nabla_{\mathbf{k}} \omega_{\mathbf{k}}.} \tag{3.64}$$

  This is the speed at which a physical signal may be transmitted.

Part II

Wave Optics

4. The Huygens-Fresnel Principle

4.1 General remarks

Following the philosophy of Chapter 3, we continue with our exploration of general wave phenomena without necessitating any recourse to the physical specifics of the waves. In this pursuit, the general features of wave propagation may be derived from a handful of simple concepts.

In the interests of simplicity, we confine ourselves to media that are linear, isotropic and (locally) homogeneous. Here, we briefly remind ourselves of the meanings of these terms as introduced in the Chapter 3:

- *Linearity*

  In a linear medium, wave propagation in the medium may be accurately modeled by a *linear differential equation*. In particular, the propagation of the disturbance depends only on the *amplitude* of the oscillations. Higher order terms may be considered negligible.

- *Isotropy*

  In an isotropic medium, the material properties are the same in every direction. This means that the refractive index will not depend on direction of propagation or the orientation of the polarisation.

- *Homogeneity*

  In a homogeneous medium, the material properties, and hence the refractive index, are the same everywhere. That is, they do not depend on spatial position.

The last point concerning homogeneity will need a little modification, since we shall be concerned with propagation between media of different refractive index, which could be interpreted as an inhomogeneous system. However, such cases will be restricted to a sharp discontinuity in the refractive indices. Within specific regions, the refractive index (for a given frequency) stays constant. Hence, we shall describe this situation as *locally* homogeneous.

The first principle we shall apply is due to the Dutch physicist Christiaan Huygens (1629 - 1695). *Huygens' Principle* [1] explains wave propagation in terms of spherical wavelets radiating from all points on a wavefront. At some later time, a new wavefront may be constructed from the sum of these

wavelets. We therefore begin this chapter with a consideration of *spherical* waves, so that we may better understand the concepts to follow.

Huygens' Principle may be thought of as more of a rule of thumb than a law of physics. Its principal merit is that it generally works, although there are a few situations that it fails to describe adequately, such as the backwards propagation of the spherical wavelets. The reason for this lack in explanatory power comes from the neglect of *interference effects*, which truly characterise wave phenomena (although, in some texts, interference effects *are* included in the definition of Huygens' Principle.

Our understanding of interference effects comes largely from the work of the French physicist Augustin-Jean Fresnel (1788 - 14 July). When this theory is combined with Huygens' Principle, we have a truly powerful explanatory tool to hand in the form of the *Huygens-Fresnel Principle* [2]. From this principle, the general principles of wave propagation may be derived. As in Chapter 3, we again emphasise that these principles are not restricted to optical propagation but have universal application throughout physics where-ever wave phenomena is encountered.

4.2 Learning objectives

The aims of this section are to gain understanding of

- The formulation of spherical waves
- Huygens' Principle
- Interference and coherence
- The Huygens-Fresnel Principle
- Application of these principles to find:
  - The Law of Rectilinear Propagation
  - The Law of Reflection
  - The Law of Refraction (Snell's Law)

4.3 Spherical waves

Previously we looked at the propagation of waves in terms of plane waves. True plane waves, however, are an idealism and are only ever represented approximately in reality. A more realistic example of a wave propagating in an isotropic and homogeneous medium is that of a *spherical wave*. In such a medium, the phase velocity is the same in all directions and the wavefronts emanating from a point source (i.e. surfaces of constant phase) will be spherical. The phase term of a spherical wave may then be given by $(\omega t - kr)$, where r is the radial distance from the source of the wave and k is the magnitude of the wave-vector.

The amplitude of the spherical wave must be modified, however, to ensure energy conservation. Since the energy of the disturbance originates from the point source, the total energy crossing any given spherical wavefront centred on that source must be a constant. At the same time, the *intensity* (energy flow across unit area) must be proportional to the squared modulus $|\mathbf{E}|^2$ of the wave. Thus, we must have

$$4\pi r^2 |\mathbf{E}|^2 = C, \tag{4.1}$$

where C is a constant. Hence, we have

$$|\mathbf{E}|^2 = \frac{C}{4\pi r^2}. \tag{4.2}$$

This then implies

$$\mathbf{E} = \frac{\mathbf{E}_0}{r} e^{i\phi}. \tag{4.3}$$

where ϕ is a phase factor. Substituting $(\omega t - kr)$ for ϕ then gives

$$\mathbf{E}(r,t) = \frac{\mathbf{E}_0}{r} e^{i(\omega t - kr)}. \tag{4.4}$$

We require one further condition on the form of a spherical wave, namely that it must remain finite at $r = 0$. This condition is met by setting the $\cos\phi$ component of the complex exponential to zero and imposing the form

$$\mathbf{E}(r,t) = \frac{\mathbf{E}_0}{r} \sin(\omega t - kr). \tag{4.5}$$

A plot of this function at a given time t is shown in Fig. 4.1.

Alternatively, we could allow the temporal factor to remain complex and put

$$\mathbf{E}(r,t) = \frac{\mathbf{E}_0}{r} e^{i\omega t} \sin(kr). \tag{4.6}$$

The particular choice of the form may be made based on mathematical convenience.

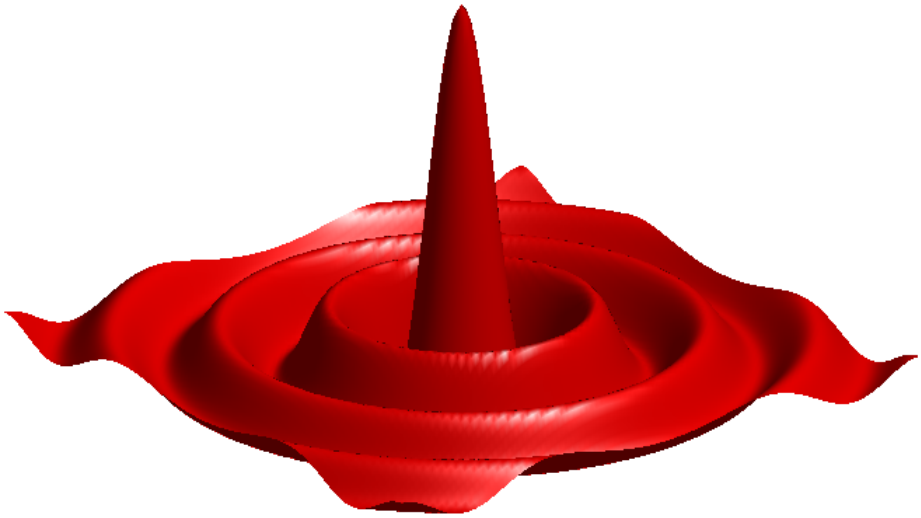


Figure 4.1: Illustration of a spherical wave (shown propagating in two-dimensions). This wave propagates outwards in all directions from the central peak. The form shown here is given by Eq. (4.5).

4.4 The geometrical wavefront

In Chapter 3 we defined a *wavefront* as the locus of points in space having the same phase. Here, we are in a position to find the propagation of a wavefront due to a disturbance in an oscillating field via application of *Huygen's Principle*. This may be stated as

> **Huygens' Principle**
>
> *Each point on a wavefront acts as a source of secondary, spherical wavelets. At a later time, t, a new wavefront is constructed from the sum of these wavelets*.

Since we shall be assuming (for the time being) that the medium through which the disturbance is propagating is *linear*, *isotropic* and (locally) *homogeneous*, we can construct a wavefront for a particular frequency geometrically. This is because the time of propagation and distance traveled is related via the phase velocity, which for a given frequency is a constant. Specifically, for a spherical wave propagating from a point source, after a time t, the radius of the spherical wavefront is

$$r = vt. \tag{4.7}$$

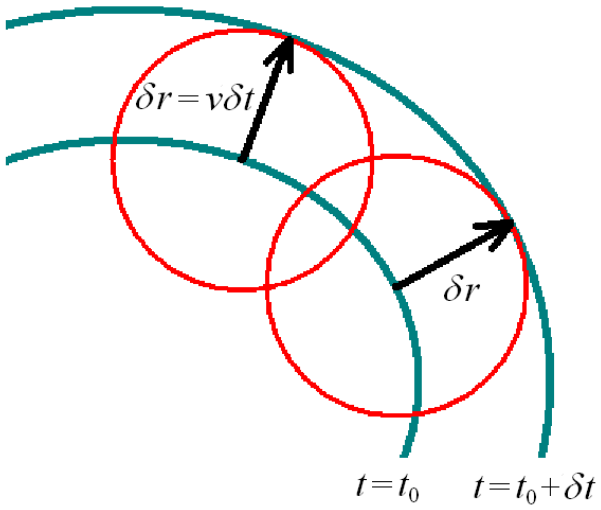


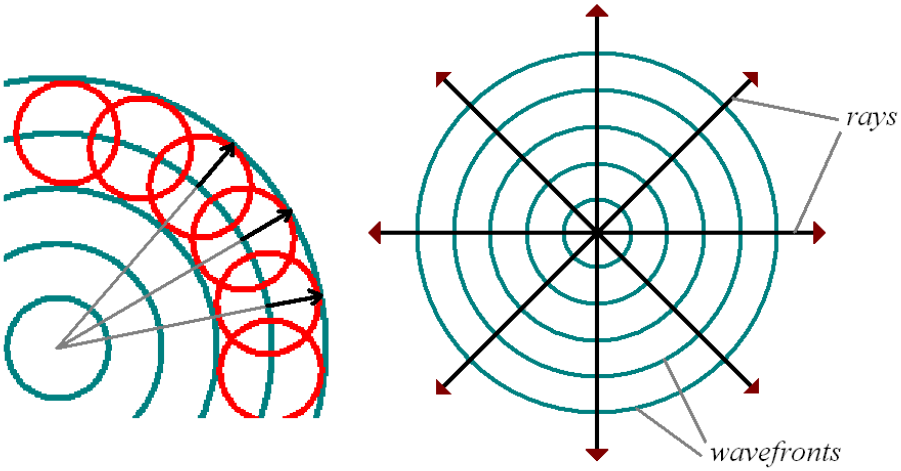
Figure 4.2: Illustration of Huygens' Principle in the general case. The initial wavefront at $t = t_0$ acts as a source of secondary, spherical wavelets. After an infinitesimal time δt, the wavelets have acquired radii $\delta r = v\delta t$, which, in the isotropic and homogeneous case, are all equal. The new wavefront at $t = t_0 + \delta t$ is then tangential to each wavelet (note that only two such wavelets are shown for the sake of clarity).

Figure 4.2 illustrates Huygens' Principle in the general case. The initial wavefront at $t = t_0$ acts as a source of secondary, spherical wavelets (of the form given by Eq. (4.5)). After an infinitesimal time δt, the wavelets have acquired radii $\delta r = v\delta t$. Since we are restricting our consideration to isotropic, homogeneous media, these radii are all equal. The new wavefront at $t = t_0 + \delta t$ is then tangential to each wavelet (note that Fig. 4.2 only shows two such wavelets for the sake of clarity).

We may note that the two wavefronts have a shortest distance between them δr. It follows then that the vector from a point on the initial wavefront along this shortest distance is at right angles to the wavefront. Picturing this wavefront as a surface of constant phase embedded in three-dimensional space, it should be appreciated that this vector is then parallel with the gradient of the phase - i.e. points in the direction of the fastest change in phase.

A line crossing a wavefront perpendicularly to it is known as a *ray*. In the isotropic case, a ray also indicates the direction of wave propagation. The relationship between wavefronts and rays is further elucidated in the simplified cases of spherical and planar waves. These are discussed in the next sub-sections.

4.4.1 Spherical wavefronts



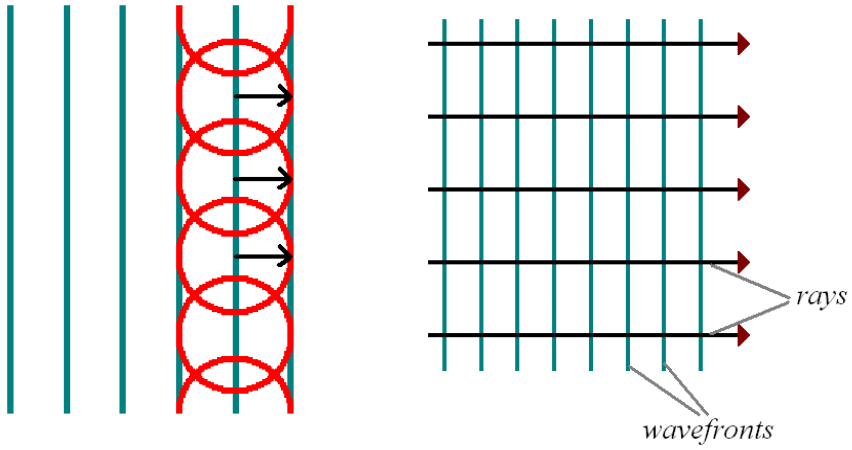
(a) (b)

Figure 4.3: Illustration of Huygen's Principle for spherical wavefronts. In (a), an initial spherical wavefront acts as a source of secondary wavelets. Since the new wavefront is everywhere displaced by the same distance from the original, it must also be spherical and centred on the same source. Thus, in (b), we see a representation of spherical wavefronts in two-dimensions as a sequence of concentric rings. The rays radiating out from the central point are everywhere perpendicular to the wavefronts.

Figure 4.3 shows an illustration of Huygen's Principle for spherical wavefronts. As described before for the general case, an initial spherical wavefront acts as a source of secondary wavelets. Since the new wavefront is everywhere displaced by the same distance from the original, it must also be spherical and centred on the same source. We may therefore visualise a spherical wave as a sequence of concentric spherical surfaces of constant phase. In Fig. 4.3 (b), this is portrayed in two dimensions in terms of concentric rings. Note that the rays radiating out from the central point are everywhere perpendicular to the wavefronts.

4.4.2 Planar wavefronts

Figure 4.4 illustrates Huygens' Principle for plane waves. In Fig. 4.4 (a), we see the initial wavefront acting as a source of secondary, spherical wavelets. Since the wavelets all have the same radii and the new wavefront must be tangential to these, we see that the wavefronts must all be parallel to each other. Moreover, these wavefronts move in the direction of



(a) (b)

Figure 4.4: Illustration of Huygens' Principle for plane waves. (a) At an initial time, a wavefront acts as a source of secondary, spherical wavelets. Since the wavelets all have the same radii and the new wavefront must be tangential to these, we see that the wavefronts must all be parallel to each other. Moreover, these wavefronts move in the direction of the rays, which are perpendicular to them. Thus, as may be seen in (b), a plane wave must propagate in a straight line at right angles to the planes of constant phase.

the rays, which are perpendicular to them. Thus, as illustrated in Fig. 4.4 (b), a plane wave must propagate in a straight line at right angles to the planes of constant phase.

Note that this result could have been obtained from the limiting case of the spherical wavefront as the radius of the leading wavefront $r \to \infty$. In this case, the wavefronts then tend to infinitely extended planes. In both cases wave propagation is in the direction of the rays, which are perpendicular to wavefronts.

This then yields the law of rectilinear propagation for waves in an isotropic and homogeneous medium - i.e. the *waves travel in straight lines*.

4.5 The laws of wave propagation

4.5.1 Rectilinear propagation

Here, we reiterate the conclusion of the last section for an infinite plane-wave propagating in a isotropic and homogeneous medium. The direction of propagation of the wavefront will be orthogonal to the plane. Since ev-

ery point on the plane is the source of a secondary wavelength, at a later time the new wavefront must be at a constant orthogonal distance from the original wavefront (i.e. parallel to it). Thus the wavefronts must continue to propagate in the same orthogonal direction. In other words:

In a isotropic and homogeneous medium, light travels in straight lines.

4.5.2 Reflection

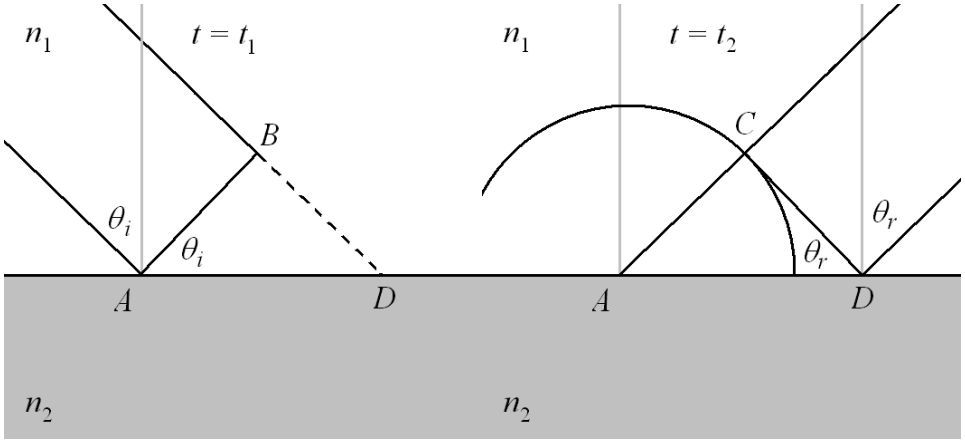


Figure 4.5: (a) A plane wave approaching the boundary between two homogeneous media with refractive indices n_1 and n_2 at time $t = t_1$, just as the edge of the wavefront touches the boundary at A. (b) At a later time $t = t_2$, the wavefront emanating from point A has a radius equal to the length of BD. The reflected wavefront now lies along the line CD.

Figure 4.5 (a) shows the wavefront of a plane-wave approaching the boundary between two homogeneous media with refractive indices n_1 and n_2. The direction of propagation is at an angle of θ_i to the normal of the boundary. This is the *angle of incidence*. At the time $t = t_1$ shown, the wavefront has just reached the boundary at point A.

At a later time $t = t_2$, the point on the wavefront at B reaches the boundary at D, shown in Fig. 4.5 (b). The length of BD is then

$$BD = \frac{cT}{n_1}, \tag{4.8}$$

where $T = t_2 - t_1$. During the same time, the wavefront emanating from A has traversed the same radial distance since this is also propagating in the same medium. Hence

$$AC = BD. \tag{4.9}$$

The reflected wavefront must therefore pass through D and be tangential to the spherical wavefront centred on A. The direction of propagation then makes an angle θ_r to the normal. This is the *angle of reflection*.

From elementary trigonometry, in Fig. 4.5 (a) we have

$$BD = AD \sin \theta_i \tag{4.10}$$

and, from Fig. 4.5 (b)

$$AC = AD \sin \theta_r. \tag{4.11}$$

Equation. (4.9) tells us that we may equate equations (4.10) and (4.11), giving

$$\sin \theta_i = \sin \theta_r \tag{4.12}$$

or, simply,

$$\boxed{\theta_i = \theta_r.} \tag{4.13}$$

In other words:

The angle of incidence is equal to the angle or reflection.

4.5.3 Refraction

Figure 4.6 (b) shows the transmitted wave propagating in the medium with refractive index n_2. During the time T, the spherical wavefront at A has propagated a radial distance of

$$AE = \frac{cT}{n_2} = AD \sin \theta_t. \tag{4.14}$$

Substituting for cT from Eq. (4.8), we have

$$BD \frac{n_1}{n_2} = AD \sin \theta_t. \tag{4.15}$$

Using Eq. (4.10) for BD and rearranging, we arrive at

$$\boxed{n_1 \sin \theta_i = n_2 \sin \theta_t.} \tag{4.16}$$

This is *Snell's Law* for refraction.

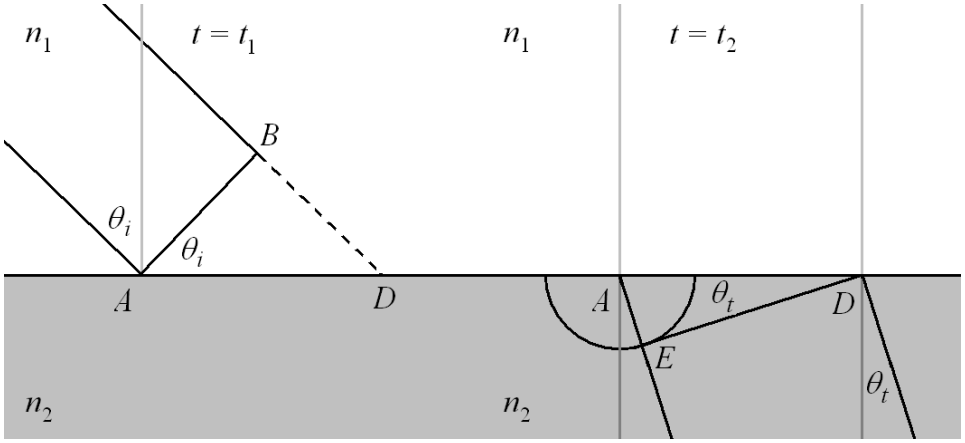


Figure 4.6: (a) A plane wave approaching the boundary between two homogeneous media with refractive indices n_1 and n_2 at time $t = t_1$. (b) At a later time $t = t_2$, the wavefront emanating from point A within the second medium n_2 has propagated a radial distance of AE. The new plane-wave wavefront then passes through D and E.

4.6 Total internal reflection

4.6.1 The critical angle

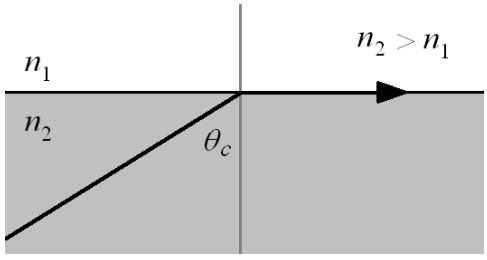


Figure 4.7: A ray of light approaching an interface between two media at the critical angle of incidence θ_c.

Snell's law gives the relation of the angles to the normal θ_1 and θ_2 of a ray of light crossing a boundary between media of refractive indices n_1 and n_2

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{4.17}$$

Now, if $n_2 > n_1$, then for light approaching the boundary on the n_2 side, there will be some range of incident angles such that the condition for

transmission is not met (since such cases would require $\sin\theta_1 > 1$).

The minimum angle θ_c at which light approaching from the medium with the higher refractive index (n_2) propagates exactly along the interface without being transmitted into the medium of refractive index n_1 is known as the *critical angle* . For this, we require $\theta_1 = \pi/2$, so $\sin\theta_1 = 1$ and

$$\theta_c = \sin^{-1}\left(\frac{n_1}{n_2}\right). \tag{4.18}$$

If the angle of incidence (from the n_2 medium) is greater than θ_c, there appears to be no transmission and the ray is entirely reflected at the interface. This is known as *total internal reflection*. In fact, we shall see in Chapter 8 that there is, in fact, a decaying wave transmitted into the n_1 known as the *evanescent wave*.

4.6.2 The slab waveguide

As an application of the phenomenon of total internal reflection, we shall consider the confinement of light within a *slab waveguide*, i.e. a slab of transparent material sandwiched between layers of a material with a lower refractive index. The proper analysis of such a waveguide requires solution of Maxwell's equations but we may still garner an intuitive picture of things using the present tools at our disposal.

Figure. 4.7 shows a schematic of a slab waveguide, with light approaching the interface between media from the medium of higher refractive index. The *numerical aperture* (NA) of the guide is defined as

$$\mathrm{NA} = n_0 \sin\theta_0, \tag{4.19}$$

where θ_0 is the maximum angle of incidence for light entering the system from a medium of refractive index n_0 that the guide will accept (light approaching at angles greater than this will be lost in the cladding layers).
Using Snell's law, we see from Fig. 4.8 that the numerical aperture is given by

$$n_0 \sin\theta_0 = n_2 \sin\alpha = n_2 \sin\left(\pi/2 - \theta_c\right). \tag{4.20}$$

Using standard trig identities, we have

$$n_2 \sin\left(\pi/2 - \theta_c\right) = n_2 \cos\theta_c = n_2 \left(1 - \sin^2\theta_c\right)^{1/2}. \tag{4.21}$$

Inserting the expression for θ_c found in (a), this becomes

$$n_2 \sin\left(\pi/2 - \theta_c\right) = n_2 \left(1 - \left(\frac{n_1}{n_2}\right)^2\right)^{1/2}. \tag{4.22}$$

Hence, taking n_2 inside the outer brackets, the required result for the numerical aperture is

$$n_0 \sin\theta_0 = \left(n_2^2 - n_1^2\right)^{1/2}. \tag{4.23}$$

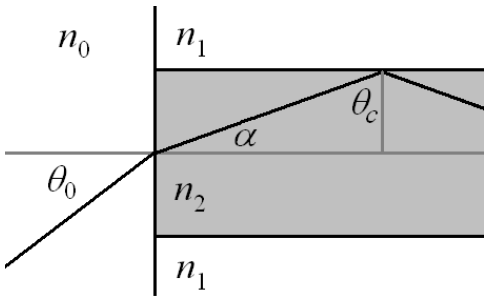


Figure 4.8: Construction for the numerical aperture (NA) of a slab waveguide consisting of cladding layers with refractive index n_1 and a guide layer of refractive index n_2. Light enters the guide from a medium of refractive index n_0.

4.7 Interference and coherence

4.7.1 Interference effects

The *Principle of Linear Superposition* asserts that, for a linear medium, two separate waveforms may be added together linearly to produce a single displacement of the field. In terms of electromagnetic waves, this means that the electric and magnetic fields associated with the propagating optical field add linearly via normal vector addition. (It should be pointed out that this no longer applies in the field of non-linear optics, where the induced electrical polarisation (or magnetisation) of a material medium is no longer a linear function of the applied fields).

This linear superposition can lead to enhancement or cancellation of the field - known respectively as *constructive* and *destructive interference*. More generally, the phenomena constitutes what we call *interference effects*.

For interference effects to appear over observable spatial or temporal intervals, the underlying waveforms must be *coherent*. That is, there must be some fixed relative phase between them. In practice, of course, such coherence does not persist indefinitely and over all space. We therefore speak of the *degree of coherence*, which may be quantified in terms of interference effects. Moreover, if we are considering the coherence of waves displaced in space, we talk of *spatial coherence* (i.e. the waveforms at different points in space stay in phase). At a particular point in space, we are

concerned with *temporal coherence* (the waveforms at different points in time stay in phase).

4.8 Huygens-Fresnel Principle

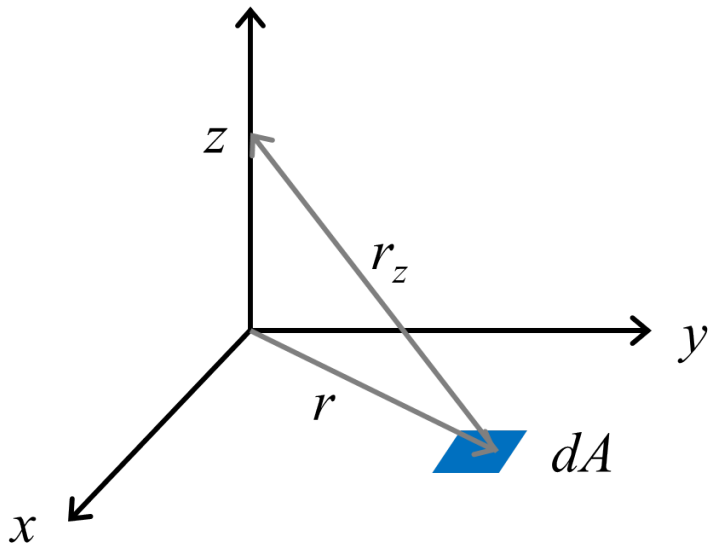


Figure 4.9: Geometric construction for the superposition of waves over the $x - y$ plane at point z, via the Huygens-Fresnel Principle.

With these points in place, we are now in a position to state the *Huygens'-Fresnel Principle*. This is a subtle modification of Huygens' Principle to explicitly include the effects of interference via the principle of superposition and may be stated as follows.

> **The Huygens'-Fresnel Principle**
>
> *For light of a given frequency, every point on a wavefront acts as a secondary source of spherical wavelets with the same frequency and the same initial phase. The wavefront at a later time and position is then the linear superposition of all of these wavelets.*

4.8.1 Propagation of a plane wave

We saw previously that plane-wave propagation, a hence the Law of Rectilinear Propagation, could be analysed in terms of Huygen's Principle. However, that principle does not forbid propagation of an arbitrary wavefront in both the forwards *and backwards* direction. To see how this problematic,

imagine two wavefronts propagating in either direction away from a central wavefront. An arbitrary time later, waves from these daughter wavefronts could propagate back to the original position, arriving there in phase. Thus the central wavefront *would never move*. This problem is resolved via the Huygens'-Fresnel Principle through the introduction of *interference*. First, let us see how this principle predicts the Law of Rectilinear Motion.

Consider the set up illustrated in Fig. 4.9, where it is assumed that the wavefront is spread out over the $x-y$ plane. The optical field at a point along the z axis is then the superposition of contributions dE_P from all infinitesimal area elements dA. We may write this for convenience as,

$$dE_P = \frac{E_0}{\lambda}\frac{\sin(kr_z)}{r_z}\,dA, \tag{4.24}$$

where E_0 is the optical field per unit area, k is the wave vector and

$$r_z = \left(x^2+y^2+z^2\right)^{1/2}. \tag{4.25}$$

In cylindrical polar coordinates,

$$r_z = \left(r^2+z^2\right)^{1/2} \tag{4.26}$$

and

$$dE_P = \frac{E_0}{\lambda}\frac{\sin(kr_z)}{r_z}\,rdrd\phi. \tag{4.27}$$

Integrating Eq. (4.27), we have

$$E_P = \frac{2\pi E_0}{\lambda}\int_0^\infty \frac{\sin(kr_z)}{r_z}\,rdr. \tag{4.28}$$

Note that

$$dr_z = r\left(r^2+z^2\right)^{-1/2}dr = \frac{rdr}{r_z}. \tag{4.29}$$

Hence,

$$E_P = \frac{2\pi E_0}{\lambda}\int_{r=0}^\infty \sin(kr_z)\;dr_z = \frac{2\pi E_0}{k\lambda}\left[-\cos(kr_z)\right]_{r=0}^\infty. \tag{4.30}$$

Since $k=2\pi/\lambda$ and $r_z=z$ when $r=0$, we have

$$E_P(z) = E_0\cos(kz). \tag{4.31}$$

This is the equation of a *plane wave* propagating in the z direction. Thus, the propagation is along a straight line.

4.8.2 Interference of plane waves

Let us now reconsider the problem of plane wave propagation from a central wave front. Allowing for interference, wavefronts that travelled some arbitrary distance z in either direction before returning, would yield a phase

$$E_P(z) = E_0 \cos(2kz). \tag{4.32}$$

Integrating this over all possible values of z gives

$$\int_{r=0}^{\infty} E_P(z)\ dz = E_0 \int_{r=0}^{\infty} \cos(2kz)\ dz. \tag{4.33}$$

The total interfering field is therefore

$$E_0 \int_{r=0}^{\infty} \cos(2kz)\ dz = \frac{E_0}{2k} \left[\sin(2kz)\right]_0^{\infty} = 0. \tag{4.34}$$

In other words, the spurious back propagation of the wave front is cancelled out via interference.

4.9 Summary

- **Spherical waves**

  Spherical waves have the form

  $$\mathbf{E}(r,t) = \frac{\mathbf{E}_0}{r} \sin(\omega t - kr) \tag{4.35}$$

  or

  $$\mathbf{E}(r,t) = \frac{\mathbf{E}_0}{r} e^{i\omega t} \sin(kr), \tag{4.36}$$

  where r is the distance from the centre of the sphere.

- **Huygen's Principle**

  Each point on a wavefront acts as a source of secondary, spherical wavelets. At a later time, t, a new wavefront is constructed from the sum of these wavelets.

- **The laws of optical propagation**

  Huygen's Principle may be applied to derive:

  - *The law of rectilinear propagation*
    In a homogeneous medium, light travels in straight lines.
  - *The law of reflection*
    In a homogeneous incident medium, the angle of incidence equals the angle of reflection.
  - *The law of refraction*
    When light passes from a homogeneous medium with refractive index n_i into another homogeneous medium with refractive index n_t , the angles of incidence and refraction, θ_i and θ_t, are given by Snell's Law

$$\boxed{n_i \sin\theta_i = n_t \sin\theta_t.}$$

- **Coherence**

  Two waveforms are said to be *coherent* if there is a fixed relative phase between them.

- **The Huygens-Fresnel Principle**

  The Huygens-Fresnel Principle may be stated as

  - *For light of a given frequency, every point on a wavefront acts as a secondary source of spherical wavelets with the same frequency and the same initial phase. The wavefront at a later time and position is then the linear superposition of all of these wavelets.*

  The Huygens-Fresnel Principle introduces *interference* that cancels out the spurious back propagation of a wave front allowed by Huygens' Principle alone.

4.10 References

[1] *Trait de la lumire*, Christiaan Huygens, Leiden, Netherlands: Pieter van der Aa (1690)

[2] *Memoir on the Diffraction of Light*, Augustin Fresnel, Academe of Science (1819)

5. Diffraction

5.1 General remarks

In Chapter 4 we considered the phenomenon of the *interference* of coherent waves. A term that may, in principle, be used synonymously with interference is *diffraction*. However, in practice, diffraction is more often used to describe interference effects that exhibit clearly delineated regions of constructive or destructive interference. In this chapter, we shall focus on the diffraction of light as it passes through an array of narrow apertures in a screen. The same analysis may also be used to describe diffraction from scattering features on a surface. In both cases, we often refer to the screen (or surface) as a *diffraction grating*.

The diffraction grating was discovered by the Scottish mathematician James Gregory around 1670, a year or so after Newton's prism experiments, by passing sunlight through a bird's feather. In a letter to John Collins [1], Gregory recommends his experiment for Newton's consideration:

> *If ye think fit, ye may signify to Mr. Newton a small experiment, which (if he know it not already) may be worthy of his consideration. Let in the sun's light by a small hole to a darkened house, and at the hole place a feather, (the more delicate and white the better for this purpose,) and it shall direct to a white wall opposite to it a number of small circles and ovals, (if I mistake them not,) whereof one is somewhat white, (to wit, the middle, which is opposite to the sun,) and all the rest severally coloured.*

Another example of diffraction in nature (best explained together with the closely related phenomenon of *thin-film interference*) is the iridescence of a butterflies wing, the colours of which are seen to change as the wing is moved.

Many other examples and applications can be mentioned, some of which will be covered in this chapter.

5.2 Learning objectives

- Diffraction regimes: Fresnel (near-field) and Fraunhofer (far-field)

- Analysis of Fraunhofer diffraction
  - Single-slit
  - Double-slit
- Fraunhofer diffraction from a circular aperture
  - The Airy disc
  - Rayleigh criterion
- Analysis of multiple slit diffraction in the far field
- The grating equation
- Use of diffraction gratings in monochromators.

5.3 Light passing through a narrow aperture

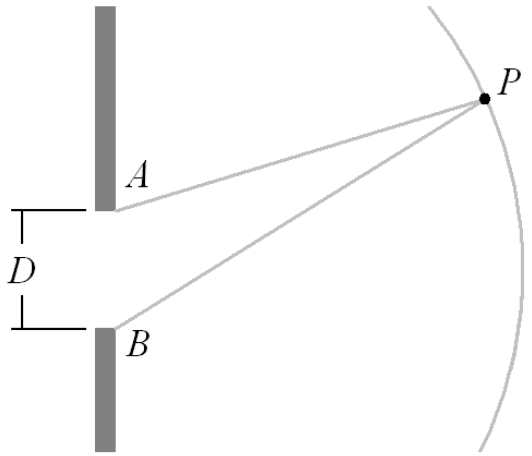


Figure 5.1: Light passing through a narrow slit of width D in a screen.

Imagine a plane waves of a single wavelength approach a screen in which there is a narrow aperture. As light passes through the aperture, each point on the emerging wavefront acts as a source of secondary wavelets in accordance with the Huygen's-Fresnel Principle. Beyond the screen, these secondary waves interfere to some extent or other to produce the phenomenon of *diffraction*.

Figure 5.1 illustrates the situation for a narrow slit of width D. We consider a general point at P and consider how the contributions from each

wavelet at the slit combines to produce the resultant optical field E_P. Now at the aperture, all points will be in phase. In a medium of constant refractive index, the phase at P from each wavelet will be a function of the geometric path alone. The resultant interference at P will then be a linear superposition of the phases of each contribution.

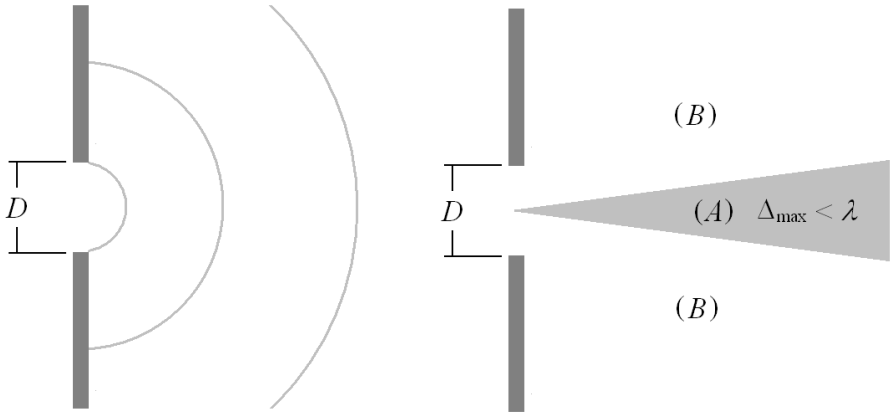


Figure 5.2: Diffraction cases for (a) $\lambda \gg D$ and (b) $\lambda \gg D$.

Consider the maximum possible path difference $\Delta_{\max}$ due to the path-lengths from the end points of the slit at A and B to P, as shown in Fig. 5.1.

$$\Delta_{\max} = \left|\vec{AP} - \vec{BP}\right| = \left|\vec{AB}\right| = D. \tag{5.1}$$

Let us consider two limiting cases.

Case: $\lambda \gg D$.

In this case, the maximum path length difference $\Delta_{\max}$ is always less than λ. The wavelets emanating from the slit therefore add constructively in all directions and the emergent optical field looks like a point source. This is illustrated in Fig. 5.2 (a).

Case: $\lambda \ll D$.

When $\lambda \ll D$, the wavelets only add constructively within a narrow solid angle subtended by the slit where $\Delta_{\max} < \lambda$. This is the shaded region labeled A in Fig. 5.2 (b). Outside of this region (in the areas labeled B) the optical field adds both constructively and destructively leading to a complicated diffraction pattern.

In the general case, we may differentiate two distinct regions of the optical field beyond the aperture. Closer to the screen, we have the *near field*

region, where the diffraction pattern varies considerably with increasing distance from the aperture. This region is characterised by *Fresnel diffraction*. Further from the screen as the radial distance from the aperture R becomes very large compared to D, we have the *far field* region, in which the diffraction pattern settles down to a constant profile. This is *Fraunhofer diffraction*, which we shall be analysing in the next sections. A more precise criterion for the onset of Fraunhofer diffraction is given in Sec. 5.4.

5.4 Single slit diffraction

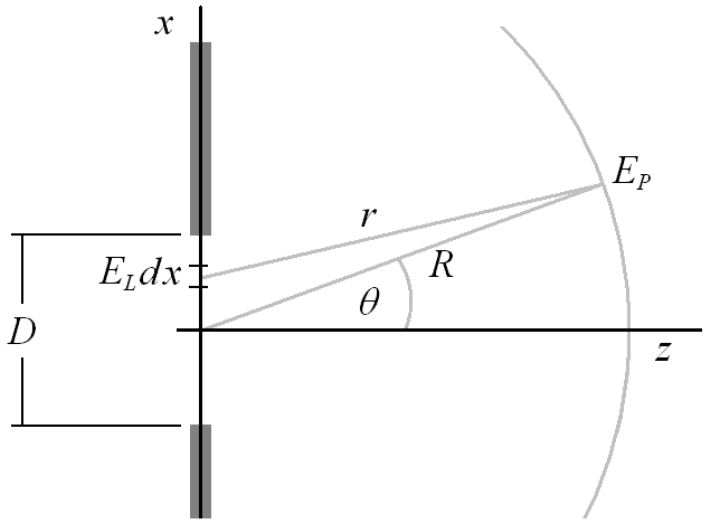


Figure 5.3: Geometry of single slit diffraction in the far field for light normally incident on the screen.

In this section, we shall analyse far-field (Fraunhofer) diffraction from a single slit in a screen. Figure 5.3 shows the geometry for a slit of width D, through which monochromatic light of wavelength λ passes. In this case, the light approaching the screen on the left is *normally incident* to it. That is, the wavefronts of the incident light are *parallel* to the screen. Note that since the electric field E_P is for a point lying on a semicircle a distance R from the centre of the slit, the angle θ must take a value on the interval $[-\pi/2, \pi/2]$.

We can characterise the electric field at the slit by defining the field strength per unit length E_L, so that the actual field at x is $dE(x) = E_L dx$. The contribution to the electric field E_P is then just the spherical wave emanating from the infinitesimal region dx

$$dE_P = \frac{E_L}{r(x)} \sin\left[\omega t - kr\left(x\right)\right]\ dx, \tag{5.2}$$

where $k = 2\pi/\lambda$. The total field E_P is then the integral of Eq. (5.2) over the slit width D

$$E_P = \int_{-D/2}^{D/2} \frac{E_L}{r(x)} \sin\left[\omega t - kr(x)\right]\ dx. \tag{5.3}$$

From Fig. 5.3 we see that $r(x)$ is given by the cosine rule

$$r^2(x) = R^2 + x^2 - 2Rx\cos\left(\tfrac{\pi}{2} - \theta\right), \tag{5.4}$$

so

$$r(x) = R\left[1 + \frac{x^2}{R^2} - \frac{2x}{R}\sin\theta\right]^{1/2}. \tag{5.5}$$

Now the Taylor expansion for a function $(1+\xi)^{1/2}$ is

$$(1+\xi)^{1/2} = 1 + \frac{\xi}{2} - \frac{\xi^2}{8} + \ldots, \tag{5.6}$$

so

$$\begin{aligned} r(x) &= R\left\{1 + \frac{1}{2}\left[\frac{x^2}{R^2} - \frac{2x}{R}\sin\theta\right] - \frac{1}{8}\left[\frac{x^2}{R^2} - \frac{2x}{R}\sin\theta\right]^2 + \ldots\right\}, \\ &= R\left\{1 + \frac{x^2}{2R^2} - \frac{x}{R}\sin\theta - \frac{x^2}{2R^2}\sin^2\theta + \ldots\right\}, \\ &= R\left\{1 - \frac{x}{R}\sin\theta + \frac{x^2}{2R^2}\cos^2\theta + \ldots\right\}. \end{aligned} \tag{5.7}$$

5.4.1 The Fraunhofer condition

We may make the condition for Fraunhofer diffraction a little more precise by inspecting the phase term $kr(x)$ in more detail. From Eq. (5.7), we have

$$kr(x) = kR - kx\sin\theta + \frac{kx^2}{2R}\cos^2\theta + \ldots. \tag{5.8}$$

Now the third term in Eq. (5.8) takes its maximum value when $x = \pm D/2$ and $\theta = 0$. That is

$$\frac{kD^2}{8R} = \frac{\pi D^2}{4\lambda R}. \tag{5.9}$$

The condition that this term makes a negligible contribution to the phase is

$$\frac{\pi D^2}{4\lambda R} \ll \pi. \tag{5.10}$$

Neglecting the factor of 4 in the denominator on the left-hand-side, this may be re-written

$$\boxed{\frac{D}{R} \ll \frac{\lambda}{D}.} \tag{5.11}$$

This then is the *Fraunhofer condition*.

5.4.2 Far field approximation

Assuming that the Fraunhofer condition applies, we may approximate $kr(x)$ as

$$kr(x) \approx kR - kx\sin\theta. \tag{5.12}$$

In Eq. (5.3), it is only the phase of the $\sin$ function that is particularly sensitive to variations in $r(x)$ and for this we use the approximation above. For the $1/r(x)$ term, we may simply substitute $1/R$. With these approximations, Eq. (5.3) becomes

$$E_P = \int_{-D/2}^{D/2} \frac{E_L}{R} \sin\left[\omega t - kR + kx\sin\theta\right]\ dx. \tag{5.13}$$

To perform this integral, we note that

$$\sin\left[\omega t - kR + kx\sin\theta\right] = \mathrm{Im}\left\{e^{i[\omega t - kR + kx\sin\theta]}\right\}, \tag{5.14}$$

so, performing the integration over the x-dependent part of the complex exponent, we have

$$\int_{-D/2}^{D/2} e^{ikx\sin\theta}\ dx = \left[\frac{e^{ikx\sin\theta}}{ik\sin\theta}\right]_{-D/2}^{D/2} = D\frac{\sin\beta}{\beta}, \tag{5.15}$$

where

$$\beta = \frac{kD}{2}\sin\theta. \tag{5.16}$$

Hence, the total field E_P is

$$E_P = \frac{E_L D}{R}\frac{\sin\beta}{\beta}\sin\left(\omega t - kR\right). \tag{5.17}$$

Now the squared modulus of E_P is proportional to the intensity, so

$$I(\theta,t) \propto |E_P|^2 = \left|\frac{E_L D}{R}\right|^2 \left|\frac{\sin\beta}{\beta}\right|^2 \sin^2(\omega t - kR). \tag{5.18}$$

Hence, taking the time average of this, we obtain

$$\boxed{I(\theta) = I(0)\left|\frac{\sin\beta}{\beta}\right|^2}. \tag{5.19}$$

Figure 5.4 shows plots of Eq. (5.19) for λ/D ratios of 1/2 and 1/4. Both curves exhibit a large central peak of intensity surrounded by smaller peaks going out to larger angles. The zeros between the peaks occur at values of

$$\beta = \frac{kD}{2}\sin\theta = m\pi, \tag{5.20}$$

where m is an integer. Hence, the zeros around the central peak are given by

$$\sin\theta = \frac{\lambda}{D}. \tag{5.21}$$

Note that this result is only valid for $\lambda \leq D$. For ratios of $\lambda/D > 1$, β never reaches $\pm\pi$ due to the limits on θ. Hence Eq. (5.19) has no zeros. In this case, we have for $\lambda/D \gg 1$

$$\lim_{\lambda/D\to\infty}\left|\frac{\sin\beta}{\beta}\right|^2 = \lim_{\beta\to 0}\left|\frac{\sin\beta}{\beta}\right|^2 = 1. \tag{5.22}$$

Taken with the result of Eq. (5.22), Eq. (5.21) therefore gives a measure of the angular spread of the central diffraction peak. For $\lambda/D \gg 1$, the diffraction pattern approaches that for a point source, i.e. a spherical wave, and the intensity varies little across all values of θ. As λ becomes less than D, the central peak becomes narrower, approaching a delta function for $\lambda/D \ll 1$.

5.5 Diffraction limited imaging

5.5.1 Diffraction from a circular aperture

The analysis for a circular aperture proceeds along the same lines as that for the single slit, except that we now need to integrate over an area to get the total contribution to the field. Taking the origin to be at the centre of the aperture and the z-axis to be perpendicular to plane of the aperture, the problem may be rendered in spherical polar coordinates, with θ being the polar angle.

The result for the intensity is then found to be

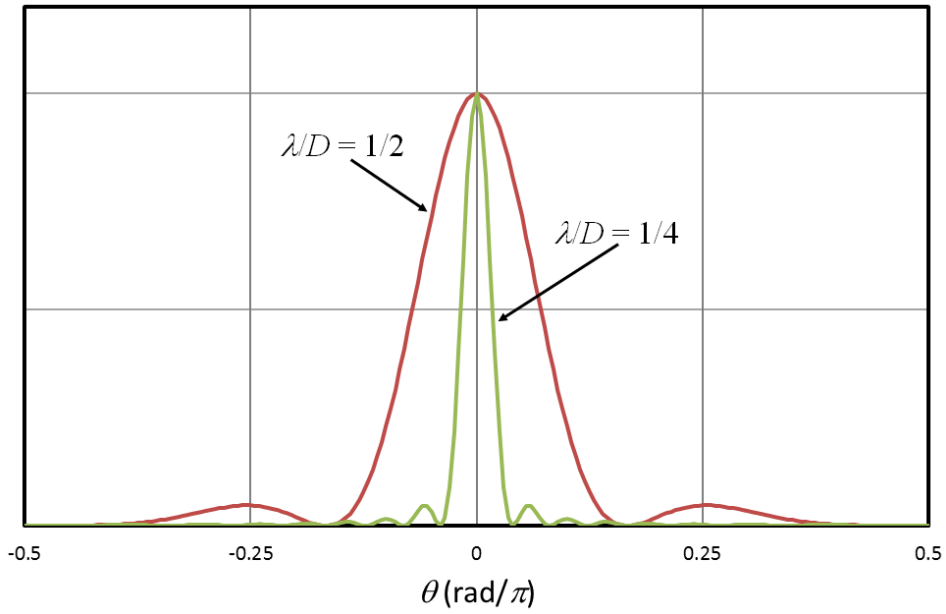


Figure 5.4: Far field diffraction patterns for $\lambda/D = 1/2$ and $\lambda/D = 1/4$.

$$\boxed{I(\theta) = I(0)\left[\frac{2J_1(kD\sin(\theta/2))}{kD\sin(\theta/2)}\right]^2}, \tag{5.23}$$

Figure 5.5 (a) shows an example of the diffraction pattern for a circular aperture. The central bright peak is known as the *Airy disc*, whilst the rings around it are referred to as *Airy rings*.
where D is now the diameter of the aperture and J_1 is the *first order Bessel function*.

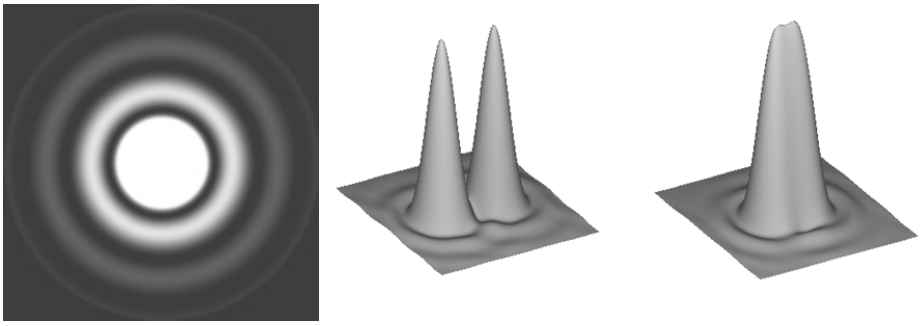
5.5.2 The Rayleigh criterion

The first zero of the intensity profile of Eq. (5.49) occurs at

$$\boxed{\sin\theta \approx 1.22\frac{\lambda}{D}}. \tag{5.24}$$

This has the same form as Eq. (5.21) for the single slit and the same observations about the spread of the central peak apply here.

Note that a distant point source will have the same diffraction pattern as that for a circular aperture. Thus, Eq. (5.24) may be used to establish limitations for the ability to resolve two such point sources, based on the overlap of the central peaks. Using the small angle approximation, E. (5.24) becomes



(a) (b) (c)

Figure 5.5: (a) Airy rings for a diffraction from a circular aperture. (b) The intensity profiles for two resolvable distant point sources. (c) Merged intensity profiles for unresolvable distant point sources.

$$\boxed{\theta_{\min} \approx 1.22 \frac{\lambda}{D}}. \tag{5.25}$$

The *Rayleigh criterion* for the resolution of two points is then that the angular separation between them must be greater than $\theta_{\min}$. This is illustrated in Figs. 5.5 (b) and (c).

5.6 Multiple slit diffraction

5.6.1 Normal incidence

Figure 5.6 shows the geometry for diffraction through multiple slits. The analysis of this situation in the far field proceeds similarly to that of single slit diffraction. Taking $N = 2$ then produces Young's well-known case for the double slit.

For N slits, the total contribution to the field E_P is

$$E_P = \sum_{n=0}^{N-1} \int_{na-D/2}^{na+D/2} dE_P. \tag{5.26}$$

Applying the approximations for Fraunhofer diffraction, this becomes

$$E_P = \frac{E_L}{R} \sum_{n=0}^{N-1} \int_{na-D/2}^{na+D/2} \sin\left(\omega t - kR + xk \sin\theta\right)\, dx. \tag{5.27}$$

Focusing on the x-dependent part of this integral, we have

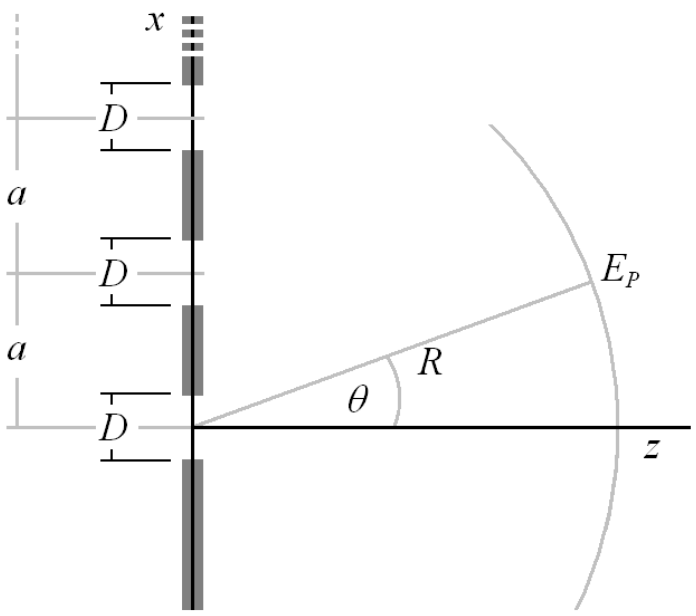


Figure 5.6: Geometry for multiple slit diffraction in the far field.

$$\sum_{n=0}^{N-1}\left[\frac{e^{ikx\sin\theta}}{ik\sin\theta}\right]_{na-D/2}^{na+D/2} = \sum_{n=0}^{N-1} e^{iknd\sin\theta} D\frac{\sin\left[(kD/2)\sin\theta\right]}{(kD/2)\sin\theta} = \sum_{n=0}^{N-1} e^{in2\alpha} D\frac{\sin\beta}{\beta}, \tag{5.28}$$

where we have pulled out the factor for the single slit and defined

$$\alpha = \frac{ka}{2}\sin\theta \tag{5.29}$$

and

$$\beta = \frac{kD}{2}\sin\theta. \tag{5.30}$$

The remaining factor in Eq. (5.28) is a geometric progression with common factor $e^{i2\alpha}$

$$S_N = \sum_{n=0}^{N-1} e^{i2n\alpha}. \tag{5.31}$$

Multiplying this by $e^{i2\alpha}$ gives

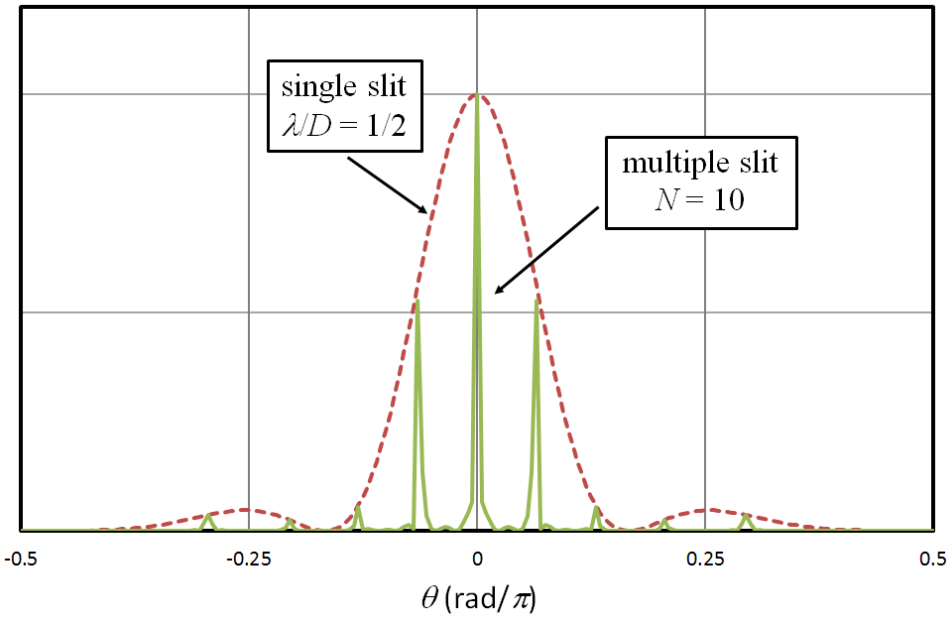


Figure 5.7: Far field diffraction patterns for a single slit with $\lambda/D = 1/2$ (dashed line) and a multiple slit with $a/D = 5/2$ and $N = 10$.

$$S_N e^{i2\alpha} = \sum_{n=1}^{N} e^{i2n\alpha}. \tag{5.32}$$

Subtracting Eq. (5.32) from Eq. (5.31), we have

$$S_N \left(1 - e^{i2\alpha}\right) = 1 - e^{i2N\alpha} \tag{5.33}$$

which gives for S_N

$$\begin{aligned} S_N &= \frac{1 - e^{i2N\alpha}}{1 - e^{i2\alpha}} \\ &= \frac{e^{iN\alpha}\left(e^{-iN\alpha} - e^{iN\alpha}\right)}{e^{i\alpha}\left(e^{-i\alpha} - e^{i\alpha}\right)} = e^{i(N-1)\alpha}\frac{\sin N\alpha}{\sin \alpha}. \end{aligned} \tag{5.34}$$

The phase factor $e^{i(N-1)\alpha}$ may be dropped out of Eq. (5.34) once the squared modulus is taken. Note that since

$$\lim_{\alpha \to 0} \frac{\sin N\alpha}{\sin \alpha} = N, \tag{5.35}$$

it is useful to explicitly incorporate a normalising factor $1/N$ into this ratio (rather than to implicitly include it in $I(0)$). Hence, the intensity takes the form

$$\boxed{I(\theta) = I(0)\left(\frac{\sin N\alpha}{N\sin\alpha}\right)^2\left(\frac{\sin\beta}{\beta}\right)^2.} \tag{5.36}$$

An example of this diffraction pattern is shown in Fig. 5.7 against the pattern for a single slit with $\lambda/D = 1/2$ (dashed line) for values of $a/D = 5/2$ and $N = 10$.

As λ/D becomes very large, the enveloping single slit pattern broadens out and the interference pattern becomes a series of sharp peaks. Now it can be shown that the maxima of

$$\left(\frac{\sin N\alpha}{N\sin\alpha}\right)^2 \tag{5.37}$$

occur when α is an integral multiple of π, i.e. $\alpha = m\pi$, where m is an integer known as the *order*. Adding m as an index to θ, we therefore have

$$\alpha = \frac{ka}{2}\sin\theta_m = \frac{a\pi}{\lambda}\sin\theta_m = m\pi, \tag{5.38}$$

which may be rearranged to give the *condition for constructive interference*

$$\boxed{m\lambda = a\sin\theta_m.} \tag{5.39}$$

This is the *grating equation* for a transmission grating with light of normal incidence (see Fig. 5.8).

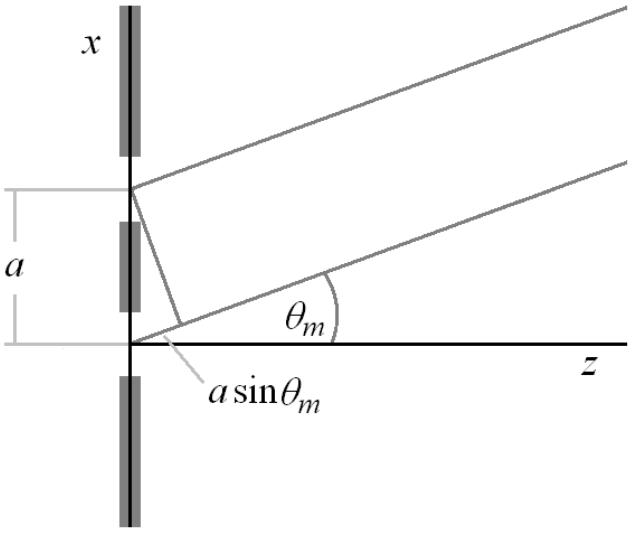


Figure 5.8: The condition for constructive interference of normally incident light passing through a transmission grating.

5.6.2 Off-axis incidence

In the foregoing analysis, the incident light was normal to the grating (which we may refer to as 'on-axis'). In the more general case, the incident light may be at an angle to the grating normal (i.e. 'off-axis'). This is illustrated in Fig. 5.9. Note that the angles θ_i' and θ_i'' are equal, giving the magnitude of the angle between the incident light and the grating normal. Measuring the angle anti-clockwise from the z-axis, we have

$$\theta_i = \pi - \theta_i'. \tag{5.40}$$

However, this makes no difference to the general result for *transmission*

$$\boxed{m\lambda = a\left(\sin\theta_m + \sin\theta_i\right).} \tag{5.41}$$

Putting $\theta = \pi$ then gives the previous result for normal incidence.

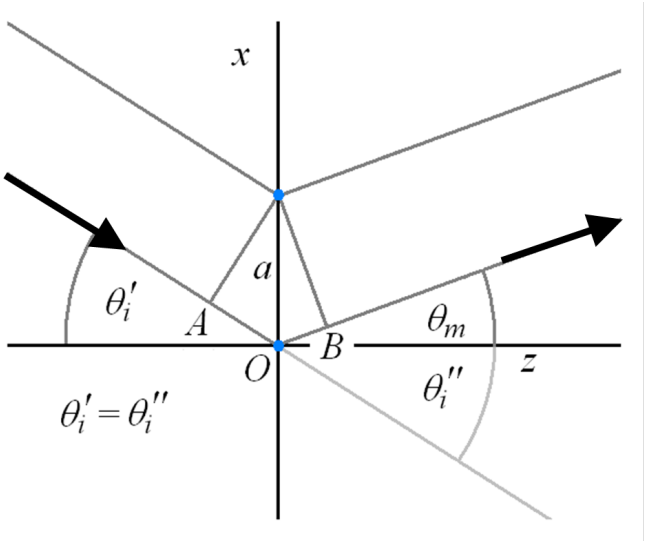


Figure 5.9: The condition for constructive interference for off-axis incident light passing through a transmission grating.

5.7 Diffraction gratings

For a *reflection grating*, the incident light arrives on the same side as the diffracted light. Here, the sources of interfering spherical waves will be perturbations on the grating that give rise to *scattering*.

Equation (5.41) still holds in this case. However, since θ_i is now negative, the sign is usually brought out of the sinusoidal term to give

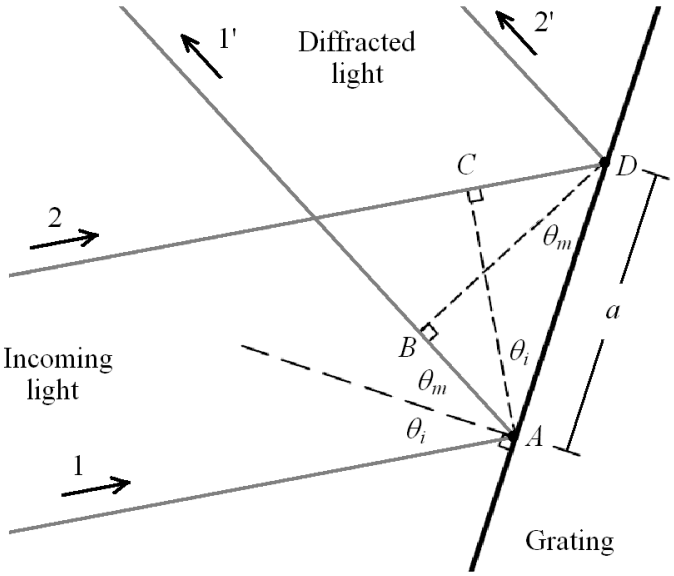


Figure 5.10: Sketch of a diffraction grating showing incoming light being diffracted from two successive rulings of the grating (at A and D).

$$\boxed{m\lambda = a\left(\sin\theta_m - \sin\theta_i\right).} \tag{5.42}$$

This is referred to as the the *reflection grating equation*.

Figure 5.10 shows the geometric set-up for part of a diffraction grating consisting of lines (or rulings) separated by a distance a (only two rulings are illustrated, located at points A and D in the figure). The width of these rulings should be small enough that light striking them is re-radiated as a cylindrical wave in the same way as light passing through narrow slits. Thus, for a given wavelength, there will be certain angles for which light from these cylindrical waves will be in-phase and may therefore create a diffracted wavefront.

As common example of a diffraction grating, Fig. 5.11 shows light diffracted into its spectral components by the grooves of a compact disc. Angling the disc relative to the incident light changes the observed colours.

Although this phenomenon is often referred to as *iridescence*, iridescence is more accurately defined in terms of *thin-film interference*. Although closely related to diffraction, thin-film interference concerns the reflection and refraction of light through layers of material with different refractive indices. This is discussed in detail in the next chapter, where we see that, for instance, the shimmering colours of a butterfly's wings are due to this effect.

Figure 5.11: Diffraction of light from the grooves of a compact disc.

5.7.1 Resolving power

We may quantify the ability of a diffraction grating to distinguish different wavelengths of light via the definition of the *resolving power* R. Earlier, we saw that the ability to resolve spatially separated points of light was given by the *Rayleigh criterion*. We may define the wavelength resolution analogously. Peaks associated with different wavelengths will be diffracted at different angles, leading to a spatial displacement. Applying Rayleigh's criterion to these peaks allows us to define a minimum resolvable displacement $\lambda_2 - \lambda_1 = \Delta\lambda$ in wavelength space. If λ is the wavelength between λ_2 and λ_1, then the *resolving power* of the grating is given by

$$\boxed{R = \frac{\lambda}{\Delta\lambda}.} \tag{5.43}$$

In words, this relation says

> *The higher the resolving power of a grating, the smaller the wavelength difference it can distinguish*

As a particular example, the resolving power of an N-ruling grating is given by

$$R = mN. \tag{5.44}$$

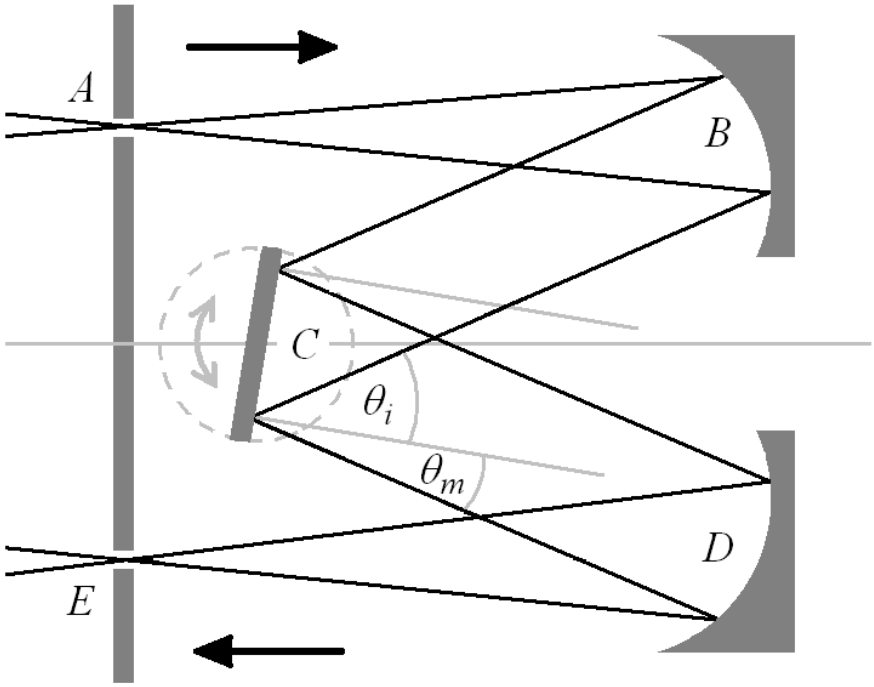


Figure 5.12: Schematic of a Czerny-Turner monochromator.

5.7.2 Monochromators

Czerny-Turner monochromator

Figure 5.12 shows a schematic of a Czerny-Turner monochromator. Polychromatic light (light containing many frequencies) enters the monochromator via a slit at A. This is then focused into plane waves by a concave mirror known as the *collimator* at B and directed towards the diffraction grating at C. Ideally, the collimator should have a parabolic surface to perfectly image the input light without aberration. At the grating, only light meeting the diffraction condition for the input angle θ_i and output angle θ_m are diffracted in a collimated beam towards the centre of a second mirror at D known as the *camera*. The camera provides the inverse operation of the collimator, focusing the light to a point at the exit slit E. Wavelengths that did not satisfy the diffraction condition for these angles are reflected away from the output path. The output light will therefore only contain components within a narrow wavelength range.

The slits and mirrors are kept fixed in this design, whilst the grating may be rotated to meet the diffraction condition for different wavelengths. Thus the wavelength (or frequency) decomposition of the input light may be

ascertained from the output of a detector placed at the exit slit. Calibration of the monochromator is achieved using monochromatic light of a known wavelengths and adjusting the rotation of the grating until the maximum intensity output is obtained.

Optical spectrum analyser

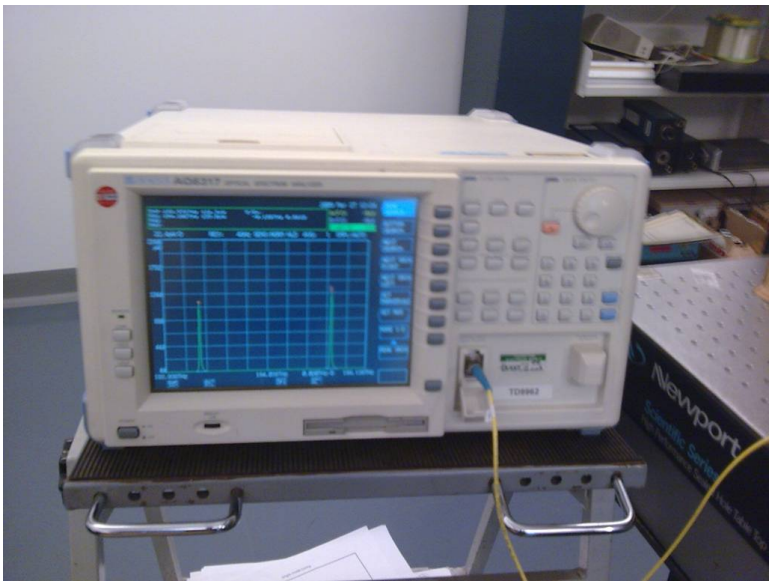


Figure 5.13: Photograph of an optical spectrum analyser in the laboratory setting.

The monochromator is the basis for the more sophisticated *optical spectrum analyser* (OSA). At its heart, an OSA just has a diffraction grating (or set of gratings) mounted on an automated motor. Light is coupled into the device via fibre optic patch cable and focused onto the grating. As the grating is rotated, an optical detector measures the optical signal at some fixed angle and plots the intensity of the signal against wavelength (or frequency). Various additional facilities may also be integrated into the device. Figure 5.13 shows an OSA as part of an experimental setup to characterise semiconductor lasers.

5.8 Diffraction around objects

We have seen how a wave may spread out as it is diffracted through a narrow aperture. An analogous situation pertains when light encounters an opaque object: the wavefronts may now diffract *around* the object. Some

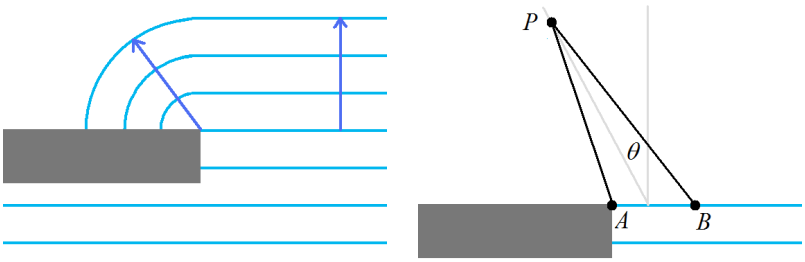


Figure 5.14: (a) Diffraction around the side of an object according to Huygens' Principle. (b) Including interference via the Huygens-Fresnel Principle.

insight into this phenomenon is furnished by Huygens' Principle, as illustrated in Fig. 5.14 (a). Here we see how light at the edge of the object, acting as a point source of a wavelet, propagates around the object. This however, is not the full story. According to such a picture , *all* waveforms ought to be diffracted around the object by $\pi/2$. The fact that this does not happen is explained via *interference*.

In Fig. 5.14 (b) we see a similar picture to the earlier diagrams of light passing through a slit, in which the condition for constructive interference at P is met by the wavelength of light being greater than the maximum path difference

$$\lambda > \left|\vec{AB}\right|. \tag{5.45}$$

Now as λ is reduced, so must the maximum path difference to maintain constructive interference. Hence, the angle θ will also tend to be reduced, limited the angular spread of the wavefront.

For an object in a train of wavefronts, this means that if the size of the object is comparable to the wavelength, the waves will tend to diffract right around the object, as illustrated in Fig. 5.15.

5.9 Summary

- For the diffraction pattern beyond an aperture in a screen through which light passes, we may define to distinct regions:
  - *Near field or Fresnel diffraction*
    The diffraction pattern varies considerably with increasing distance from the aperture.

Figure 5.15: (a) Waves encountering an object with dimensions that are greater than the wavelength. (b) Diffraction around an object of dimensions comparable with the wavelength.

  - *Far field or Fraunhofer diffraction*
    The diffraction pattern settles down to a constant profile.

- **Analysis of Fraunhofer diffraction**

  - *Single-slit*
    The intensity of the far-field diffraction pattern from a slit of width D in a barrier may be written as

$$\boxed{I(\theta) = I(0)\left|\frac{\sin\beta}{\beta}\right|^2}, \tag{5.46}$$

    where

$$\beta = \frac{kD}{2}\sin\theta, \tag{5.47}$$

    k is the wave-vector and θ is the angle from the normal to the barrier.
  - *The Fraunhofer condition*
    The condition for the validity of the Fraunhofer treatment is given by

$$\boxed{\frac{D}{R} \ll \frac{\lambda}{D}.} \tag{5.48}$$

    where λ is the wavelength and R is the radial distance from the slit.

- **Fraunhofer diffraction from a circular aperture**

– *The Airy disc*

The intensity for far-field diffraction from a circular aperture of diameter D is given by

$$\boxed{I(\theta) = I(0)\left[\frac{2J_1(kD\sin(\theta/2))}{kD\sin(\theta/2)}\right]^2}, \tag{5.49}$$

where J_1 is the first order Bessel function.

The central peak of this diffraction pattern is known as the *Airy disc* and the bright interference rings around it are called *Airy rings*.

– *The Rayleigh criterion*

In order to resolve two points sources, then the angular separation between them must be greater than

$$\boxed{\theta_{\min} \approx 1.22\frac{\lambda}{D}}. \tag{5.50}$$

- **Multiple slit diffraction**

The intensity profile of multiple slit diffraction in the far field is given by

$$\boxed{I(\theta) = I(0)\left(\frac{\sin N\alpha}{N\sin\alpha}\right)^2\left(\frac{\sin\beta}{\beta}\right)^2}. \tag{5.51}$$

where N is the number of slits,

$$\alpha = \frac{ka}{2}\sin\theta, \tag{5.52}$$

$$\beta = \frac{kD}{2}\sin\theta, \tag{5.53}$$

k is the wave-vector, a is the slit spacing and D is the slit width.

- **The grating equation**

$$\boxed{m\lambda = a(\sin\theta_m + \sin\theta_i)}. \tag{5.54}$$

where m is an integer and λ is the wavelength, gives the condition for the local maxima for multiple slit diffraction.

- **Resolving power of a grating**

$$\boxed{R = \frac{\lambda}{\Delta\lambda}.} \tag{5.55}$$

- **Monochromators**

  Diffraction gratings may be used in **monochromators**, such as the *Czerny-Turner* design, to measure the frequency composition of polychromatic light.

- **Optical spectrum analyser**

  The monochromator may be integrated into an automated unit known as an *optical spectrum analyser*.

5.10 References

[1] *Correspondence of Scientific Men of the Seventeenth Century....,* vol **2**, Ed. Stephen Jordan Rigaud, Oxford, England: Oxford University Press (1841)

Part III

Electromagnetic Waves

6. Wave Solutions to Maxwell's Equations

6.1 General remarks

Alongside the quantum mechanical explanation of light, Maxwell's prediction of the generation of electromagnetic (EM) waves is key to our understanding of optical phenomena. Furnishing Thomas Young's conclusive observations of the wave nature of light with a physical explanation, Maxwell predicted [1] in 1865 that electric and magnetic fields would mutually induce each other, propagating with a constant speed in a vacuum given by $c = (\varepsilon_0 \mu_0)^{-1/2}$, reproducing the measured speed of light. In turn, ε_0 and μ_0 are both fundamental physical constants, determining the response of the vacuum to electric and magnetic fields respectively. Maxwell's prediction was then later confirmed by Hertz in a paper of 1892 [2] reporting the generation of *radio waves* in the laboratory setting.

Although it was initially believed that EM waves would require some kind of physical medium for their propagation (the 'luminous ether'), this was later shown by Einstein to be superfluous to requirements in his 1905 paper on Special Relativity [3]. Einstein was also a leading figure in the early development of *quantum theory*. Despite the apparently contradictory nature of the so-called 'wave-particle' duality of light, the two pictures are, in fact, complimentary to one another. In this Chapter, we shall reiterate the relation between the classical theory of optical absorption and the quantum mechanical explanation.

A crucial aspect of this Chapter is the introduction to the *electric susceptibility tensor*. Although we shall currently limit our consideration to linear, isotropic and homogeneous media, the explanation of optical phenomena in *anisotropic media* (to be covered in the Chapters on *Crystal Optics*) will depend heavily on our understanding of the susceptibility tensor.

Lastly in this Chapter, we shall consider the flux of *electromagnetic energy*. We specify this in terms of the *Poynting vector*. Again, we shall anticipate subtleties in our understanding of the energy flux to arise in the context of anisotropic media.

6.2 Learning objectives

The aims of this section are to understand

- Wave solutions of Maxwell's equations in free space and dielectric media
- The electric and magnetic susceptibility tensors
- Relative permittivity and permeability
- The refractive index of a medium
- Frequency dependence of the electric susceptibility
- Optical loss in media and the relation to the photon picture
- The Poynting vector for the electromagnetic energy flux

6.3 Maxwell's equations

The physics of electromagnetism is encapsulated in Maxwell's equations. In differential form, these are

$$\nabla \cdot \mathbf{D} = \rho_f, \tag{6.1}$$
$$\nabla \cdot \mathbf{B} = 0, \tag{6.2}$$
$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{6.3}$$
$$\nabla \times \mathbf{H} = \mathbf{j}_f + \frac{\partial \mathbf{D}}{\partial t}, \tag{6.4}$$

The first of these, Eq. (6.1), is *Gauss' Law*, stating that the divergence of the *electric displacement* $\mathbf{D}$ is equal to the *free charge density* ρ_f. In other words, the free charges are the sources or sinks $\mathbf{D}$. The electric displacement is related to the *electric field* $\mathbf{E}$ and the *electrical polarisation* $\mathbf{P}$ of a material medium due to $\mathbf{E}$ via

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \tag{6.5}$$

where ε_0 is a constant known as the *permittivity of free space*.

The second of Maxwell's laws, Eq. (6.2), states that the divergence of the *magnetic field* $\mathbf{B}$ is zero. In other words $\mathbf{B}$ has no sources or sinks

(magnetic monopoles). Pictorially, this means that every field line of the magnetic field eventually joins up with itself in a loop.

Equation (6.3) is *Faraday's Law of electromagnetic induction*, which says that a changing magnetic field induces a non-conservative electric field. A conservative field $\mathbf{F}$ is a vector field related to a scalar potential Φ via $\mathbf{F} = -\nabla\Phi$. However, the curl of this would be $-\nabla \times \nabla\Phi$, which is identically zero. Meanwhile, $\nabla \cdot \nabla \times \mathbf{E}$ is also identically zero, meaning that the induced electric field has no source or sinks and that the field lines all join up in loops.

The fourth equation, Eq. (6.4) is *Ampere's Law* modified by the addition of the *displacement current density*, which Maxwell realised must be present to meet the requirement of charge conservation. Ampere's law relates the curl of the *field vector* $\mathbf{H}$ to the *free current density* $\mathbf{j}_f$. $\mathbf{H}$ is related to $\mathbf{B}$ and the *magnetisation* $\mathbf{M}$ of a material medium by

$$\mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M}, \tag{6.6}$$

where μ_0 is a constant known as the *permeability of free space*.

6.4 Electromagnetic waves in a vacuum

6.4.1 The wave equation

We consider solutions to Maxwell's equations in a vacuum in the absence of any free charge. Since there will therefore be no material polarisation or magnetisation, Eqs. (6.5) and (6.6) reduce to

$$\mathbf{D} = \varepsilon_0\mathbf{E} \tag{6.7}$$

and

$$\mathbf{H} = \frac{1}{\mu_0}\mathbf{B}. \tag{6.8}$$

In addition, the free charge density ρ_f and free current density $\mathbf{j}_f$ will both be zero. Maxwell's equations therefore reduce to

$$\nabla \cdot \mathbf{E} = 0, \tag{6.9}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{6.10}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{6.11}$$

$$\nabla \times \mathbf{B} = \varepsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t}, \tag{6.12}$$

Taking the curl of Eq. (6.11), we have

$$\nabla \times \nabla \times \mathbf{E} = -\frac{\partial (\nabla \times \mathbf{B})}{\partial t}. \tag{6.13}$$

Using Eq. (6.12) to substitute for $\nabla \times \mathbf{B}$, we then have

$$\nabla \times \nabla \times \mathbf{E} = -\varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{6.14}$$

We may then make use of the identity (see Appendix **??**)

$$\nabla \times \nabla \times \mathbf{E} = \nabla (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}, \tag{6.15}$$

noting that by Eq. (6.9) $\nabla \cdot \mathbf{E} = 0$, to obtain

$$\boxed{\nabla^2 \mathbf{E} = \varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}.} \tag{6.16}$$

A similar argument leads to the wave equation for the magnetic field

$$\boxed{\nabla^2 \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{B}}{\partial t^2}.} \tag{6.17}$$

Comparison with the general wave equation Eq. (3.24) of Chapter 3, reveals that the speed of light in a vacuum is

$$\boxed{c = (\varepsilon_0 \mu_0)^{-1/2},} \tag{6.18}$$

which is, of course, a universal constant.

6.4.2 Plane wave solutions

We shall assume plane wave solutions of Eq. (6.16) or Eq. (6.17) of the form

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \tag{6.19}$$

Here, $\mathbf{E}_0$ is known as a *Jones vector* and contains information about the polarisation of the wave. We shall discuss Jones vectors in more detail in Chapter 7.

Given solutions of this form, we are in a position to make the substitutions

$$\nabla^2 \to -|\mathbf{k}|^2, \tag{6.20}$$

$$\frac{\partial}{\partial t} \to -i\omega, \tag{6.21}$$

$$\frac{\partial^2}{\partial t^2} \rightarrow -\omega^2 \tag{6.22}$$

and

$$\nabla\times \rightarrow i\mathbf{k}\times . \tag{6.23}$$

The wave equation then becomes

$$|\mathbf{k}|^2\,\mathbf{E} = \omega^2\varepsilon_0\mu_0\mathbf{E}, \tag{6.24}$$

which gives

$$c = \frac{\omega}{|\mathbf{k}|}, \tag{6.25}$$

as we would require for plane waves in a vacuum.

We can also use Eqs. (6.21) and (6.23) with Eq. (6.11) to obtain

$$\mathbf{B} = \frac{\mathbf{k}}{\omega}\times\mathbf{E}, \tag{6.26}$$

implying that $\mathbf{B}$ is perpendicular to both $\mathbf{k}$ and $\mathbf{E}$. Similarly, from Eq. 6.12, we have

$$\mathbf{E} = c^2\mathbf{B}\times\frac{\mathbf{k}}{\omega}, \tag{6.27}$$

implying that $\mathbf{E}$ is perpendicular to both $\mathbf{k}$ and $\mathbf{B}$. Hence, $\mathbf{E}$ and $\mathbf{B}$ are both at right angles to the wavevector (and to each other), so these waves are *transverse*.

6.5 Electromagnetic waves in a material medium

6.5.1 The electric susceptibility tensor

In a material medium, we must take account of the electrical and magnetic response to an applied field. Thus, the electrical polarisation $\mathbf{P}$ induced by an applied electric field $\mathbf{E}$ is normally given by

$$\mathbf{P} = \varepsilon_0\boldsymbol{\chi}_E\mathbf{E}, \tag{6.28}$$

where $\boldsymbol{\chi}_E$ is the *electric susceptibility tensor*. Equation (6.28) may be referred to as a *constitutive equation* for the polarisation in terms of $\boldsymbol{\chi}_E$. However, the form of this equation does not make obvious that the susceptibility tensor characterises the *frequency response* of the material to an applied field. The *temporal response* of the medium may be encapsulated in a response function $\mathbf{R}(t)$ and the electrical polarisation then rendered in the form

$$\mathbf{P}(t) = \varepsilon_0 \int_{-\infty}^{\infty} \mathbf{R}(t-\tau)\,\mathbf{E}(\tau)\ d\tau. \tag{6.29}$$

Making the change of variable $\tau' = t - \tau$, this may be re-written

$$\mathbf{P}(t) = \varepsilon_0 \int_{-\infty}^{\infty} \mathbf{R}(\tau')\,\mathbf{E}(t-\tau')\ d\tau'. \tag{6.30}$$

Now, we may express $\mathbf{E}(t)$ in terms of its Fourier transform $\mathbf{E}_\omega$ via

$$\mathbf{E}(t) = \int_{-\infty}^{\infty} \mathbf{E}_\omega e^{-i\omega t}\ d\omega. \tag{6.31}$$

So, substituting this into Eq. (6.30) gives

$$\mathbf{P}(t) = \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{R}(\tau')\,\mathbf{E}_\omega e^{-i\omega(t-\tau')}\ d\omega\ d\tau'. \tag{6.32}$$

We may then define $\boldsymbol{\chi}_E$ as the *Fourier transform of* $\mathbf{R}(\tau)$

$$\boldsymbol{\chi}_E = \int_{-\infty}^{\infty} \mathbf{R}(\tau)\, e^{i\omega\tau}\ d\tau \tag{6.33}$$

to obtain

$$\mathbf{P}(t) = \varepsilon_0 \int_{-\infty}^{\infty} \boldsymbol{\chi}_E \mathbf{E}_\omega e^{-i\omega t}\ d\omega. \tag{6.34}$$

Since we also have

$$\mathbf{P}(t) = \int_{-\infty}^{\infty} \mathbf{P}_\omega e^{-i\omega t}\ d\omega, \tag{6.35}$$

we may omit the integration to obtain simply

$$\mathbf{P}_\omega = \varepsilon_0 \boldsymbol{\chi}_E \mathbf{E}_\omega. \tag{6.36}$$

Hence, we regain Eq. (6.28), which we now see is actually in the frequency domain. However, to avoid excessive use of subscripts, we shall usually take this to be tacit.

6.5.2 Linear, isotropic and homogeneous dielectric

Linearity

In general, we may express $\boldsymbol{\chi}_E$ in tensorial form via

$$P_i = \varepsilon_0 \sum_{ij} \chi^{(1)}_{ij} E_j + \varepsilon_0 \sum_{ijk} \chi^{(2)}_{ijk} E_j E_k + \varepsilon_0 \sum_{ijkl} \chi^{(3)}_{ijkl} E_j E_k E_l + \ldots \tag{6.37}$$

Here, $\chi_{ij}^{(1)}$ depends only *linearly* on the applied field whilst the components with superscripts (n) for $n > 1$ are *non-linear* in $\mathbf{E}$. We shall assume throughout that these non-linear terms are small and neglect them. Thus, we assume that the material is *linear* and put

$$(\boldsymbol{\chi}_E)_{ij} = \chi_{ij}^{(1)}. \tag{6.38}$$

In the case of very high intensity light, the electric field is large and these non-linear terms may need to be incorporated. This is the subject matter of *non-linear optics*.

There is also a constitutive equation analogous to Eq. (6.28) for the magnetisation of the material

$$\mathbf{M} = \frac{\boldsymbol{\chi}_B}{\mu_0}\mathbf{B}. \tag{6.39}$$

The *magnetic susceptibility tensor* $\boldsymbol{\chi}_B$ may be dealt with formally in the same way as $\boldsymbol{\chi}_E$, however, the magnetisation is usually very small.

Isotropy and homogeneity

The assumptions of *isotropy* (the same in all directions) and *homogeneity* (the same at every position) mean that the form of $\chi_{ij}^{(1)}$ is constant under changes of orientation and origin. Thus, $\chi_{ij}^{(1)}$ must be a *diagonal matrix* with all elements the same. That is

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_E & 0 & 0 \\ 0 & \chi_E & 0 \\ 0 & 0 & \chi_E \end{bmatrix}. \tag{6.40}$$

Note that, if $\mathbf{I}$ is the identity matrix, this relation is just

$$\boldsymbol{\chi}_E = \chi_E \mathbf{I}. \tag{6.41}$$

Wave solutions

In a *dielectric*, i.e. an electrically insulating medium, we may take ρ_f and $\mathbf{j}_f$ to be zero. Equations (6.1) and (6.4) now become

$$\nabla \cdot \mathbf{D} = 0 \tag{6.42}$$

and

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}. \tag{6.43}$$

In the meantime, Eq. (6.5) becomes

$$\mathbf{D} = \varepsilon_0 \left(\mathbf{I} + \boldsymbol{\chi}_E\right) \mathbf{E} = \varepsilon_0 \boldsymbol{\varepsilon} \mathbf{E}, \tag{6.44}$$

where we have defined the *relative permittivity* $\boldsymbol{\varepsilon}$.

Similarly, the field vector $\mathbf{H}$ becomes

$$\mathbf{H} = \frac{1}{\mu_0} \left(\mathbf{I} - \boldsymbol{\chi}_B\right) \mathbf{B} = \frac{\boldsymbol{\mu}^{-1}}{\mu_0} \mathbf{B}. \tag{6.45}$$

Here, $\boldsymbol{\mu}$ is the *relative permeability*.

In an isotropic medium, the susceptibilities reduce to scalars and Eqs. (6.42) and (6.43) simplify to

$$\nabla \cdot \mathbf{E} = 0 \tag{6.46}$$

and

$$\nabla \times \mathbf{B} = \varepsilon \varepsilon_0 \mu \mu_0 \frac{\partial \mathbf{E}}{\partial t}. \tag{6.47}$$

The wave equation is now

$$\nabla^2 \mathbf{E} = \varepsilon \varepsilon_0 \mu \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{6.48}$$

and with an analogous result for $\mathbf{B}$. This is almost in the same form as the equation for waves in a vacuum except that now the wave speed is

$$v = \left(\varepsilon \varepsilon_0 \mu \mu_0\right)^{-1/2} = \frac{c}{n}. \tag{6.49}$$

where

$$\boxed{n = (\varepsilon \mu)^{1/2}} \tag{6.50}$$

is the *refractive index*.

Note that the wave-vector $\mathbf{k}$ is related to the free space wave-vector $\mathbf{k}_0$ at the same frequency by

$$\mathbf{k} = n \mathbf{k}_0. \tag{6.51}$$

6.6 Frequency dependence of the electric susceptibility

The electric susceptibility tensor χ_E gives the response of a medium to an applied electric field (here, we will consider only the *linear* response). The effect of the applied electric field in a dielectric is to displace the atomic or molecular charges, inducing electric dipoles and hence polarising the

medium. The induced polarisation $\mathbf{P}$ is then given in terms of the driving field by Eq. (6.28), where $\boldsymbol{\chi}_E$ is a tensor.

For an electromagnetic wave propagating through a medium, the applied field is time-dependent. In general, we shall need to consider three forces acting on the charge density

- The *driving force* via the time-dependent applied field
- The *restoring force* of the harmonic oscillators
- (possibly) a *damping force* acting in opposition to the motion

The dynamics of the charge density within the medium may then be modelled classically in terms of a damped, driven oscillator. Now the polarisation $\mathbf{P}$ may be written in terms of the N *electric dipoles* $\mathbf{p}_i$ of the medium

$$\mathbf{P} = \sum_i \mathbf{p}_i = \sum_i q_i \mathbf{u}_i, \tag{6.52}$$

where q_i is the charge displaced from equilibrium by $\mathbf{u}_i$ for the ith dipole. The equation of motion of the system may now be written as

$$\sum_i \left[\frac{d^2\mathbf{p}_i}{dt^2} + \omega_i^2 \mathbf{p}_i + \frac{1}{\tau}\frac{d\mathbf{p}_i}{dt} \right] = \sum_i \frac{q_i^2 \mathbf{E}_0}{m_i} e^{i\omega t}, \tag{6.53}$$

where ω_i, q_i and m_i are the *resonant frequency*, *charge* and *mass* of the ith oscillator respectively and ω is the *driving frequency*. The time constant τ_i characterises the *dampening force* acting on the ith oscillator and, as we shall see, is associated with the *absorption* of a quantum of radiation.

Applying Eq. (6.28) to Eq. (6.53) gives

$$\varepsilon_0 \boldsymbol{\chi}_E \left[\ddot{\mathbf{E}} + \frac{1}{\tau}\dot{\mathbf{E}} + \omega_0^2 \mathbf{E} \right] = \sum_i \frac{q_i^2 \mathbf{E}_0}{m_i} e^{i\omega t}. \tag{6.54}$$

We shall assume plane wave solutions in the steady state $\mathbf{E} = \mathbf{E}_0 e^{i\omega t}$, so that Eq. (6.54) becomes

$$\varepsilon_0 \boldsymbol{\chi}_E \left[\left(\omega_0^2 - \omega^2\right) + \frac{i\omega}{\tau} \right] \mathbf{E}_0 = \sum_i \frac{q_i^2 \mathbf{E}_0}{m_i}. \tag{6.55}$$

In an *isotropic medium*, χ_E is diagonal with all diagonal elements equal. Thus, we have

$$\varepsilon_0 \chi_E = \sum_i \frac{q_i^2/m_i}{\left(\omega_i^2 - \omega^2\right) + i\omega/\tau_i}. \tag{6.56}$$

Note that χ_E is a *complex quantity*.

6.6.1 Relative permittivity

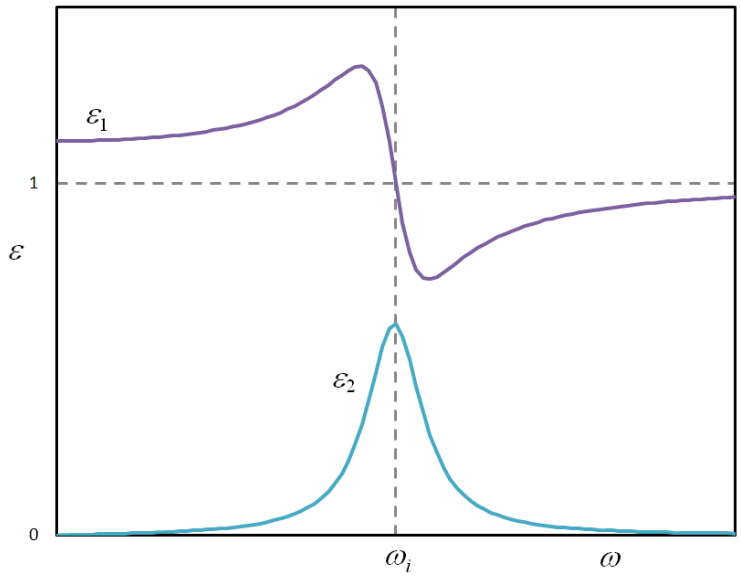


Figure 6.1: Curves showing the real and imaginary parts of the relative permittivity near an absorption resonance (at ω_i).

The relative permittivity is given in terms of χ_E by

$$\varepsilon = 1 + \chi_E. \tag{6.57}$$

Expressing this in terms of real and imaginary parts, this may be written

$$\begin{aligned} \varepsilon &= 1 + \mathrm{Re}\left(\chi_E\right) + i\mathrm{Im}\left(\chi_E\right), \\ &= \varepsilon_1 - i\varepsilon_2, \end{aligned} \tag{6.58}$$

where, putting $C_i = q_i^2/m_i$,

$$\varepsilon_1 = 1 + \varepsilon_0^{-1} \sum_i \frac{C_i\left(\omega_i^2 - \omega^2\right)}{\left(\omega_i^2 - \omega^2\right)^2 + \omega^2/\tau_i^2} \tag{6.59}$$

and

$$\varepsilon_2 = \varepsilon_0^{-1} \sum_i \frac{C_i\omega/\tau_i}{\left(\omega_i^2 - \omega^2\right)^2 + \omega^2/\tau_i^2}. \tag{6.60}$$

Curves illustrating the ω dependence of ε_1 and ε_2 are shown in Fig. 6.1. Note that the parameter $1/\tau_i$ gives a measure of the broadening of ε_2, which, as we shall see in the next sub-section, is an *absorption resonance*.

That is, the material medium shows a strong propensity to absorb some of the energy of the optical field at this frequency.

The frequency dependence of ε leads to a frequency dependence of the refractive index and, in turn, the wave speed. This leads to the phenomenon of *dispersion*, in which the different frequency components of the electromagnetic field in a medium spread out from one another due to their different wave speeds.

6.6.2 Sellmeier's equation

The dispersion of the medium may be described by giving the refractive index in terms of the free space wavelength. Assuming that the broadening factors $1/\tau_i$ are small, using Eqs. (6.50) and (6.56) and taking the relative permeability $\mu = 1$, we have

$$n^2(\lambda) = 1 + \sum_i \frac{C_i}{\omega_i^2 - \omega^2}. \tag{6.61}$$

Using $\omega = 2\pi c/\lambda$, this becomes

$$n^2(\lambda) = 1 + \sum_i \frac{C_i \lambda_i^2 \lambda^2}{(2\pi c)^2 \left(\lambda^2 - \lambda_i^2\right)}. \tag{6.62}$$

Absorbing the various factors into the constants A_i, we then have

$$n^2(\lambda) = 1 + \sum_i \frac{A_i \lambda^2}{\lambda^2 - \lambda_i^2}. \tag{6.63}$$

This is known as *Sellmeier's equation* and the parameters A_i and λ_i are found empirically. Figure (6.2) shows an example for the refractive index of borosilicate crown glass. Note that the $n(\lambda)$ decreases with increasing wavelength. This is typically the case and in fact when the condition

$$\frac{dn}{d\lambda} < 0 \tag{6.64}$$

is satisfied, the medium is said to have *normal dispersion*. In cases where Eq. (6.64) is violated, we have *anomalous dispersion*.

6.7 Optical loss

6.7.1 The absorption coefficient

If the angular frequency ω of an optical field propagating in the medium is close to one of the resonances of the electric susceptibility, there will be a propensity for the energy of the field to be absorbed into the oscillating

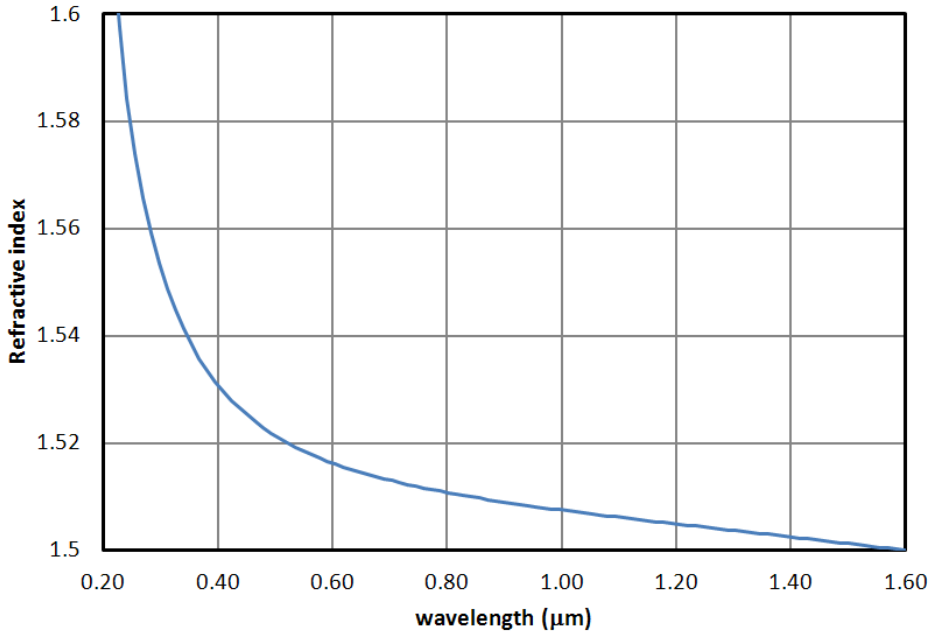


Figure 6.2: Graph of the refractive index of borosilicate crown glass using Sellmeier's equation with empirically fitted parameters.

charge. This may be formulated via consideration of the complex part of the relative permittivity as given in Eq. (6.58). If the relative permittivity is complex, then the refractive index will also become complex. We may write this as

$$n = n_0 - i\eta. \tag{6.65}$$

Consider then the wave solution

$$\mathbf{E}(z,t) = \mathbf{E}_0 \exp\left(i\omega\left[t - \frac{z}{v}\right]\right). \tag{6.66}$$

The wave speed v is given by

$$v = \frac{c}{n} = \frac{c}{n_0 - i\eta}. \tag{6.67}$$

Substituting this into Eq. (6.66) gives

$$\mathbf{E}(z,t) = \mathbf{E}_0 \exp\left(i\omega\left[t - \frac{n_0 z}{c}\right]\right) \exp\left(-\frac{\omega\eta z}{c}\right). \tag{6.68}$$

Now the intensity of the radiation $I(z)$ is proportional to the squared modulus of the field, i.e.

$$I(z) \propto |\mathbf{E}(z,t)|^2 = |\mathbf{E}_0|^2 \exp\left(-\frac{2\omega\eta z}{c}\right). \tag{6.69}$$

Hence, the intensity decreases exponentially with the distance propagated through the medium. The intensity may then be written as

$$I(z) = I(0)\, e^{-\alpha z}, \tag{6.70}$$

where

$$\boxed{\alpha = \frac{2\omega\eta}{c}} \tag{6.71}$$

is the *absorption coefficient*. Note that this is exactly the same result obtained in Chapter **??** for the photon description of light.

6.7.2 Photon absorption

The process of optical absorption just described has an immediate quantum mechanical interpretation. Recall that the energy of a photon is given by $\epsilon = \hbar\omega$. In terms of resonant frequency, we may consistently interpret the energy difference between electronic states as

$$\epsilon = \hbar\omega_i. \tag{6.72}$$

The absorption of a photon is then just the transition from an electronic state ϵ_1 to a higher state $\epsilon_2 = \epsilon_1 + \hbar\omega_i$.

Energy-time Uncertainty Principle

Inspecting Fig. 6.1, we see that the energy of the transition is not sharply defined. This may be explained in terms of the *energy-time Uncertainty Principle*. In terms of angular frequency, this is

$$\Delta\epsilon \sim \frac{1}{\tau}. \tag{6.73}$$

In other words, τ may be interpreted as the *time uncertainty*. Equivalently, τ is proportional to the *energy uncertainty*. This may be understood in terms of the *broadening* of the electronic energy levels.

6.8 Time symmetry

A clear consequence of optical loss is that we *do not have time-reversibility*. That is, changing the direction of time does not preserve the same physics. In one temporal direction, we have an exponentially *decaying* amplitude, in

the other an exponentially *increasing* amplitude. This latter situation does model the physically real phenomena of *optical gain* in a laser, although this does involve the additional requirements of optical feedback and population inversion, as briefly described in Chapter 2.

Mathematically, we can trace this to the appearance of the imaginary component of the susceptibility. On the other hand, if this component is taken to be zero then we have a *lossless* medium and we *do have* time reversibility - at least in the case of the linear susceptibility that we have investigated.

To further elucidate the consequences of time reversibility it is useful to invoke a notation sometimes used in texts on nonlinear optics. One of the characteristic features of the nonlinear optics is the emergence of new frequencies due to the coupling of nonlinear powers of the electric field. So for instance, a component of the second order susceptibility might be written

$$\chi^{(2)}\left(-\omega_i;\omega_1,\omega_2\right), \tag{6.74}$$

where

$$\omega_i = \omega_1 + \omega_2. \tag{6.75}$$

(This arises out of the conservation of energy, where the energy of a photon is $\hbar\omega$). For the linear susceptibility, we would have, by analogy

$$\chi^{(1)}\left(-\omega_i;\omega_j\right), \tag{6.76}$$

where

$$\omega_i = \omega_j. \tag{6.77}$$

Since we shall normally limit our consideration to the linear case, this notation is rather too cumbersome for general use. However, we will find it useful for making clear the implications of time reversibility. As a first step, however, let us write the angular frequency in a more general form as a complex entity

$$\omega_j = \omega_{j,R} + i\omega_{j,I} \tag{6.78}$$

and substitute this into Eq. (6.33) to obtain, for the linear susceptibility

$$\chi^{(1)}\left(-\omega_i;\omega_j\right) = \int_{-\infty}^{\infty} \mathbf{R}\left(\tau\right) e^{i\left(\omega_{j,R}+i\omega_{j,I}\right)\tau}\, d\tau \tag{6.79}$$

Taking the complex conjugate of this, we have

$$\chi^{(1)*}\left(-\omega_i;\omega_j\right) = \int_{-\infty}^{\infty} \mathbf{R}^*\left(t\right) e^{-i\left(\omega_{j,R}-i\omega_{j,I}\right)\tau}\, d\tau. \tag{6.80}$$

Now, since $\omega_j^* = \omega_{j,R} - i\omega_{j,I}$, if $\mathbf{R}(t)$ is real, this implies

$$\chi^{(1)*}(-\omega_i;\omega_j) = \chi^{(1)}\left(i\omega_i^*;-i\omega_j^*\right). \tag{6.81}$$

Since the expression on the right-hand-side is just saying $i\omega_i^* = i\omega_j^*$, we may cancel the common factor of i and rewrite the expression as

$$\chi^{(1)*}(-\omega_i;\omega_j) = \chi^{(1)}\left(-\omega_j^*;\omega_i^*\right). \tag{6.82}$$

If we now re-enforce the constraint that ω should only have real values, we then have

$$\chi^{(1)*}(-\omega_i;\omega_j) = \chi^{(1)}(-\omega_j;\omega_i). \tag{6.83}$$

Now, we can write out the Eq. (6.28) as

$$P_i e^{i\omega t} = \varepsilon_0 \sum_j \chi^{(1)}(-\omega_i;\omega_j) E_j e^{i\omega t}, \tag{6.84}$$

(remembering that this is in the frequency representation). Applying time symmetry means putting $t \to -t$. This gives

$$P_i e^{-i\omega t} = \varepsilon_0 \sum_j \chi^{(1)}(-\omega_i;\omega_j) E_j e^{-i\omega t}. \tag{6.85}$$

With ω, P_i and E_j constrained to be real, this is equivalent to

$$\left[P_i e^{i\omega t}\right]^* = \varepsilon_0 \sum_j \chi^{(1)}(-\omega_i;\omega_j)\left[E_j e^{i\omega t}\right]^*. \tag{6.86}$$

However, since

$$\left[P_i e^{i\omega t}\right]^* = \varepsilon_0 \sum_j \chi^{(1)*}(-\omega_i;\omega_j)\left[E_j e^{i\omega t}\right]^*, \tag{6.87}$$

this implies

$$\chi^{(1)}(-\omega_i;\omega_j) = \chi^{(1)*}(-\omega_i;\omega_j). \tag{6.88}$$

Using Eq. (6.83), this means

$$\chi^{(1)}(-\omega_i;\omega_j) = \chi^{(1)}(-\omega_j;\omega_i). \tag{6.89}$$

In other words

- *In a lossless material, the linear electric susceptibility tensor is a symmetric matrix*.

This result has a number of useful consequences. We will encounter one of this in Section 6.10 on the *Poynting vector*. Further consequences of more general symmetries will be explored in Chapter 11.

6.9 Dispersion

In earlier Chapters, we took it as a general feature of wave propagation that different wavelengths may see different refractive indices in transparent media. This then modifies the wave speed from c to c/n. In this Chapter, we have seen how this phenomenon may be accounted for in the case of electromagnetic waves in terms of the frequency dependency of the electric susceptibility. Moreover, we have seen that, although the frequency is unchanged by the refractive index, the *free space wavevector* $\mathbf{k}_0 \rightarrow \mathbf{k} = n\mathbf{k}_0$ (in an isotropic medium).

This frequency dependence on the refractive index leads to *dispersion*, in which the different components of light move through a medium at different speeds. This also affects the degree to which different components of the light are *refracted* as it passes from one medium to another in accordance with Snell's Law.

6.9.1 The rainbow

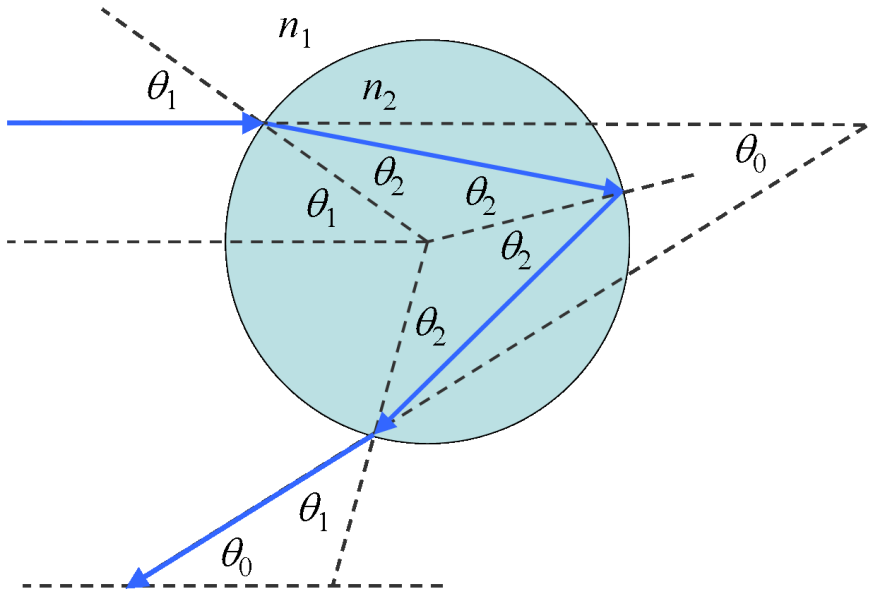


Figure 6.3: Light from the Sun transmitting into a spherical droplet of water, reflecting once and transmitting at an angle θ_0 to the horizontal.

Perhaps the most beautiful common example of dispersion is that of the *rainbow*, often seen on rainy days when the Sun also appears. Figure 6.3 shows the case for a single frequency component of the light from the Sun transmitting into a spherical droplet of water, reflecting once and transmitting out again at an angle θ_0 to the horizontal.

From inspection, θ_0 is found to be given by

$$\theta_0 = 4\theta_2 - 2\theta_1. \tag{6.90}$$

Note that θ_0 does not vary monotonically with the input angle θ_1 but has a *maximum* value. To determine this, we may differentiate Eq. (6.90) with respect to θ_1 and set the result equal to zero. That is,

$$\frac{d\theta_0}{d\theta_1} = 4\frac{d\theta_2}{d\theta_1} - 2 = 0. \tag{6.91}$$

To evaluate $d\theta_2/d\theta_1$, we note that

$$\frac{d\sin^{-1} x}{d\theta} = -\left(1-x^2\right)^{-1/2}\frac{dx}{d\theta}. \tag{6.92}$$

So, since

$$\theta_2 = \sin^{-1}\left(\frac{n_1}{n_2}\sin\theta_1\right), \tag{6.93}$$

we have

$$\begin{aligned}\frac{d\theta_2}{d\theta_1} &= \frac{(n_1/n_2)\cos\theta_1}{\left[1-(n_1/n_2)^2\sin^2\theta_1\right]^{1/2}}, \\ &= \left[\frac{1-\sin^2\theta_1}{(n_2/n_1)^2-\sin^2\theta_1}\right]^{1/2}\end{aligned} \tag{6.94}$$

Substituting this into Eq. (6.91), we have

$$4\left[\frac{1-\sin^2\theta_1}{(n_2/n_1)^2-\sin^2\theta_1}\right]^{1/2} = 2, \tag{6.95}$$

which yields

$$\theta_1 = \sin^{-1}\left\{\left[\frac{4}{3}-\frac{1}{3}\left(\frac{n_2}{n_1}\right)^2\right]^{1/2}\right\}. \tag{6.96}$$

Additionally, Snell's Law gives

$$\theta_2 = \sin^{-1}\left(\frac{n_1}{n_2}\sin\theta_1\right) = \sin^{-1}\left\{\frac{n_1}{n_2}\left[\frac{4}{3}-\frac{1}{3}\left(\frac{n_2}{n_1}\right)^2\right]^{1/2}\right\}. \tag{6.97}$$

Thus,

$$\theta_{0,\max} = 4\sin^{-1}\left(\frac{n_1}{n_2}\eta\right) - 2\sin^{-1}(\eta)\,. \tag{6.98}$$

where

$$\eta = \left[\frac{4}{3} - \frac{1}{3}\left(\frac{n_2}{n_1}\right)^2\right]^{1/2}. \tag{6.99}$$

Taking $n_1 = 1$ and $n_2 = 1.333$ (a typical average value for water), we find

$$\boxed{\theta_{0,\max} = 42^\circ.} \tag{6.100}$$

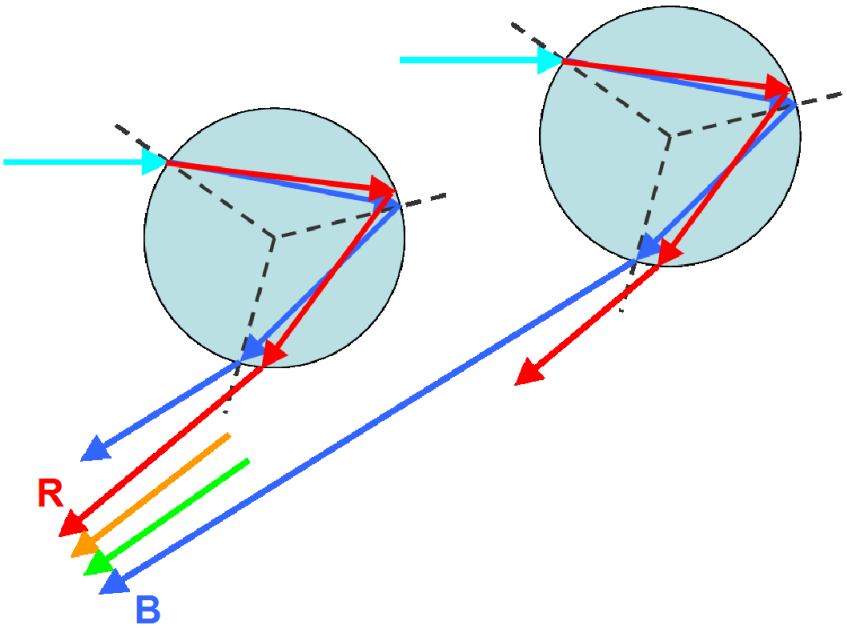


Figure 6.4: Since blue light has a greater refractive index than red in water, it is refracted to a greater degree and emerges at a shallower angle to the horizontal. Light from many droplets then builds up a pattern seen as a bow from the ground, with red light at the outer rim.

Dispersion of water

In the optical range, water exhibits *normal dispersion*, which means that violet light ($\sim 400\,\mathrm{nm}$) sees a higher refractive index than red light ($\sim 700\,\mathrm{nm}$). As a consequence, violet (and then blue) light is refracted to a greater degree than red by a spherical water droplet, as shown in Fig. 6.4. As a result, the deviation of red light has the greatest angle to the horizontal.

From a particular viewing point, an observer will see light refracted from many droplets with the net appearance of the rainbow with red light lying around the outer rim.

In fact, if the viewer was at a high enough altitude (in practice, in an aircraft), the rainbow would appear as a complete circle. On ground level, however, the lower portion of the circle is cut off and we just see a 'bow'.

6.10 The Poynting vector

The conservation of electromagnetic energy may be stated in terms of *Poynting's theorem*

$$-\frac{\partial u}{\partial t} = \nabla \cdot \mathbf{S} + \mathbf{j}_f \cdot \mathbf{E}, \tag{6.101}$$

where u is the energy density

$$u = \tfrac{1}{2}\left(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}\right), \tag{6.102}$$

and $\mathbf{S}$ is the *Poynting vector* giving the flow of energy crossing unit area. Hence, $-\partial u/\partial t$ is the rate at which the energy density decreases.

Taking the time derivative of Eq. (6.102), $\nabla \cdot \mathbf{S}$ is the divergence of the energy flow (the volume integral of this being the energy flux through the surface of the volume) and $\mathbf{j}_f \cdot \mathbf{E}$ is the work done on any free charges by the electric field.

The time derivative of Eq. (6.102) is

$$\frac{\partial u}{\partial t} = \frac{1}{2}\left(\frac{\partial \mathbf{E}}{\partial t}\cdot\mathbf{D} + \mathbf{E}\cdot\frac{\partial \mathbf{D}}{\partial t} + \frac{\partial \mathbf{B}}{\partial t}\cdot\mathbf{H} + \mathbf{B}\cdot\frac{\partial \mathbf{H}}{\partial t}\right). \tag{6.103}$$

We shall be assuming a linear and homogeneous medium but allowing it to be anisotropic. We shall therefore need to make use of the constitutive equations, Eqs. (6.44) and (6.45) for the electric displacement $\mathbf{D}$ and field vector $\mathbf{H}$ respectively.

Since the susceptibility tensors have no time dependence we have, from Eq. (6.44),

$$\frac{\partial \mathbf{D}}{\partial t} = \varepsilon_0\left(\mathbf{I} + \boldsymbol{\chi}_E\right)\frac{\partial \mathbf{E}}{\partial t}. \tag{6.104}$$

Now, the ith component of the matrix product of $\boldsymbol{\chi}_E$ and the time derivative of $\mathbf{E}$ is

$$\left(\boldsymbol{\chi}_E \frac{\partial \mathbf{E}}{\partial t}\right)_i = \sum_j \chi_{ij}\frac{\partial E_j}{\partial t}. \tag{6.105}$$

Hence

$$\mathbf{E}\cdot\boldsymbol{\chi}_E\frac{\partial\mathbf{E}}{\partial t}=\sum_i E_i\left(\boldsymbol{\chi}_E\frac{\partial\mathbf{E}}{\partial t}\right)_i=\sum_{ij}E_i\chi_{ij}\frac{\partial E_j}{\partial t}. \tag{6.106}$$

In Section 6.8, we found that in a loss-less medium, $\boldsymbol{\chi}_E$ is a symmetric matrix, i.e. $\chi_{ij}=\chi_{ji}$. So, since the scalar components commute,

$$\mathbf{E}\cdot\boldsymbol{\chi}_E\frac{\partial\mathbf{E}}{\partial t}=\sum_{ij}\frac{\partial E_j}{\partial t}\chi_{ji}E_i=\frac{\partial\mathbf{E}}{\partial t}\cdot\boldsymbol{\chi}_E\mathbf{E}. \tag{6.107}$$

Taking the dot product of $\mathbf{E}$ and Eq. (6.105), we therefore find

$$\begin{aligned}\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} &= \varepsilon_0\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t}+\varepsilon_0\mathbf{E}\cdot\boldsymbol{\chi}_E\frac{\partial\mathbf{E}}{\partial t},\\ &= \varepsilon_0\frac{\partial\mathbf{E}}{\partial t}\cdot\mathbf{E}+\varepsilon_0\frac{\partial\mathbf{E}}{\partial t}\cdot\boldsymbol{\chi}_E\mathbf{E},\\ &= \frac{\partial\mathbf{E}}{\partial t}\cdot\mathbf{D}.\end{aligned} \tag{6.108}$$

A similar argument holds for $\mathbf{B}$ and $\mathbf{H}$ yielding

$$\mathbf{B}\cdot\frac{\partial\mathbf{H}}{\partial t}=\frac{\partial\mathbf{B}}{\partial t}\cdot\mathbf{H}. \tag{6.109}$$

Using Eqs. (6.108) and (6.109), Eq. (6.103) for the time derivative of the energy density becomes

$$\frac{\partial u}{\partial t}=\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}+\frac{\partial\mathbf{B}}{\partial t}\cdot\mathbf{H}. \tag{6.110}$$

We now apply Faraday's law, Eq. (6.3), and Maxwell's modified form of Ampere's law, Eq. (6.4), to give

$$\begin{aligned}-\frac{\partial u}{\partial t} &= (\nabla\times\mathbf{E})\cdot\mathbf{H}-\mathbf{E}\cdot(\nabla\times\mathbf{H}-\mathbf{j}_f),\\ &= (\nabla\times\mathbf{E})\cdot\mathbf{H}-\mathbf{E}\cdot(\nabla\times\mathbf{H})+\mathbf{j}_f\cdot\mathbf{E}.\end{aligned} \tag{6.111}$$

Equating this to the right-hand-side of Eq. (6.101), we have

$$(\nabla\times\mathbf{E})\cdot\mathbf{H}-\mathbf{E}\cdot(\nabla\times\mathbf{H})=\nabla\cdot\mathbf{S}. \tag{6.112}$$

We now make use of the identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{H})\equiv(\nabla\times\mathbf{E})\cdot\mathbf{H}-\mathbf{E}\cdot(\nabla\times\mathbf{H}) \tag{6.113}$$

and hence arrive at the result

$$\boxed{\mathbf{S}=\mathbf{E}\times\mathbf{H}.} \tag{6.114}$$

This is the *instantaneous Poynting vector*, being the energy flowing across unit area per unit time.

Isotropic medium

In an isotropic medium, $\mathbf{S}$ is in the direction of the wave-vector $\mathbf{k}$ (although this is *not* true in the anisotropic case). It may then be shown that

$$\boxed{\mathbf{S} = \hat{\mathbf{k}} E^2 \left(\frac{\varepsilon\varepsilon_0}{\mu\mu_0} \right)^{1/2},} \tag{6.115}$$

where $\hat{\mathbf{k}}$ is the unit vector in the $\mathbf{k}$-direction. After time-averaging, we then have

$$\boxed{\langle \mathbf{S} \rangle = \tfrac{1}{2} \hat{\mathbf{k}} E_0^2 \left(\frac{\varepsilon\varepsilon_0}{\mu\mu_0} \right)^{1/2},} \tag{6.116}$$

where E_0 is the amplitude of the electric field component of the wave.

6.11 Summary

- **The speed of light**

  Maxwell's equations yield electromagnetic wave solutions. In free space, the wave speed is a universal constant

  $$\boxed{c = (\varepsilon_0\mu_0)^{-1/2}.} \tag{6.117}$$

- **The electric and magnetic susceptibility tensors**

  The electric and magnetic susceptibility tensors, $\boldsymbol{\chi}_E$ and $\boldsymbol{\chi}_B$, give the response of a medium to applied electric and magnetic fields respectively.

- **The relative permittivity ε and permeability μ**

  The relative permittivity ε and permeability μ of a medium is given in terms of the electric and magnetic susceptibilities

  $$\boxed{(\mathbf{I} + \boldsymbol{\chi}_E) = \boldsymbol{\varepsilon}} \tag{6.118}$$

  and

  $$\boxed{(\mathbf{I} - \boldsymbol{\chi}_B) = \boldsymbol{\mu}^{-1}.} \tag{6.119}$$

- **The wave speed and refractive index**

  The wave speed in a dielectric becomes modified according to

$$v = \frac{c}{n}, \tag{6.120}$$

  where

$$\boxed{n = (\varepsilon\mu)^{1/2}} \tag{6.121}$$

  is the **refractive index**.

- **Frequency dependence of the electric susceptibility**

  The electric susceptibility tensor χ_E has a frequency dependence and is complex. This means that the refractive index becomes frequency dependent and complex. The real part modifies the wave speed leading to *dispersion*. The dispersion is said to be normal if the condition

$$\frac{dn}{d\lambda} < 0 \tag{6.122}$$

  is satisfied.

- **Optical loss in media**

  The imaginary part of the refractive index implies a loss of energy from the optical field. The intensity of the field then decays exponentially as

$$I(z) = I(0)\, e^{-\alpha z}, \tag{6.123}$$

  where

$$\boxed{\alpha = \frac{2\omega\eta}{c}} \tag{6.124}$$

  is the **absorption coefficient**.

  This can be related to a transition between electronic states due to the absorption photons of the required energies.

- **Poynting vector**

  The electromagnetic power intensity (energy flow across unit area) is given by the *Poynting vector* $\mathbf{S}$ by

$$\boxed{\mathbf{S} = \mathbf{E} \times \mathbf{H}.} \tag{6.125}$$

  This result holds for an anisotropic medium.

6.12 References

[1] James Clerk Maxwell, *A Dynamical Theory of the Electromagnetic Field*, Philosophical Transactions of the Royal Society of London **155**, 459-512 (1865)

[2] Heinrich Hertz, *Untersuchungen über die Ausbreitung der elektrischen Kraft*, Johann Ambrosius Barth, Leipzig (1892)

[3] Albert Einstein, *Zur Elektrodynamik bewegter Körper*, Annalen der Physik **17**, 891 (1905)

7. Polarisation

7.1 General remarks

In Chapter 6, we saw how Maxwell's equations predict electromagnetic wave propagation. The polarisation of such radiation is characterised by the direction of the electric field. Although we have seen that the electric field directions must be transverse to the propagation direction, there is still a wealth of different possibilities to explore. In this Chapter we discuss the polarisation of EM waves in a fairly formal manner in terms of *Jones vectors*. The advantage of using Jones vectors is that

- the complex phase factor involving the spatial and temporal dependence (known as the *propagator* may often be abstracted from the formal presentation of the polarisation
- optical elements may be modelled by *Jones matrices* operating on the vectors

The physics of the optical elements is only dealt with in passing, since such materials are generally *anisotropic* and will be covered in more depth in Part V.

7.2 Learning objectives

The objectives of this section are to understand

- **Linear polarisation**
  - *Linearly x-polarised*
  - *Linearly y-polarised*
- **Linear polarisers**
  - *Dichroism*
  - *Polaroid sheets*
- **Retardation**

- **Circular polarisation**
  - $\Gamma = \pi/2$ *right circularly polarised*
  - $\Gamma = -\pi/2$ *left circularly polarised*
- **Elliptical polarisation**
- **Jones matrix**
  - *Linear polariser*
  - *Rotation of a state of polarisation by an angle* θ
- **Wave plates**
  - *Birefringence*
  - *Half-wave plate*
  - *Quarter-wave plate*
  - *General retardation plate - phase shift =* Γ
- **Analysis of polarised light**
- **Malus' Law**

7.3 Linear polarisation

7.3.1 The Jones vector

In Chapter 6, we saw that we may describe the polarisation of light in terms of the *electric field vector* $\mathbf{E}$. In an isotropic medium, this is perpendicular to the wavevector $\mathbf{E}$ and may be written, quite generally as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{i(\omega t - \mathbf{k}\cdot\mathbf{r})} \tag{7.1}$$

where $\mathbf{E}_0$ is a *Jones vector* containing the details of the polarisation. A principle advantage of using the Jones vector notation is that the exponential factor multiplying it (sometimes referred to as the *propagator*) often cancels from expressions, leaving the description of the light entirely in terms of the polarisation.

For the time being, we shall assume that we have *linearly polarised light*, although we shall see that other polarisation states are possible. Taking the wave direction to be along the z-axis for definiteness, we have, in the present case

$$\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{7.2}$$

where θ is the angle of $\mathbf{E}_0$ relative to the x-axis. Note that we did not need to specify the direction of $\mathbf{E}$ in the z-direction since this is zero by definition.

It should also be noted that the column vector (which we may refer to in short-hand as $\hat{\mathbf{p}}$) is *normalised*. That is, the dot product of $\hat{\mathbf{p}}$ with itself is unity. In terms of matrix multiplication, this can be represented by multplying the vector by its transpose. More generally, since the components of $\hat{\mathbf{p}}$ may be complex, we may write this in terms of the *complex conjugate* (see Appendix A.2). Using the dagger notation, we have

$$\hat{\mathbf{p}}^\dagger \hat{\mathbf{p}} = \begin{bmatrix} p_x^* & p_y^* \end{bmatrix} \begin{bmatrix} p_x \\ p_y \end{bmatrix} = p_x^2 + p_y^2 = 1. \tag{7.3}$$

We may quite trivially identify two special cases of Eq. (7.2)

(i) $\theta = 0$, so $\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. This is *linearly x-polarised light*.

(ii) $\theta = \pi/2$, so $\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. This is *linearly y-polarised light*.

7.3.2 Linear polarisers

Linear polarisation may be obtained in a variety of different ways. Here, we shall consider just one: the *dichroic sheet*.

Dichroism

Dichroism, a particular case of *pleochroism*, refers the phenomenon of the selective absorption of light depending on the direction of polarisation. As such, it is more properly the subject matter of *anisotropic* materials and, indeed, we shall cover the topic in greater depth in Part V on *Crystal Optics*. For the time being, however, it will help to have a real-world example of polarisation to add flesh to the mathematical description of the phenomenon.

Consider a general electromagnetic wave approaching a frame of parallel wires, as in Fig. 7.1. Since the wires are conducting, they cannot support an electric field and the free charges are redistributed until the electric field *parallel* to them is (close to) zero. The remaining electric field will therefore be *perpendicular* to the wires and the transmitted EM wave will be *linearly polarised*. The polarisation direction along which the polariser passes light is called the *transmission axis* (TA).

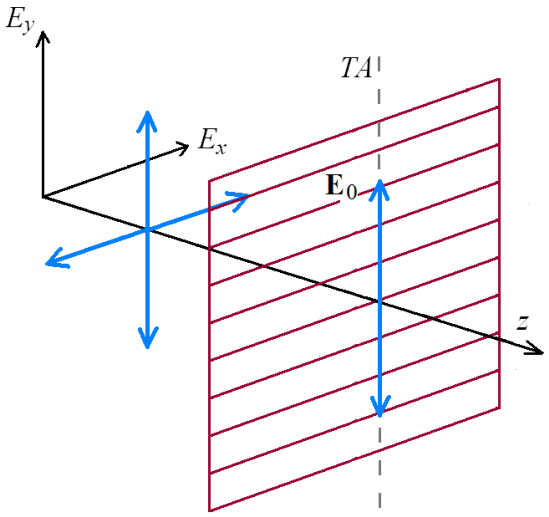


Figure 7.1: A linear polariser based on dichroism. The EM radiation approaches a grid of horizontal conductors which absorb the field in that direction. Linearly polarised light is then transmitted perpendicularly to this (along the *transmission axis* (TA) of the polariser).

Close to total attenuation of the EM field might be achieved via some cross-hatched network of wires, provided the spacings between them were on a commensurate scale to the wavelength of light. Such meshing may be seen threading the windows of microwave ovens, for instance.

Materials for optical polarisers

Such an effect may also be achieved for optical wavelengths. One possible candidate is *tourmaline*, a crystal boron silicate mineral containing various impurities. This mineral belongs to the *trigonal crystal system* (see Chapter 11). A major limitation of tourmaline as an effective polariser is that it is highly colour dependent.

A better choice is *herapathite* or *iodoquinine sulfate*. The story goes that this material was discovered by accident when a student added iodine to the urine of a dog that had been fed quinine (!) This lead to the formation of green crystals that turned out to polarise light. However, it turns out that it is very difficult to grow large crystals of herapathite.

Polaroid sheets

Progress was initially made by embedding small crystals of herapathite into a polymer. This was the first *Polaroid sheet* known as *J-sheet*. A

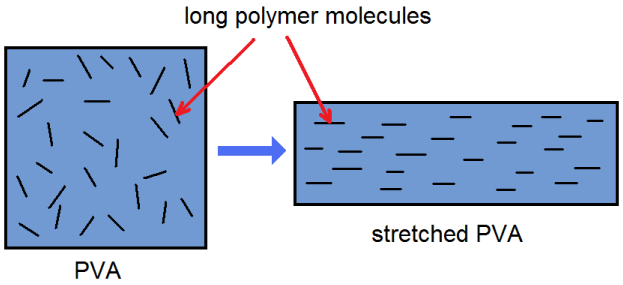


Figure 7.2: Schematic of the *H-sheet* Polaroid, showing the stretching of PVA polymer molecules into parallel, conducting chains.

second improvement was the *H-sheet* in which long polyvinyl alcohol polymer molecules are stretched during manufacture to obtain parallel chains (see Fig. 7.2). The sheet is impregnated with iodine, which attaches to the chains making them conducting. The sheet then strongly absorbs light polarised in this direction. Figure 7.4 shows a sketch of a linear polariser with its transmission axis at an arbitrary angle θ to x-axis.

One of the best known applications for linear polarisers is in the lenses of sunglasses, which cut out one of the orthogonal components of the light (Fig. 7.3).

Figure 7.3: The well-known application of linear polarisers as sunglasses.

We may summarise here the cases for linearly polarised light

- *Linearly polarised (at an angle θ to x-axis)*

$$\boxed{\mathbf{E}_0 = E_0 \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix},} \tag{7.4}$$

- *Linearly x-polarised*

$$\boxed{\mathbf{E}_0 = E_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix},} \tag{7.5}$$

- *Linearly y-polarised*

$$\boxed{\mathbf{E}_0 = E_0 \begin{bmatrix} 0 \\ 1 \end{bmatrix}}. \tag{7.6}$$

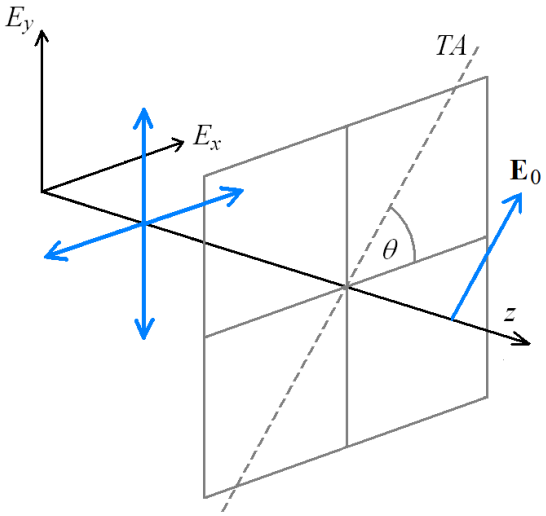


Figure 7.4: Sketch of a linear polariser with its transmission axis (TA) at an angle θ to the E_x-axis. Initially unpolarised light approaches the polariser, which then only passes light polarised along the direction of its transmission axis.

7.4 Jones matrices

The Jones calculus may be used to formally model the optical elements that prepare or change a state of polarisation. Since these operators must act on a column vector, such optical elements are represented by *matrices*.

7.4.1 Linear polariser

Firstly, let us consider an element that passes only light polarised in the E_x-direction. Formally, we are looking for an operator that satisfies the mapping

$$\begin{bmatrix} E_{0x} \\ E_{0y} \end{bmatrix} \rightarrow \begin{bmatrix} E_{0x} \\ 0 \end{bmatrix}. \tag{7.7}$$

This mapping will then involve multiplying the left-hand term by a matrix representing the optical element. For simple elements, it is often straightforward to find the elements of the matrix by inspection. Thus, we see that Eq. (7.7) is satisfied by

$$\begin{bmatrix} E_{0x} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E_{0x} \\ E_{0y} \end{bmatrix}. \tag{7.8}$$

Denoting this matrix by $\mathbf{P}_x$, the operation of an x-linear polariser is then encapsulated by

$$\boxed{\mathbf{P}_x = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.} \tag{7.9}$$

Similarly, for a y-linear polariser, we would have

$$\begin{bmatrix} 0 \\ E_{0y} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} E_{0x} \\ E_{0y} \end{bmatrix}. \tag{7.10}$$

Denoting the y-linear polariser matrix by $\mathbf{P}_y$,

$$\boxed{\mathbf{P}_y = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.} \tag{7.11}$$

In the more general case, we consider a linear polariser with its transmission axis at an angle θ to E_x-axis. We may specify this via a unit vector parallel to the transmission axis,

$$\hat{\mathbf{p}} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \tag{7.12}$$

The amplitude of the transmitted light is then

$$|\mathbf{E}_1| = \mathbf{E}_0 \cdot \hat{\mathbf{p}} = E_{0x}\cos\theta + E_{0y}\sin\theta \tag{7.13}$$

and

$$\mathbf{E}_1 = (\mathbf{E}_0 \cdot \hat{\mathbf{p}})\,\hat{\mathbf{p}}. \tag{7.14}$$

Thus, we have the mapping

$$\begin{bmatrix} E_{0x} \\ E_{0y} \end{bmatrix} \rightarrow \begin{bmatrix} E_{0x}\cos^2\theta + E_{0y}\cos\theta\sin\theta \\ E_{0x}\cos\theta\sin\theta + E_{0y}\sin^2\theta \end{bmatrix}. \tag{7.15}$$

Denoting this general polarisation matrix by $\mathbf{P}_\theta$, we see by inspection that

$$\boxed{\mathbf{P}_\theta = \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix}.} \tag{7.16}$$

7.4.2 Rotator

Consider an optical element that rotates an state of linear polarisation by an angle θ. Let us assume that the light is initially x-linearly polarised. We therefore require the mapping

$$\begin{bmatrix} E_{0x} \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} E_{0x}\cos\theta \\ E_{0x}\sin\theta \end{bmatrix}. \tag{7.17}$$

This would be accomplished via the matrix

$$\mathbf{R}_{\theta,x} = \begin{bmatrix} \cos\theta & 0 \\ \sin\theta & 0 \end{bmatrix}. \tag{7.18}$$

However, in the general case, we would also require that y-linearly polarised be mapped according to

$$\begin{bmatrix} 0 \\ E_{0y} \end{bmatrix} \rightarrow \begin{bmatrix} -E_{0y}\sin\theta \\ E_{0y}\cos\theta \end{bmatrix}, \tag{7.19}$$

which we could achieve by operating with

$$\mathbf{R}_{\theta,y} = \begin{bmatrix} 0 & -\sin\theta \\ 0 & \cos\theta \end{bmatrix}. \tag{7.20}$$

We may obtain both operations simultaneously by adding these two matrices, giving

$$\mathbf{R}_{\theta} = \mathbf{R}_{\theta,x} + \mathbf{R}_{\theta,y} \tag{7.21}$$

or

$$\boxed{\mathbf{R}_{\theta} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.} \tag{7.22}$$

Hence, this is the required *rotation matrix* (c.f. Appendix A.2). Note that the inverse operation would be a rotation by $-\theta$.

7.4.3 Combining matrices

Note that although $\mathbf{R}_{\theta}$ rotates the electric field vector $\mathbf{E}$ by θ within a given coordinate system, it may also be interpreted as rotating the coordinate system by $-\theta$ (without physically changing $\mathbf{E}$) and vice versa for $\mathbf{R}_{-\theta}$. Let us take the latter assumption and apply it to the case of a linear polariser with its transmission axis at an angle θ to the E_x axis.

We first transform to the coordinate system E'_x-E'_y which is rotated from E_x-E_y by θ. This is accomplished via the rotator $\mathbf{R}_{-\theta}$.

$$\mathbf{E}'_0 = \mathbf{R}_{-\theta}\mathbf{E}_0. \tag{7.23}$$

In this new coordinate system, the transmission axis of the polariser is aligned to the E'_x-axis. Thus, in the rotated frame, the polariser is now an linear x-polariser. To obtain the transmission through it, we therefore need to apply $\mathbf{P}_x$

$$\mathbf{E}'_1 = \mathbf{P}_x\mathbf{E}'_0 = \mathbf{P}_x\mathbf{R}_{-\theta}\mathbf{E}_0. \tag{7.24}$$

Finally, we need to transform back to the original coordinate system. We do this with $\mathbf{R}_\theta$

$$\mathbf{E}_1 = \mathbf{R}_\theta\mathbf{E}'_1 = \mathbf{R}_\theta\mathbf{P}_x\mathbf{R}_{-\theta}\mathbf{E}_0. \tag{7.25}$$

Since matrix multiplication is associative, we can encapsulate the entire process in a resultant matrix obtained by multiplying all three matrix operators together

$$\mathbf{M} = \mathbf{R}_\theta\mathbf{P}_x\mathbf{R}_{-\theta}. \tag{7.26}$$

Substituting from Eqs. (7.103) and (7.9), we have

$$\begin{aligned}
\mathbf{M} &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \\
&= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \cos\theta & \sin\theta \\ 0 & 0 \end{bmatrix}, \\
&= \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix} = \mathbf{P}_\theta,
\end{aligned} \tag{7.27}$$

as found earlier.

We see here an example of the way in which operations performed one after another may be encapsulated by multiplying the matrices for the individual processes together. However, it is important that this multiplication is carried out in order (later operations acting on the left), since matrix multiplication is not commutative.

7.5 Elliptically polarised light

7.5.1 The retardation

The most general form of polarisation is *elliptical polarisation*, in which the electric field spirals around the propagation axis tracing out an ellipse. This may be understood by resolving the electric field into orthogonal components. So long as these components remain *in phase*, the polarisation will

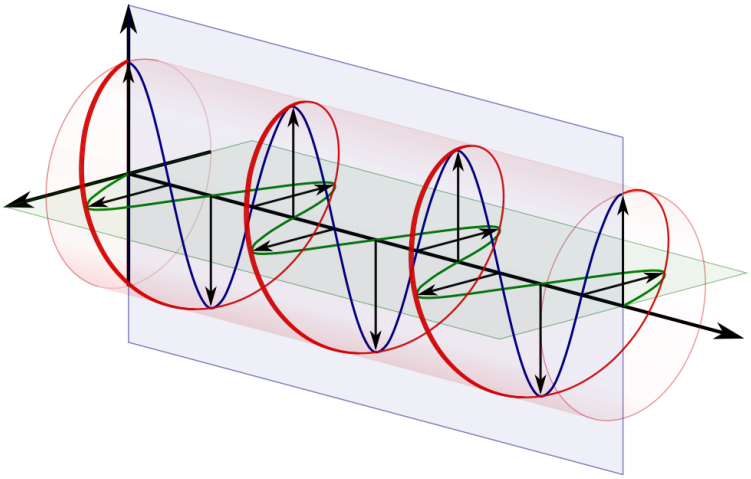


Figure 7.5: Illustration of elliptically polarised light in terms of orthogonal components with a phase shift Γ between them, giving the resultant waveform (helical line).

be linear. If, however, a phase shift is introduced on to one of the components, the polarisation will become *elliptical*, as in Fig. 7.5. This phase shift is known as the *retardation* Γ.

Writing the polarisation out in terms of Jones vectors

$$\mathbf{E} = \left(|\mathbf{E}_{0x}| \begin{bmatrix} e^{i\phi_x} \\ 0 \end{bmatrix} + |\mathbf{E}_{0y}| \begin{bmatrix} 0 \\ e^{i\phi_y} \end{bmatrix} \right) e^{i(\omega t - kz)}, \tag{7.28}$$

where the retardation is given by

$$\phi_y - \phi_x = \Gamma. \tag{7.29}$$

Multiplying Eq. (7.28) through by $e^{-i\phi_x}$ then gives

$$\mathbf{E}e^{-i\phi_x} = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ e^{i\Gamma}\sin\theta \end{bmatrix} e^{i(\omega t - kz)}, \tag{7.30}$$

where, again, θ is the angle between the electric field and the x-axis. Absorbing the common phase shift into the electric field vector, the *Jones vector for elliptical polarisation* is

$$\boxed{\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ e^{i\Gamma}\sin\theta \end{bmatrix}.} \tag{7.31}$$

We may pause to identify some special cases of Eq. 7.31 that reduce back down to linear polarisation.

(i) $\Gamma = 0$. This just yields a common phase shift on both components and the wave remains linearly polarised. Hence

$$\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \tag{7.32}$$

(ii) $\Gamma = \pm\pi$. In this case, we have

$$e^{i\Gamma} = -1, \tag{7.33}$$

so

$$\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix}. \tag{7.34}$$

Thus $\mathbf{E}$ remains linearly polarised but the direction of polarisation is rotated by an angle 2θ through the x-axis. We shall see later that this change in polarisation may be achieved by an optical element known as a *half-wave plate*.

7.5.2 Circular polarisation

Let us now consider the cases where $\Gamma = \pm\pi/2$ and impose the condition $\theta = \pi/4$, so that $\cos\theta = \sin\theta = 2^{-1/2}$.

(iii) $\Gamma = \pi/2$. This gives

$$e^{i\pi/2} = i, \tag{7.35}$$

so

$$\mathbf{E}_0 = \frac{|\mathbf{E}_0|}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}. \tag{7.36}$$

Asserting the propagator, the electric field is given by

$$\mathbf{E} = \frac{|\mathbf{E}_0|}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix} e^{i(\omega t - kz)}. \tag{7.37}$$

Considering the real part of $\mathbf{E}$

$$\mathrm{Re}\,[\mathbf{E}] = \frac{|\mathbf{E}_0|}{\sqrt{2}} \begin{bmatrix} \cos(\omega t - kz) \\ -\sin(\omega t - kz) \end{bmatrix} \tag{7.38}$$

Thus, in space the electric field vector, $\mathbf{E}$, rotates around the propagation axis (z-axis) in the same direction as a *right-handed screw*. Hence, this is known as *right circular polarisation*. Note that looking along the approaching wave towards the origin, an observer would see $\mathbf{E}$ at a fixed point in space rotating *clockwise* in time.

(iv) $\Gamma = -\pi/2$. In this case we have

$$e^{i\pi/2} = -i, \tag{7.39}$$

so

$$\mathbf{E}_0 = \frac{|\mathbf{E}_0|}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}. \tag{7.40}$$

This reverses the sign of y-component from the previous case. Hence, the electric field rotates around the propagation axis in the opposite direction to right circular polarisation. Moreover, at a given point in space, $\mathbf{E}$ rotates *anti-clockwise* in time. This is therefore known as *left circular polarisation*.

Summarising the results for circularly polarised light,

- *right circular polarised light*

$$\boxed{\mathbf{E}_+ = \frac{E_0}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix},} \tag{7.41}$$

- *left circular polarised light*

$$\boxed{\mathbf{E}_- = \frac{E_0}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}.} \tag{7.42}$$

7.5.3 Case: $\Gamma = \pm\pi/2$, $E_{0x} \neq E_{0y}$

The results of the previous section held for the case where imposed the condition that amplitudes of the x and y components of the electric field were equal. If we now relax this condition, the resultant polarisation is no longer circular. Instead, for phase shifts of $\Gamma = \pm\pi/2$, we will have

$$\mathbf{E} = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ \pm i \sin\theta \end{bmatrix} e^{i\phi}. \tag{7.43}$$

where $\phi = \omega t - kz$. Considering just the real part, we then have

$$\mathrm{Re}\,[\mathbf{E}] = |\mathbf{E}_0| \begin{bmatrix} \cos\theta\cos\phi \\ \mp i\sin\theta\sin\phi \end{bmatrix}, \tag{7.44}$$

which just traces out an ellipse in either a clockwise ($\Gamma = \pi/2$) or anti-clockwise ($\Gamma = -\pi/2$) direction. Putting

$$\begin{aligned} E_{0x} &= |\mathbf{E}_0|\cos\theta, \\ E_x &= \cos\phi \end{aligned}$$

and

$$\begin{aligned} E_{0y} &= |\mathbf{E}_0|\sin\theta, \\ E_y &= \sin\phi, \end{aligned}$$

we find

$$\left(\frac{E_y}{E_{0y}}\right)^2 + \left(\frac{E_x}{E_{0x}}\right)^2 = 1. \tag{7.45}$$

This is an ellipse in standard form, so its principle axes are aligned to the x and y coordinate axes. Hence the maximum and minimum values of E_x are $\pm E_{0x}$ and similarly for E_y.

7.5.4 Case: Γ is arbitrary, $E_{0x} \neq E_{0y}$

In the most general case, the phase shift Γ is arbitrary and there is no restriction on the relative amplitudes of E_{0x} and E_{0y}. The electric field vector may now be written as

$$\mathbf{E} = \begin{bmatrix} E_{0x} \\ e^{i\Gamma}E_{0y} \end{bmatrix} e^{i\phi}, \tag{7.46}$$

where $\phi = \omega t - kz$ as before. The real part of $\mathbf{E}$ is

$$\mathbf{E} = \begin{bmatrix} E_{0x}\cos\phi \\ E_{0y}\cos(\phi+\Gamma) \end{bmatrix}. \tag{7.47}$$

Firstly, let us consider the rotation of the field vector around the propagation direction. From Eq. (7.47), we see that the real part of $\mathbf{E}$ makes an angle

$$\theta = \tan^{-1}\left[\frac{E_{0y}\cos(\phi+\Gamma)}{E_{0x}\cos\phi}\right] \tag{7.48}$$

to the x-axis. Taking the derivative with respect to time, we have

$$\frac{\partial\theta}{\partial t} = -\frac{\omega E_{0x} E_{0y} \sin\Gamma}{E_{0x}^2 \cos^2\phi + E_{0y}^2 \cos^2(\phi+\Gamma)}. \tag{7.49}$$

In any of the cases $E_{0x} = 0$, $E_{0y} = 0$, $\Gamma = 0$ or $\Gamma = \pm\pi$, we have linear polarisation and, as expected, we have $\partial\theta/\partial t = 0$. In other words, the orientation of the electric field vector $\mathbf{E}$ stays fixed.

In other cases, the direction of rotation is given by the sign of $\partial\theta/\partial t$, i.e. anti-clockwise for $\partial\theta/\partial t > 0$ and clockwise for $\partial\theta/\partial t < 0$. Inspecting Eq. (7.49), we see that this is determined soley by the sign of $\sin\Gamma$. Assuming that ω, E_{0x} and E_{0y} are all positive, we therefore have

- $0 < \Gamma < \pi$, so $\sin\Gamma > 0$ and $\partial\theta/\partial t < 0$.

  Rotation is therefore **clockwise**.

- $-\pi < \Gamma < 0$, so $\sin\Gamma < 0$ and $\partial\theta/\partial t > 0$.

  Rotation is therefore **anti-clockwise**.

Note that the direction of rotation changes as the electric field vector $\mathbf{E}$ passes through linear states of polarisation for $\Gamma = 0$ and $\Gamma = \pi$.

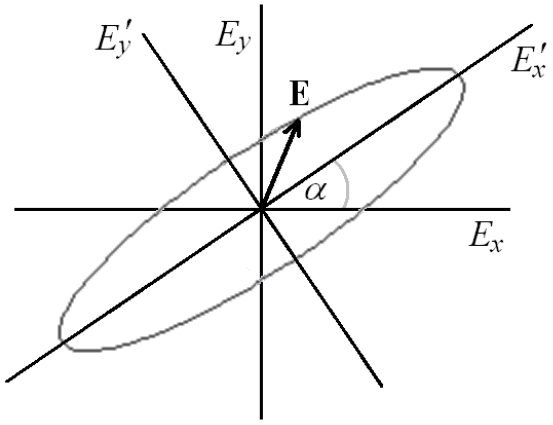


Figure 7.6: The polarisation ellipse for a general state of polarisation.

Let us now define

$$E_x = E_{0x}\cos\phi \tag{7.50}$$

and

$$E_y = E_{0y}\cos(\phi+\Gamma). \tag{7.51}$$

This gives

$$\frac{E_x}{E_{0x}} = \cos\phi \quad \text{and} \quad \frac{E_y}{E_{0y}} = \cos\phi\cos\Gamma - \sin\phi\sin\Gamma, \tag{7.52}$$

so

$$\frac{E_y}{E_{0y}} - \frac{E_x}{E_{0x}}\cos\Gamma = -\sin\phi\sin\Gamma. \tag{7.53}$$

Squaring this gives

$$\begin{aligned}\left(\frac{E_y}{E_{0y}}\right)^2 - 2\left(\frac{E_y}{E_{0y}}\right)\left(\frac{E_x}{E_{0x}}\right)\cos\Gamma \quad &+ \\ + \left(\frac{E_x}{E_{0x}}\right)^2\cos^2\Gamma \quad &= \quad \sin^2\phi\sin^2\Gamma \\ &= \quad \left(1-\cos^2\phi\right)\sin^2\Gamma. \end{aligned} \tag{7.54}$$

But from Eq. (7.53), we have $\cos^2\phi = E_x^2/E_{0x}^2$, so inserting this into Eq. (7.54) and re-arranging, we have

$$\boxed{\left(\frac{E_y}{E_{0y}}\right)^2 + \left(\frac{E_x}{E_{0x}}\right)^2 - 2\left(\frac{E_y}{E_{0y}}\right)\left(\frac{E_x}{E_{0x}}\right)\cos\Gamma = \sin^2\Gamma.} \tag{7.55}$$

This is the general equation of an ellipse.

Equation (7.55) is illustrated in Fig. 7.6. Note that in the E'_x-E'_y coordinate system (rotated from E_x-E_y by an angle α about the z-axis), the ellipse is in standard form. The coordinate systems are therefore related by the transformation

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}\begin{bmatrix} E'_x \\ E'_y \end{bmatrix}, \tag{7.56}$$

where the matrix encapsulates a rotation by α about the origin. Expressing the products of E_x and E_y in terms of the primed coordinates, we have

$$E_x^2 = E_x'^2\cos^2\alpha + E_y'^2\sin^2\alpha - 2E'_xE'_y\cos\alpha\sin\alpha, \tag{7.57}$$

$$E_y^2 = E_x'^2\sin^2\alpha + E_y'^2\cos^2\alpha + 2E'_xE'_y\cos\alpha\sin\alpha, \tag{7.58}$$

and

$$E_xE_y = \left(E_x'^2 - E_y'^2\right)\cos\alpha\sin\alpha + E'_xE'_y\left(\cos^2\alpha - \sin^2\alpha\right). \tag{7.59}$$

Since the ellipse is in standard form in the E'_x-E'_y coordinate system, the cross-terms $E'_x E'_y$ must disappear. From Eq. (7.55), we therefore require

$$\sin 2\alpha \left(\frac{1}{E_{0y}^2} - \frac{1}{E_{0x}^2} \right) - 2 \frac{\cos 2\alpha}{E_{0x} E_{0y}} \cos \Gamma = 0, \tag{7.60}$$

where we have used the identities

$$2 \cos \alpha \sin \alpha = \sin 2\alpha \tag{7.61}$$

and

$$\cos^2 \alpha - \sin^2 \alpha = \cos 2\alpha. \tag{7.62}$$

Hence, we have

$$\boxed{\tan 2\alpha = \frac{2E_{0x}E_{0y}}{E_{0x}^2 - E_{0y}^2} \cos \Gamma.} \tag{7.63}$$

Now, Eq. (7.63) may be re-written

$$\tan 2\alpha = \frac{2 \left(E_{0y}/E_{0x} \right)}{1 - \left(E_{0y}/E_{0x} \right)^2} \cos \Gamma. \tag{7.64}$$

Defining $\tan \beta = E_{0y}/E_{0x}$ and using the double angle identity for $\tan$, this may be expressed as

$$\tan 2\alpha = \frac{2 \tan \beta}{1 - \tan^2 \beta} \cos \Gamma = \tan 2\beta \cos \Gamma. \tag{7.65}$$

Note from Eq. (7.63) that if $E_{0x} = E_{0y}$, $\tan 2\alpha \to \infty$, so $\alpha = \pi/4$.

7.5.5 Limiting cases

We may now apply the special conditions considered in the previous sub-sections to our general formulation for the polarisation.

(i) $\Gamma = 0$. In this case $\cos \Gamma = 1$, $\sin \Gamma = 0$ and Eq. (7.55) reduces to

$$\left(\frac{E_y}{E_{0y}} - \frac{E_x}{E_{0x}} \right)^2 = 0, \tag{7.66}$$

yielding

$$E_y = \frac{E_{0y}}{E_{0x}} E_x. \tag{7.67}$$

This is the equation of a straight line with gradient equal to the ratio of the amplitudes of the y and x components of the field. We therefore have linear polarisation as expected. Note that this implies that it is the *phase shift that gives rise to the elliptical shape* that $\mathbf{E}$ traces out.

(ii) $\Gamma = \pm\pi$. In this case $\cos\Gamma = -1$, $\sin\Gamma = 0$. Thus Eq. (7.55) gives us

$$E_y = -\frac{E_{0y}}{E_{0x}} E_x. \tag{7.68}$$

Again, we have linear polarisation with the negative of the gradient for the $\Gamma = 0$ case.

(iii) $\Gamma = \pm\pi/2$. Now $\cos\Gamma = 0$ and $\sin\Gamma = \pm 1$. Equation. (7.55) therefore reduces to Eq. (7.45), the equation of an ellipse in standard form.

$$\left(\frac{E_y}{E_{0y}}\right)^2 + \left(\frac{E_x}{E_{0x}}\right)^2 = 1. \tag{7.69}$$

When $E_{0x} = E_{0y} = E_0$, this further simplifies to

$$E_y^2 + E_x^2 = E_0^2, \tag{7.70}$$

which is the equation of a circle. Thus the polarisation is circular.

7.6 Wave plates

7.6.1 Birefringence

A *retardation* or *wave plate* is an optical element that produces some retardation between the orthogonal components of the wave. The physical origin of this retardation is due to the phenomenon of *birefringence*, in which the components of the wave see a different refractive index (and hence have a different phase velocity) depending on the orientation of the polarisation within some *anisotropic* material. The subject of anisotropy is dealt with in detail in Part V on *Crystal Optics*. Meanwhile, we shall describe a simplified picture for the sake of insight.

Just as a linear polariser has a particular transmission axis, so we can define two transmission axes perpendicular to the propagation direction for a wave plate. These are known as the *fast* and *slow* axes, corresponding to the wave speeds of the transmitted orthogonal components. Thus, the phase of the wave along the fast axis is ahead of that along the slow axis.

The speed of the waves is, of course, determined by the refractive index that it sees. This is determined by a construction known as the *index*

ellipsoid. This concept will be fully developed in Chapter 12. For now, we shall just consider an ellipsoid defined by the equation

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1. \tag{7.71}$$

where n_o and n_e are called the *ordinary* and *extraordinary* refractive indices respectively. The fact that there are only two special refractive indexes reflects the fact that the crystal type Eq. (7.71) describes has only one *optical axis*. Hence, such a material is described as being *uniaxial*.

Consider that case of two orthogonally polarised plane waves, $\mathbf{E}_e$ and $\mathbf{E}_o$ propagating with the same wavevector $\mathbf{k}$ in this crystal, as shown in Fig. 7.7. The plane perpendicular to $\mathbf{k}$ intersects the index ellipsoid in an ellipse. One of the axes of this ellipse is always equal to the *ordinary* refractive index n_o, the other is dependent on the *extraordinary* refractive index n_e and the angle of $\mathbf{k}$ to the extraordinary axis. Thus, one of the plane waves will travel with a wave speed $v_o = c/n_o$, the other with wave speed $v_o(\theta) = c/n(\theta)$. These are known as the *ordinary* and *extraordinary* waves respectively.

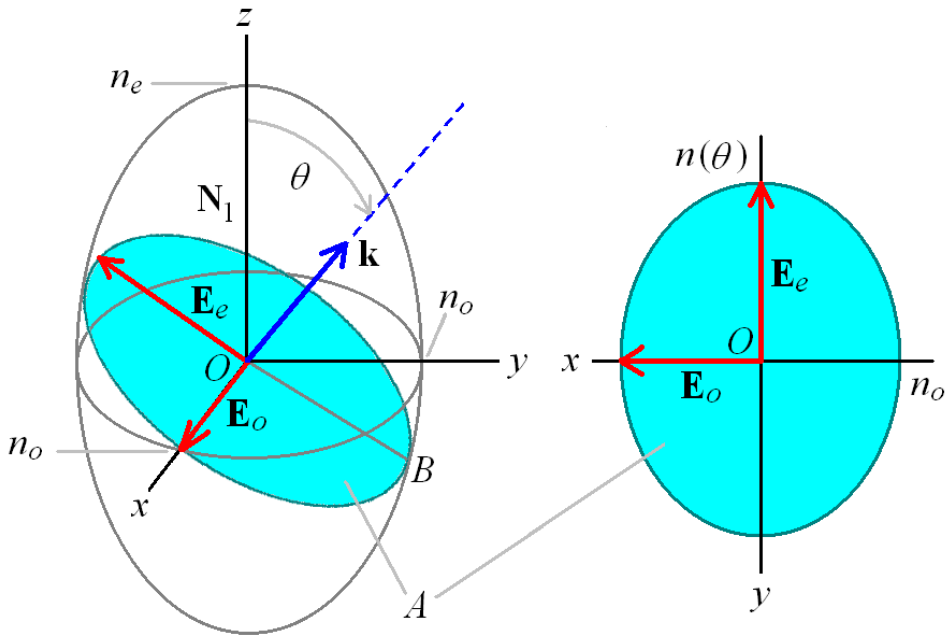


Figure 7.7: The index ellipsoid for a uniaxial crystal projected in 2D showing two orthogonally polarised plane waves, $\mathbf{E}_e$ and $\mathbf{E}_o$ propagating with the same wavevector $\mathbf{k}$. The plane perpendicular to $\mathbf{k}$ intersects the index ellipsoid in an ellipse. One of the axes of this ellipse is always equal to the *ordinary* refractive index n_o, the other is dependent on the *extraordinary* refractive index n_e and the angle of $\mathbf{k}$ to the extraordinary axis.

Uniaxial crystals may be further defined according to the relative sizes of n_e and n_o. A *negative uniaxial crystal* has $n_e < n_o$. Examples include calcite ($\mathrm{CaCO_3}$) and ruby ($\mathrm{Al_2O_3}$). For $n_e > n_o$, we have a *positive uniaxial crystal*, for example, quartz ($\mathrm{SiO_2}$).

In a retardation plate, in a negative uniaxial crystal, the extraordinary axis is aligned with the *fast axis* of the plate, since $c/n_e > c/n_o$. For a positive uniaxial crystal, we have the opposite case and the extraordinary axis is therefore aligned with the *slow* axis.

The *birefringence* is defined as

$$\Delta n(\theta) = n(\theta) - n_0. \tag{7.72}$$

For a wave with extraordinary and ordinary components, E_e and E_o respectively, propagating in a direction r, we may write

$$E_e = E_0 \exp\left(i\omega\left[t - \frac{n(\theta)r}{c}\right]\right) \tag{7.73}$$

and

$$E_o = E_0 \exp\left(i\omega\left[t - \frac{n_o r}{c}\right]\right) \tag{7.74}$$

where we have used $k = \omega/v = n\omega/c$. The second of these equations may be re-written

$$E_o = E_0 \exp\left(i\omega\left[t - \frac{n(\theta)r}{c}\right]\right) \exp\left(i\frac{\omega}{c}\left[n(\theta) - n_o\right]r\right). \tag{7.75}$$

Hence, after a distance r, the ordinary wave acquires a retardation

$$\Gamma(r) = \frac{\omega\Delta n(\theta)}{c} r. \tag{7.76}$$

7.6.2 Half-wave plate

A half-wave plate introduces a retardation of $\Gamma = \pm\pi$. Suppose, then, that the initial polarisation of a propagating EM wave is given according to Eq. (7.2) by

$$\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \tag{7.77}$$

After transmission through a half-wave plate the new polarisation will be

$$\mathbf{E}_1 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix}. \tag{7.78}$$

We can express this in terms of the action of a matrix $\mathbf{M}_\pi$ acting on $\mathbf{E}_0$. Clearly, this matrix is given by

$$\boxed{\mathbf{M}_\pi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.} \tag{7.79}$$

Thus, this has the effect of reversing the direction of y-component $\sin\theta$ (where θ is the initial angle of the polarisation to the x-axis). In other words, the half-wave plate has the effect of rotating a state of linear polarisation by -2θ through the x-axis. Note that this is equivalent to a rotation of $2\theta'$ through the y-axis, where $-\theta'$ is the angle of the electric field vector to the y-axis. Thus, whilst the fast axis may be aligned in either the x or y directions, we may say unambiguously that

- *a half-wave plate has the effect of rotating a linear state of polarisation by an angle -2θ through the fast (or slow) axis, where θ is the initial angle of the electric field vector to the fast (or slow) axis.*

7.6.3 Quarter-wave plate

For a quarter-wave plate, we have $\Gamma = \pm\pi/2$. Applying the same reasoning as before, we find that the Jones matrix for a quarter-wave plate is given by

$$\boxed{\mathbf{M}_{\pm\pi/2} = \begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix}.} \tag{7.80}$$

Applying this to the Jones vector of Eq. (7.2) we have

$$\mathbf{E}_1 = |\mathbf{E}_0| \begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ \pm i\sin\theta \end{bmatrix}. \tag{7.81}$$

Comparing this result to Eq. (7.43), we see that this yields *elliptically polarised* light. In this case, the principle axes of the ellipse are aligned with the E_x and E_y axes. In the specific case where

$$\cos\theta = \sin\theta, \tag{7.82}$$

we have $\theta = \pi/4$ and Eq. (7.81) reduces to

$$\mathbf{E}_1 = \frac{|\mathbf{E}_0|}{\sqrt{2}} \begin{bmatrix} 1 \\ \pm i \end{bmatrix}. \tag{7.83}$$

This is the Jones vector for *circularly polarised* light. Specifically, as given in the last section, for $\Gamma = \pi/2$ we have *right* circular polarisation and for $\Gamma = -\pi/2$ we have *left* circular polarisation. Thus

- *a quarter-wave plate introduces a phase shift of $\pi/2$ between the components of the optical field, producing elliptically polarised light. In the specific case where E_x and E_y are initially equal, the quarter-wave plate produces circularly polarised light.*

7.6.4 General retardation plate

A general retardation plate introduces an arbitrary phase shift $\Gamma = \pm\pi/2$. It is straightforward to see that the Jones matrix for such a plate will be given by

$$\boxed{\mathbf{M}_\Gamma = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\Gamma} \end{bmatrix}.} \tag{7.84}$$

7.6.5 3D glasses

Figure 7.8: 3D glasses based on circularly polarised light.

A recent use of wave-plates has been in the design of 3D glasses used in conjunction with *stereoscopic* filming and projection of video to give the impression of visual depth. Stereoscopic visualisation is achieved by taking two images of the same scene from the position of both eyes. These images are then brought together again in such a way that the brain interprets them as different perspectives on the same view and producing the impression of a three dimensional scene.

One way of achieving this effect is to project the two images in oppositely handed circularly polarised light. The lenses of special 3D glasses

then pass one of these projections and block the other. This is achieved by first passing the light through a quarter wave plate and then a linear polariser, the linear polarisers being rotated at 90° to each other in each lens.

The two projections are right and left circularly polarised. Passing each through a quarter wave plate gives

$$\mathbf{M}_{\pi/2}\mathbf{E}_{+} = \frac{E_0}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}\begin{bmatrix} 1 \\ i \end{bmatrix} = \frac{E_0}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{7.85}$$

and

$$\mathbf{M}_{\pi/2}\mathbf{E}_{-} = \frac{E_0}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}\begin{bmatrix} 1 \\ -i \end{bmatrix} = \frac{E_0}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{7.86}$$

Note that Eq. (7.85) yields linearly polarised light aligned along the $y = -x$ axis, whilst Eq. (7.86) gives linearly polarised light along the $y = x$ axis. Hence, if this is followed by a linear polariser with its transmission axis aligned along $y = x$, the combination will *pass* left circularly polarised light and *block* right circularly polarised light.

For the other lens, the opposite effect may be achieved either with a $\mathbf{M}_{-\pi/2}$ quarter wave plate and the same orientation linear polariser or the same wave plate and an orthogonally orientated linear polariser. Note, that due to the rotational symmetry of the circularly polarised light, tilting the glasses does not effect the analysis of the light (an advantage over just using linear polarisers).

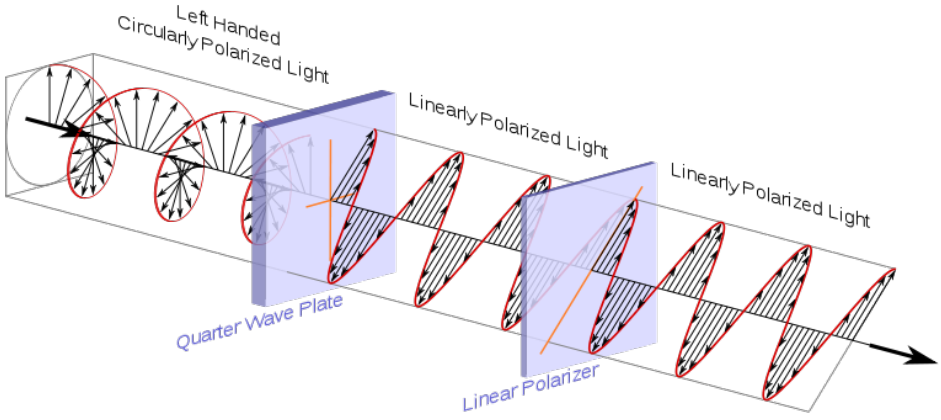


Figure 7.9: Analysis of circularly polarised light into linearly polarised light.

7.7 Analysis of polarised light

7.7.1 Linear analyser

In addition to its role preparing linear states of polarisation, a linear polariser may also be used to analyse light in a general state of polarisation into its linear components. In this role, it is often referred to as an *analyser*.

Consider light in a general state of polarisation

$$\mathbf{E} = |\mathbf{E}_0| \begin{bmatrix} \cos\theta_1 \cos\omega t \\ \sin\theta_1 \cos(\omega t + \Gamma) \end{bmatrix}. \tag{7.87}$$

This is to be passed through an analyser with its transmission axis at an angle θ_2 to the x-axis. The unit vector in this direction is then

$$\hat{\mathbf{p}} = \begin{bmatrix} \cos\theta_2 \\ \sin\theta_2 \end{bmatrix}. \tag{7.88}$$

The analyser will only pass the component of light polarised in this direction. Thus, the amplitude of the transmitted light $E_{\mathbf{p}}$ is found by taking the dot product

$$E_{\mathbf{p}} = \mathbf{E} \cdot \hat{\mathbf{p}} = |\mathbf{E}_0| \left[\cos\omega t \cos\theta_1 \cos\theta_2 + \cos(\omega t + \Gamma) \sin\theta_1 \sin\theta_2\right]. \tag{7.89}$$

Putting

$$\begin{aligned} E_{0x} &= |\mathbf{E}_0| \cos\theta_1, \\ E_{0y} &= |\mathbf{E}_0| \sin\theta_1, \end{aligned} \tag{7.90}$$

this may be rewritten

$$E_{\mathbf{p}} = E_{0x} \cos\omega t \cos\theta_2 + E_{0y} \cos(\omega t + \Gamma) \sin\theta_2. \tag{7.91}$$

The intensity of the transmitted light will be proportional to the squared modulus of this, that is

$$\begin{aligned} |E_{\mathbf{p}}|^2 &= \left[E_{0x}^2 \cos^2\omega t \cos^2\theta_2 + E_{0y}^2 \cos^2(\omega t + \Gamma) \sin^2\theta_2 + \right. \\ &+ \left. 2E_{0x}E_{0y} \cos\omega t \cos(\omega t + \Gamma) \cos\theta_2 \sin\theta_2\right]. \end{aligned}$$

This expression is still time dependent. Averaging over a period of the optical oscillation $T = 2\pi/\omega$, we have

$$\begin{aligned}\left\langle |E_{\mathbf{p}}|^2 \right\rangle &= \frac{\omega}{2\pi}\int_0^{2\pi/\omega} |E_{\mathbf{p}}|^2 \; dt \\ &= \frac{1}{2}\left[E_{0x}^2\cos^2\theta_2 + E_{0y}^2\sin^2\theta_2 + 2E_{0x}E_{0y}\cos\theta_2\sin\theta_2\cos\Gamma\right].\end{aligned}$$

The time-averaged intensity may then be given as

$$\boxed{I(\theta) = I_0\left(\cos^2\theta + r^2\sin^2\theta + 2r\cos\theta\sin\theta\cos\Gamma\right),} \tag{7.92}$$

where $r = E_{0y}/E_{0x}$.

Consider the case where the incident light is linearly x-polarised, that is $E_{0y} = 0$. In this case, Eq. (7.92) reduces to

$$\boxed{I(\theta) = I_0\cos^2\theta.} \tag{7.93}$$

This is known as *Malus' Law* for the transmission of linearly polarised light. For circularly polarised light, $r = 1$ and $\Gamma = \pm\pi/2$. In this case, we have

$$I(\theta) = I_0\left(\cos^2\theta + \sin^2\theta\right) = I_0. \tag{7.94}$$

In other words, the time-averaged intensity is constant.

7.8 Summary

- **Linear polarisation**

  Light polarised in a fixed direction all along the propagation direction is said to be *linearly polarised*. For light travelling in the z-direction, the general expression for linearly polarised light is

  $$\boxed{\mathbf{E}_0 = E_0\begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix},} \tag{7.95}$$

  – *Linearly x-polarised*

  $$\boxed{\mathbf{E}_0 = E_0\begin{bmatrix}1 \\ 0\end{bmatrix},} \tag{7.96}$$

  – *Linearly y-polarised*

  $$\boxed{\mathbf{E}_0 = E_0\begin{bmatrix}0 \\ 1\end{bmatrix}.} \tag{7.97}$$

- **Retardation**

  In anisotropic media, orthogonal components of the light may see different refractive indices. This introduces a phase shift between the components known as the *retardation* Γ.

- **Circular polarisation**

  If the y component of a linearly polarised plane wave is multiplied by a factor $e^{i\pm\pi/2}$, then it will acquire a phase shift of $\Gamma = \pm\pi/2$. When $E_{0x} = E_{0y}$, this leads to *circularly polarised light*. There are two cases to consider:

  – $\Gamma = \pi/2$ *right circularly polarised*
  In this case, the electric field vector at a particular point along the z-axis rotates in a *clockwise* direction in the E_x-E_y plane. This is known as *right circularly polarised* light.

$$\boxed{\mathbf{E}_+ = \frac{E_0}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}.} \tag{7.98}$$

  – $\Gamma = -\pi/2$ *left circularly polarised*
  In this case, the electric field vector at a particular point along the z-axis rotates in an *anti-clockwise* direction in the E_x-E_y plane. This is known as *left circularly polarised* light.

$$\boxed{\mathbf{E}_- = \frac{E_0}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}.} \tag{7.99}$$

- **Elliptical polarisation**

  In the general case, light is *elliptically polarised*, with the components of the electric field E_x and E_y satisfying

$$\boxed{\left(\frac{E_y}{E_{0y}}\right)^2 + \left(\frac{E_x}{E_{0x}}\right)^2 - 2\left(\frac{E_y}{E_{0y}}\right)\left(\frac{E_x}{E_{0x}}\right)\cos\Gamma = \sin^2\Gamma.} \tag{7.100}$$

  The rotation of the electric field vector about the propagation direction depends on the *retardation* Γ.

  – Case: $0 < \Gamma < \pi$. Rotation is **clockwise**.
  – Case: $-\pi < \Gamma < 0$. Rotation is **anti-clockwise**.

The Jones vector for the general case is

$$\boxed{\mathbf{E}_0 = |\mathbf{E}_0| \begin{bmatrix} \cos\theta \\ e^{i\Gamma}\sin\theta \end{bmatrix}.} \tag{7.101}$$

- **Jones matrix**
  A Jones matrix represents the operation of an optical element on a state of polarisation (represented by a Jones vector). Examples covered are

  – *Linear polariser (with the transmission axis at an angle θ to the E_x axis)*

$$\boxed{\mathbf{P}_\theta = \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix}.} \tag{7.102}$$

  – *Rotation of a state of polarisation by an angle θ*

$$\boxed{\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.} \tag{7.103}$$

  – *Half-wave plate*

$$\boxed{\mathbf{M}_\pi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.} \tag{7.104}$$

  – *Quarter-wave plate*

$$\boxed{\mathbf{M}_{\pm\pi/2} = \begin{bmatrix} 1 & 0 \\ 0 & \pm i \end{bmatrix}.} \tag{7.105}$$

  – *General retardation plate - phase shift = Γ*

$$\boxed{\mathbf{M}_\Gamma = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\Gamma} \end{bmatrix}.} \tag{7.106}$$

- **Analysis of polarised light**

  A linear polariser is an optical element that passes only linearly polarised light.

  Such an optical element may also be used to analyse light into its linearly polarised components. It may then be refered to as an *analyser*.

For the general case of arbitrarily polarised light passing through an analyser with its transmission at an angle θ to the E_x axis, the time-averaged intensity of the transmitted light is given by

$$\boxed{I(\theta) = I_0\left(\cos^2\theta + r^2\sin^2\theta + 2r\cos\theta\sin\theta\cos\Gamma\right),} \tag{7.107}$$

where $r = E_{0y}/E_{0x}$.

- **Malus' Law**

  For linearly x-polarised light, the transmitted intensity through the analyser is given by

$$\boxed{I(\theta) = I_0\cos^2\theta.} \tag{7.108}$$

  This is known as **Malus' Law** for the transmission of linearly polarised light.

8. The Fresnel Equations

8.1 General remarks

A crucial application of optics involves the reflection and transmission of light at the boundary between media of different refractive indices. Although, strictly, this violates our notion of *homogeneity*, we may still think of the media involved being *locally* homogeneous.

The quantitative model of reflection and transmission arises out of Maxwell's equations and are generally referred to as the *Fresnel equations*. The starting point for this analysis is to understand the boundary conditions of the different electromagnetic fields $\mathbf{E}$, $\mathbf{H}$, $\mathbf{D}$ and $\mathbf{B}$, which we revise in the first section of this Chapter. Thereafter, however, we shy away from a full derivation of the Fresnel equations, referring the reader to standard texts on electromagnetism, and simply quote the results.

In the later sections, we extend our treatment to consider *irradiance*, i.e. the incident, reflected and transmitted *intensities*. Along the way, we shall also consider reflection and transmission from a wavevector picture, obtaining alternative derivations of the Laws of Reflection and Refraction, as well as predicting the existence of an *evanescent wave* in the case of total internal reflection.

The results obtained here contain plenty of interesting physics. Some of the possible applications are mentioned briefly along the way. Other applications that we shall cover later include

- **Thin-film interference**
- **Anti-reflection coatings**
- **Resonant cavities**

8.2 Learning objectives

- **Boundary conditions**

  The boundary conditions of the electromagnetic fields at interfaces between media of different refractive indices.

- **Reflection and refraction**

  Reflection and refraction via wavevector.

- **Fresnel equations**

  The Fresnel equations for the reflection and transmission coefficients r and t.

  – *Brewster angle*

- **Time reversibility**

  Use of the principle of time reversibility for analysis of reflection and transmission coefficients.

- **Stokes treatment**

  Alternative derivation of the results of time reversibility via the Stoke's treatment.

- **Irradiance**

  Analysis of power reflection and transmission between different media.

- **Total internal reflection**

  Wavevector analysis of total internal reflection to obtain the *evanescent wave*.

  – *Optical couplers*

8.3 Boundary conditions

We shall consider the boundary conditions that must prevail when an electromagnetic wave crosses the boundary between media of different refractive index. To facilitate this, we shall make use of *Stoke's theorem*

$$\int_A (\nabla \times \mathbf{F}) \cdot d\mathbf{A} = \oint_C \mathbf{F} \cdot d\mathbf{l}, \tag{8.1}$$

where the line integral on the right-hand-side is around the curve C enclosing the area A.

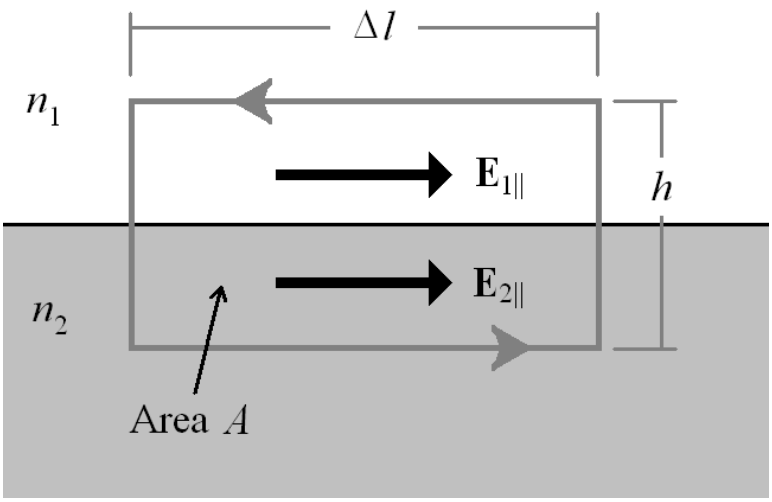


Figure 8.1: Line integral taken around a section of the interface between two media of different refractive index, showing the components of the electric field parallel to the boundary.

8.3.1 The electric field E

Faraday's law is given by

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{8.2}$$

Integrating this over a surface enclosing a volume, we have

$$\int_A (\nabla \times \mathbf{E}) \cdot d\mathbf{A} = -\int_A \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{A}. \tag{8.3}$$

Using Stoke's theorem,

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_A \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{A}. \tag{8.4}$$

Now let us perform the line integral on the left-hand-side of Eq. (8.4) around a rectangular loop set into the interface between two media, as illustrated in Fig. 8.1. Since the integration is taken in the anti-clockwise direction, as the height h of the rectangle tends to zero, we have

$$\lim_{h\to 0} \oint_C \mathbf{E} \cdot d\mathbf{l} = \left(E_{2\|} - E_{1\|}\right) \Delta l. \tag{8.5}$$

Here $E_{1\|}$ and $E_{2\|}$ are the components of the electric field either side of the interface parallel to it and Δl is the length of the rectangle. Now as h tends to zero, so does the area of the rectangle enclosed by the line integral. So,

$$\lim_{h\to 0}\int_A \frac{\partial \mathbf{B}}{\partial t}\cdot d\mathbf{A} = 0 \tag{8.6}$$

and hence, from Eq. (8.4),

$$\left(E_{2\|} - E_{1\|}\right)\Delta l = 0. \tag{8.7}$$

This implies that

$$E_{2\|} = E_{1\|} \tag{8.8}$$

or, expressing this more generally

$$\boxed{E_{\|}\ is\ continuous\ across\ a\ boundary}$$

8.3.2 The field vector H

If there are no free currents, then from Maxwell's law of induction,

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}. \tag{8.9}$$

We may apply exactly the same argument here for the parallel components of $\mathbf{H}$ as we did for the electric field, with the result

$$\boxed{H_{\|}\ is\ continuous\ across\ a\ boundary.}$$

When $\mathbf{j}_f \neq 0$, we have the modified result that

$$H_{2\|} - H_{1\|} = j_f, \tag{8.10}$$

where j_f is the *surface current*.

8.3.3 The electric displacement D

In the absence of any free charges, Gauss' law gives

$$\nabla \cdot \mathbf{D} = 0. \tag{8.11}$$

Using the *divergence theorem*, the volume integral of the left-hand-side is

$$\int_V \nabla \cdot \mathbf{D}\ dV = \int_A \mathbf{D}\cdot d\mathbf{A}. \tag{8.12}$$

Let us perform the surface integral over the surface of cylindrical volume sunk into the boundary between two media of different refractive index, as illustrated in Fig. 8.2. As the height h of the cylinder is taken to zero, we have

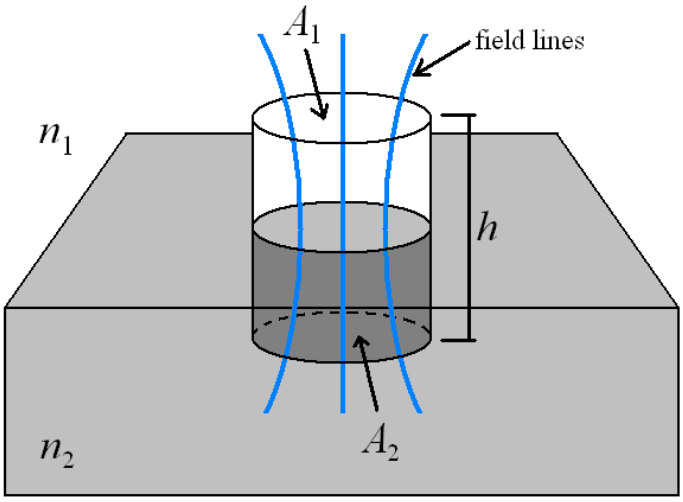


Figure 8.2: A cylindrical volume sunk into the boundary between two media of different refractive index, showing the field lines of $\mathbf{D}$ passing through the circular ends of the volume.

$$\lim_{h\to 0}\int_A \mathbf{D}\cdot d\mathbf{A} = \int_{A_1}\mathbf{D}\cdot d\mathbf{A_1} - \int_{A_2}\mathbf{D}\cdot d\mathbf{A_2}. \tag{8.13}$$

This integrations over the opposite ends of the cylinder have opposite signs since the normal to the surface always points outwards whereas $\mathbf{D}$ points into the volume on one side of the boundary and outwards on the other. Now, in the absence of charge, Eq. (8.11) tells us that

$$\int_{A_1}\mathbf{D}\cdot d\mathbf{A_1} - \int_{A_2}\mathbf{D}\cdot d\mathbf{A_2} = 0. \tag{8.14}$$

Therefore, keeping the circular ends of the cylinder parallel to the boundary, we have

$$D_{1\perp}A_1 = D_{2\perp}A_2, \tag{8.15}$$

where $D_{1\perp}$ and $D_{2\perp}$ are the components of $\mathbf{D}$ either side of the boundary perpendicular to it. Since the areas of the ends of the cylinder, A_1 and A_2, are equal, we have

$$D_{1\perp} = D_{2\perp}. \tag{8.16}$$

In other words,

$$\boxed{D_\perp \textit{ is continuous across a boundary.}}$$

Again, this result requires modification if the free charge density is not zero and we have

$$D_{1\perp} - D_{2\perp} = \sigma_f, \tag{8.17}$$

where σ_f is the *surface charge density*.

8.3.4 The magnetic field B

Since magnetic field lines are always closed, the Maxwell equation

$$\nabla \cdot \mathbf{B} = 0 \tag{8.18}$$

is always valid. Applying the same argument as used for the electric displacement, we then have

$$\boxed{B_\perp \textit{ is continuous across a boundary.}}$$

Note that this is *always* true.

8.4 Reflection and refraction revisited

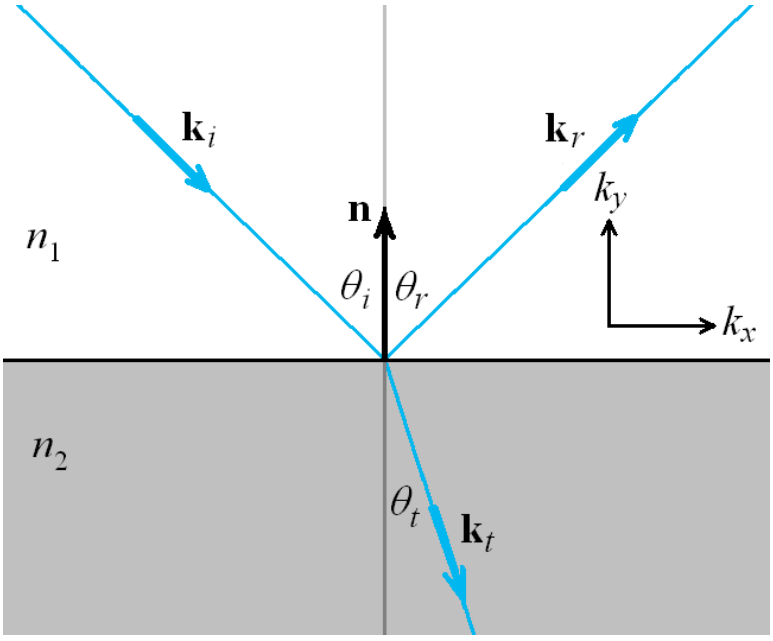


Figure 8.3: Illustration of reflection and refraction in terms of the wavevectors of the incident, reflected and transmitted waves, $\mathbf{k}_i$, $\mathbf{k}_r$ and $\mathbf{k}_t$ respectively.

Armed with the conceptual tools of the previous subsections, we may now return again to the topic of light propagation between different media. Figure 8.3 illustrates the familiar concept of reflection and refraction at the interface between different media pictured, in this case, in terms of the wave-vectors $\mathbf{k}$ (so we may think of this as a visualization of '$\mathbf{k}$-space'). The subscripts i, r and t label the wave-vectors of the incident, reflected and transmitted waves respectively. The plane of incidence is defined by the incident wave-vector $\mathbf{k}_i$ and the normal to the boundary $\mathbf{n}$. Since this defined as being the $k_x - k_y$ plane, $\mathbf{k}_i$ has no k_z-component.

We now apply the boundary condition that the component of the electric field parallel to the interface $\mathbf{E}_\parallel$ is continuous. Therefore

$$\mathbf{E}_\parallel = \mathbf{E}_{i\parallel} e^{i(\omega_i t - \mathbf{k}_i \cdot \mathbf{r})} + \mathbf{E}_{r\parallel} e^{i(\omega_r t - \mathbf{k}_r \cdot \mathbf{r})} = \mathbf{E}_{t\parallel} e^{i(\omega_t t - \mathbf{k}_t \cdot \mathbf{r})}. \tag{8.19}$$

If this is to hold true at all times, the angular frequencies must all be equal, i.e.

$$\omega_i = \omega_r = \omega_t. \tag{8.20}$$

Thus

$$\mathbf{E}_{i\parallel} e^{-i\mathbf{k}_i \cdot \mathbf{r}} + \mathbf{E}_{r\parallel} e^{-i\mathbf{k}_r \cdot \mathbf{r}} = \mathbf{E}_{t\parallel} e^{-i\mathbf{k}_t \cdot \mathbf{r}}. \tag{8.21}$$

We have already chosen our coordinate system so that $\mathbf{k}_i$ has no k_z-component. If we also define the interface between the media, in real space, to be at $y = 0$, then the parallel component of any wave-vector will have no y-component (since this would be transverse to the boundary). Applying these conditions to Eq. (8.21) and re-arranging, we have

$$\mathbf{E}_{i\parallel} e^{-ik_{ix}x} = \mathbf{E}_{t\parallel} e^{-ik_{tx}x} e^{-ik_{tz}z} - \mathbf{E}_{r\parallel} e^{-ik_{rx}x} e^{-ik_{rz}z} \tag{8.22}$$

Now, for the boundary conditions to hold, any phase due to the k_z component on any wave must be common to all terms. For this to be true, these components must all be zero. Hence

$$\mathbf{E}_{i\parallel} e^{-ik_{ix}x} = \mathbf{E}_{t\parallel} e^{-ik_{tx}x} - \mathbf{E}_{r\parallel} e^{-ik_{rx}x}. \tag{8.23}$$

So for this to hold for all x in real space we must have

$$k_{ix} = k_{rx} = k_{tx}. \tag{8.24}$$

The magnitude of the incident wave-vector is

$$|\mathbf{k}_i| = n_1 |\mathbf{k}_0|, \tag{8.25}$$

where n_1 is the refractive index of the incident medium and $\mathbf{k}_0$ is the free space wave-vector. Since the reflected wave travels in the same medium, we also have

$$|\mathbf{k}_r| = n_1 |\mathbf{k}_0| = |\mathbf{k}_i| . \tag{8.26}$$

Similarly, the magnitude of the transmitted wave is

$$|\mathbf{k}_t| = n_2 |\mathbf{k}_0| , \tag{8.27}$$

where n_2 is the refractive index of the second medium. Inspecting Fig. 8.3 and using the result of Eq. (8.24) for the equality of the k_x components, we see that

$$k_{ix} = |\mathbf{k}_i| \sin\theta_i = k_{ir} = |\mathbf{k}_r| \sin\theta_r. \tag{8.28}$$

Using Eq. (8.26), we then have

$$n_1 |\mathbf{k}_0| \sin\theta_i = n_1 |\mathbf{k}_0| \sin\theta_r, \tag{8.29}$$

which implies

$$\boxed{\theta_i = \theta_r,} \tag{8.30}$$

i.e. the *Law of Reflection*.

Proceeding along similar lines, Eq. (8.24) also implies that

$$k_{ix} = |\mathbf{k}_i| \sin\theta_i = k_{it} = |\mathbf{k}_t| \sin\theta_t, \tag{8.31}$$

which, upon substitution from Eqs. (8.26) and (8.27) gives

$$n_1 |\mathbf{k}_0| \sin\theta_i = n_2 |\mathbf{k}_0| \sin\theta_t. \tag{8.32}$$

Thus, dividing by $|\mathbf{k}_0|$ we arrive at *Snell's Law* once again

$$\boxed{n_1 \sin\theta_i = n_2 \sin\theta_t.} \tag{8.33}$$

8.5 Fresnel equations

8.5.1 s and p-type polarisation

Before proceeding, we briefly introduce a common scheme for denoting the polarisation of waves propagating across a dielectric boundary. The notation is only valid for linearly polarised light and is given in reference to the plane of incidence defined by the incident wave-vector and the normal to the boundary.

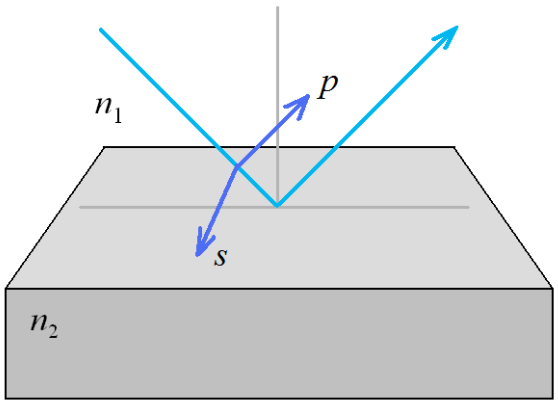


Figure 8.4: Sketch showing the orthogonal components of the polarisation parallel (p) and perpendicular (s) to the plane of incidence (the plane containing the incident and reflected ray of light).

s-polarised

This refers to light polarised *perpendicular* to the plane of incidence. The 's' stands for the German word *senkrecht*.

This is also often referred to as *transverse electric* (TE). Note that for linearly polarised light, the electric field will be perpendicular, or transverse, to the plane of incidence.

p-polarised

This refers to light polarised *parallel* to the plane of incidence.

This is also often referred to as *transverse magnetic* (TM). Note that for linearly polarised light, the magnetic field will be perpendicular, or transverse, to the plane of incidence.

8.5.2 Derivation of the Fresnel equations

s-polarised light

Figure 8.5 illustrates s-type polarisation (with the electric field $\mathbf{E}$ normal to the page) showing the magnetic field $\mathbf{B}$, with the incident light taken to be from the medium with refractive index n_1. From the continuity of the $H_{\|}$ at the boundary, we have

$$H_{si}\cos\theta_i - H_{sr}\cos\theta_r = H_{st}\cos\theta_t. \tag{8.34}$$

Similarly, the continuity of $B_{\perp}$ gives

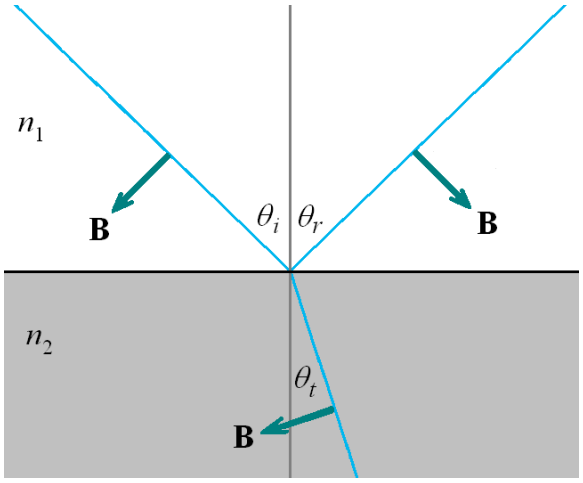


Figure 8.5: s-type polarisation (with the electric field $\mathbf{E}$ normal to the page) showing the magnetic field $\mathbf{B}$.

$$B_{si}\sin\theta_i + B_{sr}\sin\theta_r = B_{st}\sin\theta_t. \tag{8.35}$$

Using $\theta_i = \theta_r = \theta_1$ and $\theta_t = \theta_2$, we can rewrite these expressions, putting H in terms of B in Eq. (8.34) via the relative permeability. This gives

$$\frac{1}{\mu_1}(B_{si} - B_{sr})\cos\theta_1 = \frac{1}{\mu_2}B_{st}\cos\theta_2 \tag{8.36}$$

and

$$(B_{si} + B_{sr})\sin\theta_1 = B_{st}\sin\theta_2. \tag{8.37}$$

Now, we may use *Snell's Law* to substitute for $\sin\theta_2$ so that Eq. (8.37) becomes

$$(B_{si} + B_{sr}) = B_{st}\frac{n_1}{n_2}. \tag{8.38}$$

Eliminating B_{st} from Eqs. (8.36) and (8.38) gives

$$\frac{\mu_2}{\mu_1}(B_{si} - B_{sr})\frac{\cos\theta_1}{\cos\theta_2} = (B_{si} + B_{sr})\frac{n_2}{n_1}. \tag{8.39}$$

or

$$B_{si}\left(\frac{\mu_2}{\mu_1}\frac{\cos\theta_1}{\cos\theta_2} - \frac{n_2}{n_1}\right) = B_{sr}\left(\frac{\mu_2}{\mu_1}\frac{\cos\theta_1}{\cos\theta_2} + \frac{n_2}{n_1}\right). \tag{8.40}$$

The ratio B_{sr}/B_{si} is therefore

$$\frac{B_{sr}}{B_{si}} = \frac{n_1'\cos\theta_1 - n_2'\cos\theta_2}{n_1'\cos\theta_1 + n_2'\cos\theta_2}. \tag{8.41}$$

where $n_i' = n_i/\mu_i$. Now, from Eqs. (6.26) and (6.51) of Chapter 6, we have

$$|\mathbf{B}| = \frac{nk_0}{\omega} |\mathbf{E}| . \tag{8.42}$$

Since ω is the same in either medium we have

$$\frac{B_{sr}}{B_{si}} = \frac{E_{sr}}{E_{si}} \tag{8.43}$$

and thus the reflection coefficient for s-polarised light is

$$\boxed{r_s = \frac{E_{sr}}{E_{si}} = \frac{n_1' \cos\theta_1 - n_2' \cos\theta_2}{n_1' \cos\theta_1 + n_2' \cos\theta_2}.} \tag{8.44}$$

If we now eliminate $B_s r$ from Eqs. (8.36) and (8.38), we get

$$(B_{si} - B_{sr}) = \frac{\mu_1}{\mu_2} B_{st} \frac{\cos\theta_2}{\cos\theta_1} \tag{8.45}$$

$$(B_{si} + B_{sr}) = B_{st} \frac{n_1}{n_2}. \tag{8.46}$$

$$B_{si} - B_{st} \frac{\mu_1}{\mu_2} \frac{\cos\theta_2}{\cos\theta_1} = B_{st} \frac{n_1}{n_2} - B_{si}, \tag{8.47}$$

which gives

$$2B_{si} = B_{st} \left(\frac{n_1}{n_2} + \frac{\mu_1}{\mu_2} \frac{\cos\theta_2}{\cos\theta_1} \right). \tag{8.48}$$

and therefore

$$\frac{B_{st}}{B_{si}} = \frac{2(n_2/\mu_1)\cos\theta_1}{n_1' \cos\theta_1 + n_2' \cos\theta_2}. \tag{8.49}$$

In this case, we have

$$\frac{E_{st}}{E_{si}} = \frac{n_1 B_{st}}{n_2 B_{si}}. \tag{8.50}$$

Thus,

$$\boxed{t_s = \frac{E_{st}}{E_{si}} = \frac{2n_1' \cos\theta_1}{n_1' \cos\theta_1 + n_2' \cos\theta_2}.} \tag{8.51}$$

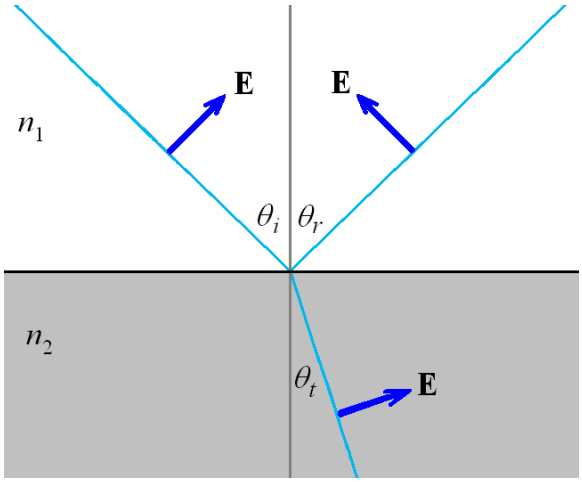


Figure 8.6: p-type polarisation, in which $\mathbf{E}$ is parallel to the plane of incidence.

p-polarised light

In this case, we consider p-type polarisation in terms of the electric field. Figure 8.6 illustrates the situation. Now, the continuity of $E_{\|}$ at the boundary gives

$$E_{pi}\cos\theta_i - E_{pr}\cos\theta_r = E_{pt}\cos\theta_t \tag{8.52}$$

and the continuity of $D_{\perp}$ gives

$$D_{pi}\sin\theta_i + D_{pr}\sin\theta_r = D_{pt}\sin\theta_t. \tag{8.53}$$

Using $\theta_i = \theta_r = \theta_1$ and $\theta_t = \theta_2$, we can rewrite these expressions, with Eq. (8.53) in terms of $\mathbf{E}$, as

$$(E_{pi} - E_{pr})\cos\theta_1 = E_{pt}\cos\theta_2 \tag{8.54}$$

and

$$\epsilon_1 (E_{pi} + E_{pr})\sin\theta_1 = \epsilon_2 E_{pt}\sin\theta_2. \tag{8.55}$$

Using *Snell's Law* in conjunction with

$$\frac{n_1}{n_2} = \frac{n_2\epsilon_1\mu_1}{n_1\epsilon_2\mu_2}, \tag{8.56}$$

Eq. (8.55) becomes

$$\epsilon_1 (E_{pi} + E_{pr}) = \epsilon_2 E_{pt}\frac{n_1}{n_2} = E_{pt}\frac{n_2\epsilon_1\mu_1}{n_1\mu_2}. \tag{8.57}$$

Hence

$$(E_{pi} + E_{pr}) = E_{pt}\frac{n_2\mu_1}{n_1\mu_2} = E_{pt}\frac{n_2'}{n_1'}, \tag{8.58}$$

where, again, $n_i' = n_i/\mu_i$.

Eliminating E_{pt} from Eqs. (8.54) and (8.58), we have

$$(E_{pi} - E_{pr})\,\frac{\cos\theta_1}{\cos\theta_2} = (E_{pi} + E_{pr})\,\frac{n_1'}{n_2'}, \tag{8.59}$$

which may be rearranged to give

$$\boxed{r_p = \frac{E_{pr}}{E_{pi}} = \frac{n_2'\cos\theta_1 - n_1'\cos\theta_2}{n_2'\cos\theta_1 + n_1'\cos\theta_2}.} \tag{8.60}$$

Eliminating E_{pr} from Eqs. (8.54) and (8.58), we have

$$2E_{pi} = E_{pt}\left(\frac{n_2'}{n_1'} + \frac{\cos\theta_2}{\cos\theta_1}\right) \tag{8.61}$$

or

$$\boxed{t_p = \frac{E_{pt}}{E_{pi}} = \frac{2n_1'\cos\theta_1}{n_2'\cos\theta_1 + n_1'\cos\theta_2}.} \tag{8.62}$$

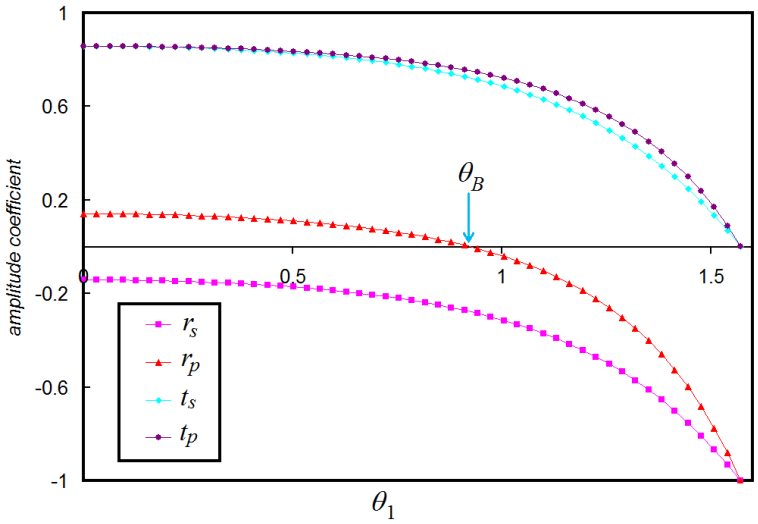


Figure 8.7: Graph of the s and p-polarised reflection and transmission coefficients for light incident from the lower n_1 medium. Note that r_p goes through zero at the *Brewster angle* θ_B.

8.5.3 Alternative forms

Eliminating the refractive indices using *Snell's Law* the Fresnel equations for the reflection and transmission coefficients for s and p-polarised light may be given in the following alternative forms.

s-polarised

Reflection

$$\boxed{r_s = \frac{E_{sr}}{E_{si}} = \frac{n_1 \cos\theta_i - n_2 \cos\theta_t}{n_1 \cos\theta_i + n_2 \cos\theta_t} = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)},} \tag{8.63}$$

Transmission

$$\boxed{t_s = \frac{E_{st}}{E_{si}} = \frac{2n_1 \cos\theta_i}{n_1 \cos\theta_i + n_2 \cos\theta_t} = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t)},} \tag{8.64}$$

p-polarised

Reflection

$$\boxed{r_p = \frac{E_{pr}}{E_{pi}} = \frac{n_2 \cos\theta_i - n_1 \cos\theta_t}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)},} \tag{8.65}$$

Transmission

$$\boxed{t_p = \frac{E_{pt}}{E_{pi}} = \frac{2n_1 \cos\theta_i}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}.} \tag{8.66}$$

Phase change on reflection

These results are plotted in Fig.8.7 for light incident from the n_1 (lower refractive index) medium. Note that r_s is *negative* over every angle of incidence and r_p becomes negative at oblique angles. This corresponds to a *phase change* of π as the light passes from the lower refractive medium to the higher. In general, it is possible for the coefficiencts to exhibit other phase changes and must therefore be considered to be *complex* quantities.

8.5.4 Brewster angle

Equation (8.65) for the reflection coefficient r_p may be written as

$$r_p = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)} = \frac{\sin(\theta_i - \theta_t)\cos(\theta_i + \theta_t)}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}. \tag{8.67}$$

Now when $\theta_i + \theta_t = \pi/2$, $\cos(\theta_i + \theta_t) = 0$, so $r_p = 0$. In other words, *no p-polarised light is reflected*. The angle of incidence θ_B at which this occurs is called the *Brewster angle*. This is a case of polarisation on reflection, since the reflected light will all be s-polarised.

We may derive an expression for the Brewster angle by noting that, using Snell's Law, $\theta_B + \theta_t = \pi/2$ implies

$$\theta_t = \sin^{-1}\left(\frac{n_1}{n_2}\sin\theta_B\right) = \frac{\pi}{2} - \theta_B. \tag{8.68}$$

Taking the $\sin$ of this, we have

$$\frac{n_1}{n_2}\sin\theta_B = \sin\left(\frac{\pi}{2} - \theta_B\right) = \sin\frac{\pi}{2}\cos\theta_B + \sin\theta_B\cos\frac{\pi}{2} = \cos\theta_B. \tag{8.69}$$

Rearranging and taking the arctangent of both sides then gives us our required result

$$\boxed{\theta_B = \tan^{-1}\left(\frac{n_2}{n_1}\right).} \tag{8.70}$$

Road glare

The phenomenon of polarisation by reflection at the Brewster angle gives us a strategy for reducing glare. For example, Fig. 8.8 illustrates the case on a sunny day when light reflected from the road or bonnet of a car can have an adverse effect on visibility for the motorist. Most of the reflected light will be s-polarised (aligned parallel to the ground). Polarising sunglasses are usually designed with the transmission axis in the vertical direction. Hence, such glasses will cut out much of the glare.

Occasionally, car wind screens are also treated with a polarising film. So long as the transmission axes of the screen and glasses are aligned, there will be little problem. However, if the driver were to rotate his or her lenses by 90° then the polarisers would become *crossed* and almost all the light would be blocked!

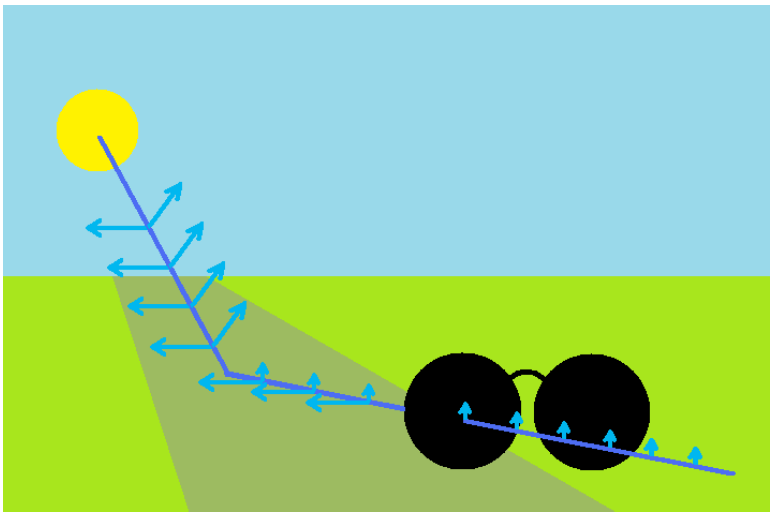
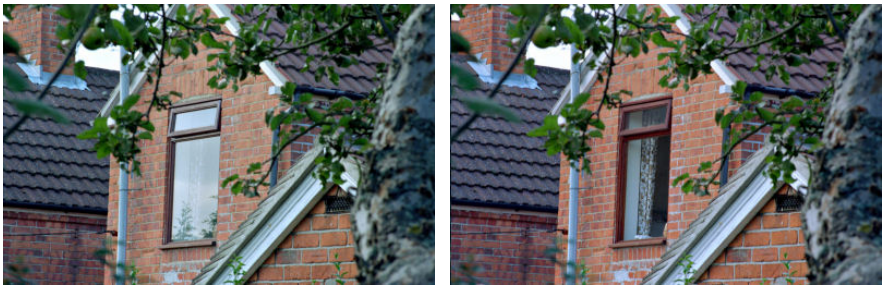


Figure 8.8: On a sunny day, there may be significant glare from the road or the bonnet of a car. Since this is reflected light, it will tend to be s-polarised. This glare can then be greatly reduced by wearing polarising sunglasses to block the s-polarised component.



(a) (b)

Figure 8.9: An example of the low reflection coefficient near the Brewster angle for p-polarised light. In (a), the window is practically opaque due to the bright reflection from the Sun. In (b), however, a polarising filter is used on the camera to block s-polarised light. Since the amplitude of p-polarised light is already low, most of the light seen from the window is transmitted through it from the room inside.

Polarizing filter on camera

A similar application is shown in Fig. 8.9. Here, a polarising filter attached to a camera blocks most of the reflected light from a window. This has the result that most of the light from the window is now transmitted from inside the room. In other words, the interior of the room is now made visible.

8.6 Time reversibility

In Eqs. (8.63) to (8.66), the light begins in the medium labelled with the 'i' subscript and is transmitted into the medium labelled by 't'. Let us consider the case where time is reversed, so that the light begins in the medium labelled 't' and is transmitted to the 'i' medium. To evaluate the reflection and transmission coefficients, we simply need to swap the positions of the angles θ_i and θ_t. We shall denote the time-reversed coefficients using prime $(')$ notation.

For reflection, on time reversal we have

$$r'_s = -\frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)} = -r_s \tag{8.71}$$

and

$$r'_p = \frac{\tan(\theta_t - \theta_i)}{\tan(\theta_t + \theta_i)} = -r_p. \tag{8.72}$$

We see that, in general

$$\boxed{r' = -r.} \tag{8.73}$$

Note, this tells us that if there is a phase change of pi associated with r, there is no such phase change for r'. In particular, in comparison with Fig. 8.7 this means that there is no phase change on reflection for s-polarised light when it is incident from the higher refractive index medium.

It will prove useful to derive a further result here for $1 - r^2$. For s-polarised waves

$$1 - r_s^2 = 1 - \frac{\sin^2(\theta_i - \theta_t)}{\sin^2(\theta_i + \theta_t)} = \frac{\sin^2(\theta_i + \theta_t) - \sin^2(\theta_i - \theta_t)}{\sin^2(\theta_i + \theta_t)},$$

giving

$$\boxed{1 - r_s^2 = \frac{4\sin\theta_i \cos\theta_i \sin\theta_t \cos\theta_t}{\sin^2(\theta_i + \theta_t)}.} \tag{8.74}$$

Similarly, we have

$$\begin{aligned} 1 - r_p^2 &= 1 - \frac{\tan^2(\theta_i - \theta_t)}{\tan^2(\theta_i + \theta_t)} = \frac{\tan^2(\theta_i + \theta_t) - \tan^2(\theta_i - \theta_t)}{\tan^2(\theta_i + \theta_t)}, \\ &= \frac{\sec^2(\theta_i + \theta_t) - \sec^2(\theta_i - \theta_t)}{\tan^2(\theta_i + \theta_t)}, \\ &= \frac{\cos^2(\theta_i - \theta_t) - \cos^2(\theta_i + \theta_t)}{\cos^2(\theta_i - \theta_t)\sin^2(\theta_i + \theta_t)}, \end{aligned}$$

giving

$$\boxed{1 - r_p^2 = \frac{4 \sin\theta_i \cos\theta_i \sin\theta_t \cos\theta_t}{\cos^2(\theta_i - \theta_t) \sin^2(\theta_i + \theta_t)}.} \tag{8.75}$$

Now, for transmission, we have

$$t_s = \frac{2 \cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t)} \tag{8.76}$$

and

$$t_p = \frac{2 \cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t) \cos(\theta_i - \theta_t)}. \tag{8.77}$$

So

$$t_s' = \frac{2 \cos\theta_t \sin\theta_i}{\sin(\theta_t + \theta_i)} \tag{8.78}$$

and

$$t_p' = \frac{2 \cos\theta_t \sin\theta_i}{\sin(\theta_t + \theta_i) \cos(\theta_t - \theta_i)}. \tag{8.79}$$

Hence, for s-polarised waves

$$t_s' t_s = \frac{4 \sin\theta_i \cos\theta_i \sin\theta_t \cos\theta_t}{\sin^2(\theta_i + \theta_t)} = 1 - r_s^2, \tag{8.80}$$

whilst for p-polarised waves

$$t_p' t_p = \frac{4 \sin\theta_i \cos\theta_i \sin\theta_t \cos\theta_t}{\cos^2(\theta_i - \theta_t) \sin^2(\theta_i + \theta_t)} = 1 - r_p^2. \tag{8.81}$$

So, in general, we have

$$\boxed{t't = 1 - r^2.} \tag{8.82}$$

8.7 Stoke's treatment

The results of the previous section may also be derived following a somewhat more intuitive approach via the *Stoke's treatment*. In Fig 8.10 (a), we see the reflection and transmission of an initial ray of amplitude E_0. The time-symmetric counterpart of this situation is shown in Fig 8.10 (b). In the latter case, there are two initial waves: one with amplitude rE_0 incident from the n_1 side of the boundary and another with amplitude tE_0 incident from the n_2 side.

Figure 8.11 shows the combined effect of these two situations. We note the following:

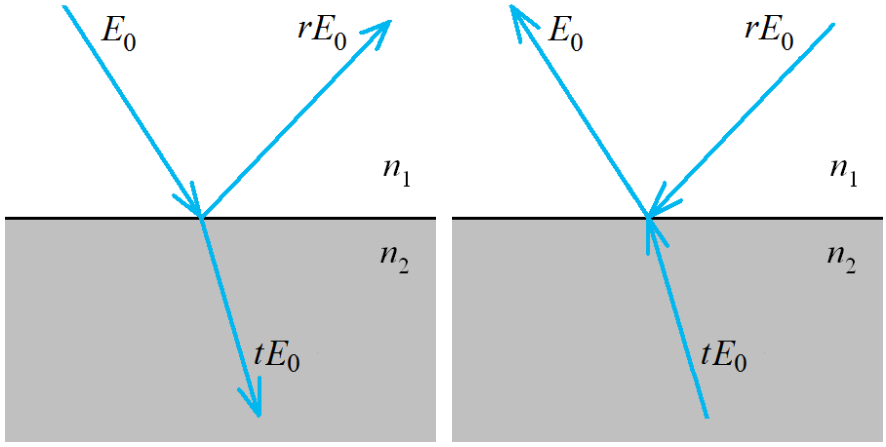


Figure 8.10: (a) The reflection and transmission of an initial ray of amplitude E_0. (b) The time-symmetric counterpart to (a).

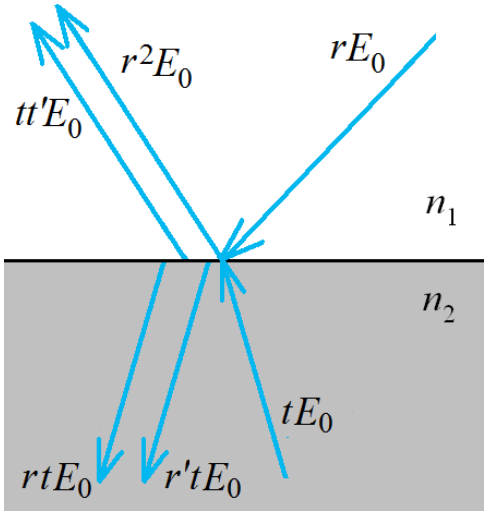


Figure 8.11: The combined effect of the cases in Fig. 8.10.

- The reflected component of rE_0 is r^2E_0, since this is reflection in the n_1 medium (i.e. we do *not* multiply by r')
- The reflected component of tE_0 in the n_2 medium is $r'tE_0$, since this is the *opposite* case to reflection in the n_1 medium.

We may now draw the following conclusions. Firstly, since the total amplitude of the rays on the left-hand-side of Fig. 8.11 in the n_2 medium must cancel out, we have

$$\left(rt + r't\right) E_0 = 0, \tag{8.83}$$

which implies

$$\boxed{r' = -r.} \tag{8.84}$$

Secondly, we must also have

$$\left(r^2 + t't\right) E_0 = E_0, \tag{8.85}$$

which implies

$$\boxed{t't = 1 - r^2.} \tag{8.86}$$

Thus, we have reproduced the results of the previous section.

8.8 Irradiance

In Chapter 6, we found that the time averaged *Poynting vector* in an isotropic medium is given by

$$\langle \mathbf{S} \rangle = \tfrac{1}{2} \hat{\mathbf{k}} E_0^2 \left(\frac{\epsilon \epsilon_0}{\mu \mu_0} \right)^{1/2}. \tag{8.87}$$

This may be written in terms of the speed of light $c = (\varepsilon_0 \mu_0)^{-1/2}$ and the refractive index $n = (\varepsilon \mu)^{1/2}$ to obtain

$$\langle \mathbf{S} \rangle = \frac{n E_0^2}{2 \mu \mu_0 c} \hat{\mathbf{k}}. \tag{8.88}$$

Note that this gives the *intensity* of the light or, equivalently, the *irradiance*. We can therefore write Eq. (8.88) as

$$\langle \mathbf{S} \rangle = I \hat{\mathbf{k}}. \tag{8.89}$$

8.8.1 Reflectance and transmittance

Figure 8.12 shows a beam of light incident on the boundary between media of refractive indices n_1 and n_2. Also shown is the reflected beam with $\theta_r = \theta_i$ as required by the Law of Reflection. The intensity over the surface at A with normal vector $\mathbf{n}$ is given by

$$I_A = -I_i \hat{\mathbf{k}}_i \cdot \mathbf{n} = I_i \cos \theta_i. \tag{8.90}$$

This also equals

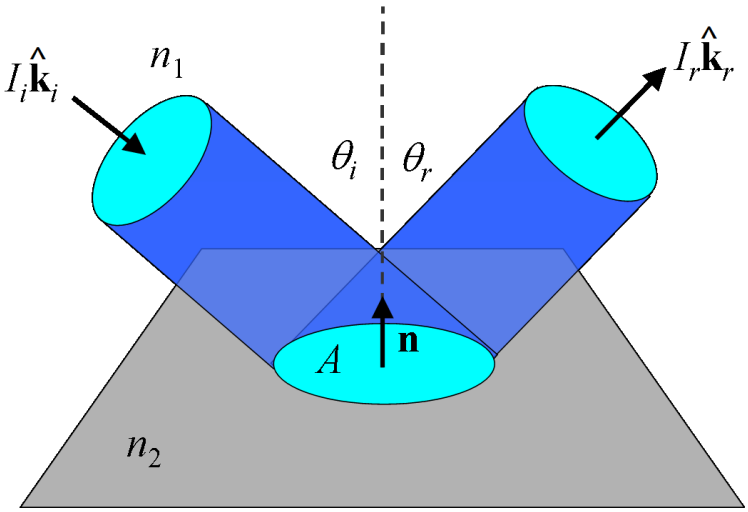


Figure 8.12: A beam of light incident on the boundary between media of refractive indices n_1 and n_2 together with the reflected beam. Note that $\theta_r = \theta_i$ as required by the Law of Reflection. The area over which the incident intensity is spread is denoted by A and has normal vector $\mathbf{n}$.

$$I_A = I_r\hat{\mathbf{k}}_r \cdot \mathbf{n} = I_r \cos\theta_r. \tag{8.91}$$

We may then define the *reflectance* R as the ratio of the reflected and incident intensities.

$$R = \frac{I_r \cos\theta_r}{I_i \cos\theta_i} = \frac{I_r}{I_i} = \left(\frac{E_{0r}}{E_{0i}}\right)^2. \tag{8.92}$$

Recalling the results of the previous section, we may express this in terms of the reflection coefficient r

$$\boxed{R = r^2.} \tag{8.93}$$

An example of intensity reflectance is shown in Fig. 8.13. Note that the reflectance is low for small angles, only climbing significantly beyond the Brewster angle. This has a familiar consequence when viewing a reflective surface from a low angle and the surface mirrors almost perfectly the surrounding environment (see Fig. 8.14).

The *transmittance* T may be defined as the ratio of the transmitted intensity to the incident intensity. By a similar argument as before, we then have

$$T = \frac{I_t \cos\theta_t}{I_i \cos\theta_i} = \frac{n_2 \cos\theta_t}{n_1 \cos\theta_i}\left(\frac{E_{0t}}{E_{0i}}\right)^2. \tag{8.94}$$

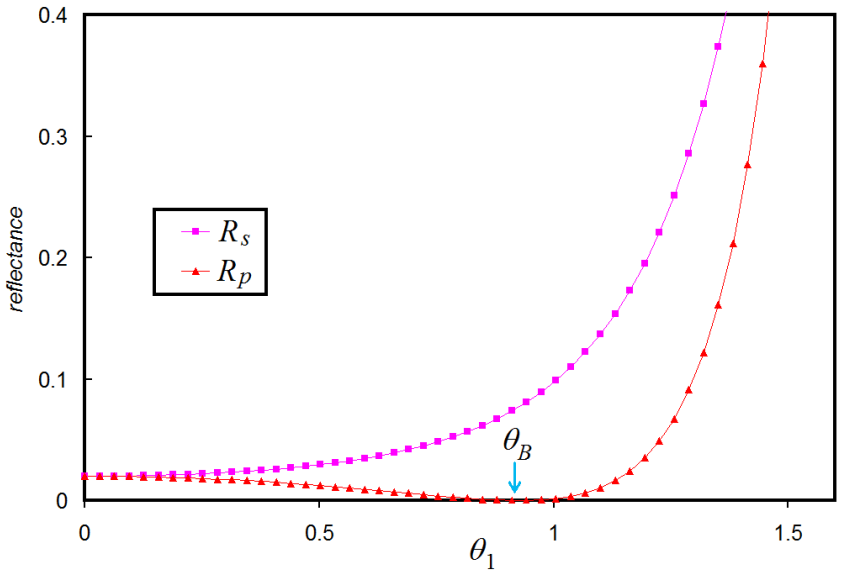


Figure 8.13: Chart of the intensity reflectance for s and p-polarised light. Here, the refractive index of the incident medium $n_1 = 1$ and the transmitted medium has $n_2 = 1.33$ (note that this is the refractive index of water).

Again, we may express this in terms of the transmission coefficient t

$$\boxed{T = \frac{n_2 \cos\theta_t}{n_1 \cos\theta_i} t^2.} \tag{8.95}$$

The conservation of energy

Imposing the principle of the conservation of energy (energy in = energy out), we have

$$I_i \cos\theta_i = I_r \cos\theta_r + I_t \cos\theta_t \tag{8.96}$$

or, dividing by $I_i \cos\theta_i$

$$1 = \frac{I_r \cos\theta_r}{I_i \cos\theta_i} + \frac{I_t \cos\theta_t}{I_i \cos\theta_i}. \tag{8.97}$$

Thus,

$$\boxed{R + T = 1.} \tag{8.98}$$

Figure 8.14: View of a water surface from a low angle. Most of the incident light at such an angle is reflected.

8.9 Total internal reflection

In Chapter 4 we encountered the phenomenon of *total internal reflection* in which there is no transmission from a higher refractive index medium to a lower for incident angles less than the *critical angle*. Here, we shall modify this conclusion slightly to show that there is, in fact, a decaying wave transmitted into the lower refractive index medium.

Let us reconsider Fig. 8.3. We shall consider the case where $n_1 > n_2$, so the total internal reflection will take place on the n_1 side of the boundary. Now the y component of $\mathbf{k}_t$ is given by

$$k_{ty} = k_t \cos\theta_t = k_t \left(1 - \sin^2\theta_t\right)^{1/2}. \tag{8.99}$$

Using Snell's Law, this becomes

$$k_{ty} = k_t \cos\theta_t = k_t \left(1 - \left[\frac{n_1}{n_2}\right]^2 \sin^2\theta_t\right)^{1/2}. \tag{8.100}$$

This expression equals zero (implying zero transmission) when

$$\sin\theta_i = \frac{n_2}{n_1} = \sin\theta_c, \tag{8.101}$$

where θ_c is the *critical angle*. So, for $\theta_i > \theta_c$, we have

$$k_{ty} = ik_t \left(\left[\frac{n_1}{n_2} \right]^2 \sin^2 \theta_t - 1 \right)^{1/2} \equiv i\gamma_{ty}. \tag{8.102}$$

In general, we can write the transmitted wave as

$$E_t = E_{t0} e^{i(\omega t - k_{tx}x - k_{ty}y)}. \tag{8.103}$$

Substituting $i\gamma_{ty}$ for k_{ty}, we have

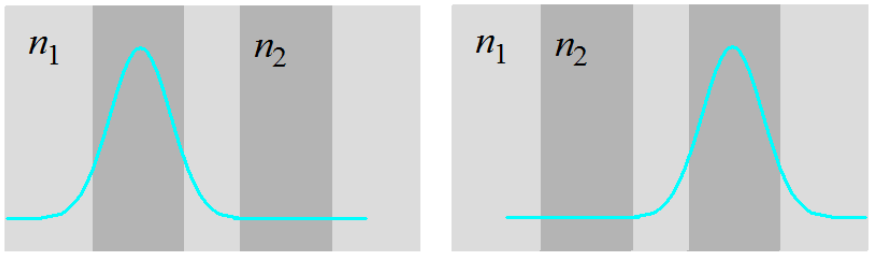
$$E_t = E_{t0} e^{i(\omega t - k_{tx}x - ik_{ty}y)}, \tag{8.104}$$

which becomes

$$\boxed{E_t = E_{t0} e^{i(\omega t - k_{tx}x) + \gamma_{ty}y}.} \tag{8.105}$$

Since we have set the problem up so that the transmitted wave travels in the $-$ve y-direction, this represents an *exponentially decaying wave*. This is known as the *evanescent wave*.

8.9.1 Optical coupling in waveguides



(a) (b)

Figure 8.15: Optical coupling in two adjacent waveguides ($n_2 > n_1$). If the waveguides are close together, the evanescent wave may overlap the adjacent guide, leading to a coupling of power from one guide to the other.

The existence of the evanescent wave has a real-world application in *optical coupling*. Figure 8.15 shows an example for two adjacent slab waveguides. As is suggested in the diagram, the electric field is not entirely contained within a given waveguide, with the evanescent wave spreading out from it. If the waveguides are close enough, there may be significant overlap of the evanescent wave in the adjoining guide. It can be shown that this leads to a *coupling* of power between the guides.

A common example of an optical coupler arises in fibre optics. In this case, the central cores of two optical fibres are brought close together leading to power transfer between the fibres (see Fig 8.16).

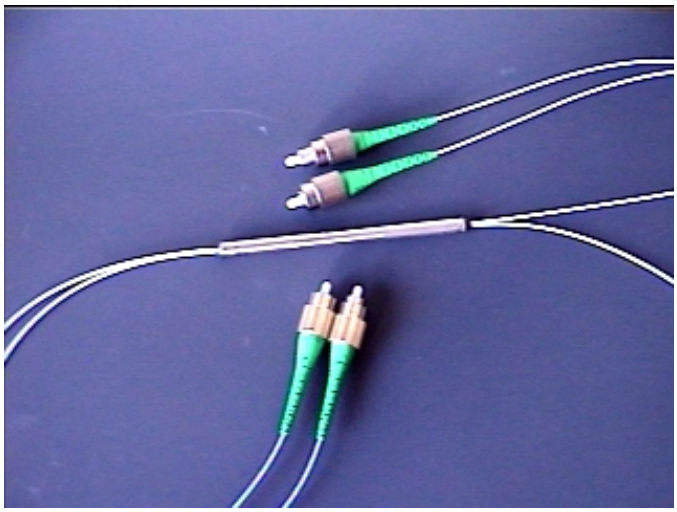


Figure 8.16: Fibre optic coupler. Note that (as is common) the second input port has been removed.

Another use of couplers is in interferometry, when optical couplers are used as the basis of the fibre optic *Mach-Zehnder interferometer*. In this case, the power from an input fibre is divided equally between to output fibres. A phase shift is then applied to one arm of the Mach-Zehnder interferometer before the power is recombined (with interference effects) at a second coupler.

8.10 Summary

- **Boundary conditions**

  The boundary conditions of the electromagnetic fields at interfaces between media of different refractive indices.

$$\boxed{E_{\|} \textit{ is continuous across a boundary}}$$

$$\boxed{H_{\|} \textit{ is continuous across a boundary.}}$$

When $\mathbf{j}_f \neq 0$, we have the modified result that

$$H_{2\|} - H_{1\|} = j_f, \tag{8.106}$$

$$\boxed{D_{\perp} \textit{ is continuous across a boundary.}}$$

If the free charge density is not zero and we have

$$D_{1\perp} - D_{2\perp} = \sigma_f, \tag{8.107}$$

where σ_f is the *surface charge density*.

$$\boxed{B_\perp \; is \; continuous \; across \; a \; boundary.}$$

This result is *always* true.

- **Reflection and refraction**

  The Laws of Reflection and Refraction may be derived by considering wavevector.

- **Polarisation of incident wave**

  - s*-polarised*
    This refers to light polarised *perpendicular* to the plane of incidence. The 's' stands for the German word *senkrecht*.
  - p*-polarised*
    This refers to light polarised *parallel* to the plane of incidence.

- **Fresnel equations**

  The Fresnel equations for the reflection and transmission coefficients r and t.

  - s*-polarised*
    - *Reflection*

$$\boxed{r_s = \frac{n_1 \cos\theta_i - n_2 \cos\theta_t}{n_1 \cos\theta_i + n_2 \cos\theta_t} = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}.} \tag{8.108}$$

    - *Transmission*

$$\boxed{t_s = \frac{2n_1 \cos\theta_i}{n_1 \cos\theta_i + n_2 \cos\theta_t} = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t)}.} \tag{8.109}$$

  - p*-polarised*
    - *Reflection*

$$\boxed{r_p = \frac{n_2 \cos\theta_i - n_1 \cos\theta_t}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}.} \tag{8.110}$$

∗ *Transmission*

$$\boxed{t_p = \frac{2n_1 \cos\theta_i}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}.} \tag{8.111}$$

– *Brewster angle*
When the angle of incidence equals the *Brewster angle*, the reflected light is entirely s-polarised.

$$\boxed{\theta_B = \tan^{-1}\left(\frac{n_2}{n_1}\right).} \tag{8.112}$$

– *Phase change on reflection* s-polarised light incurs a phase change of π on reflection from a lower refractive index medium for every angle of incidence. r_p-polarised light also has a phase change of π when the angle of incidence is greater than the *Brewster angle*.

- **Time reversibility**
Use of the principle of time reversibility for analysis of reflection and transmission coefficients.

$$\boxed{r' = -r.} \tag{8.113}$$

$$\boxed{t't = 1 - r^2.} \tag{8.114}$$

- **Stokes treatment**
Alternative derivation of the results of time reversibility via the Stoke's treatment.

- **Irradiance**
Analysis of power reflection and transmission between different media.

– *Reflectance*

$$\boxed{R = r^2.} \tag{8.115}$$

– *Transmittance*

$$\boxed{T = \frac{n_2 \cos\theta_t}{n_1 \cos\theta_i} t^2.} \tag{8.116}$$

  - *Conservation of energy*

$$\boxed{R + T = 1.} \tag{8.117}$$

- **Total internal reflection**

  Wavevector analysis of total internal reflection to obtain the *evanescent wave*.

  The evanescent wave decays as $e^{-\gamma y}$, where γ is the *imaginary* wavevector.

  - *Optical couplers*
    The evanescent wave may be exploited to obtain power transmission between adjacent waveguides.

Part IV

Geometrical Optics

9. Fermat's Principle

9.1 General remarks

Although the subject of *geometrical optics* could be approached from wave optics, it is more typically presented in terms of *ray tracing*. For this, the appropriate theoretical tool is *Fermat's Principle*.

This is a variational principle similar to Hero of Alexandria's assertion that light travels by the shortest geometrical path. In fact, in an isotropic and homogeneous media, Fermat's Principle reduces to Hero's. More generally, however, Fermat's Principle states that light travels the path of *shortest time*. This is rendered in terms of the *optical path length*, which, although having dimensions of space, is actually proportional to the propagation time.

Using Fermat's Principle, we begin in this chapter by deriving

- The Law of Rectilinear Propagation
- The Law of Reflection
- The Law of Refraction (Snell's Law)

9.2 Learning objectives

The aims of this section are to gain understanding of

- Optical path length
- Fermat's Principle
- Application of this principles to find:
  - The Law of Rectilinear Propagation
  - The Law of Reflection
  - The Law of Refraction (Snell's Law)
- Perfect imaging: imaging all rays perfectly from a point or plane onto another point or plane
- Application of Fermat's Principle to analyse

  - Perfect mirrors
  - Perfect lenses

- The concept of curvature
- Finding the curvature of a surface

9.3 Geometric wavefront

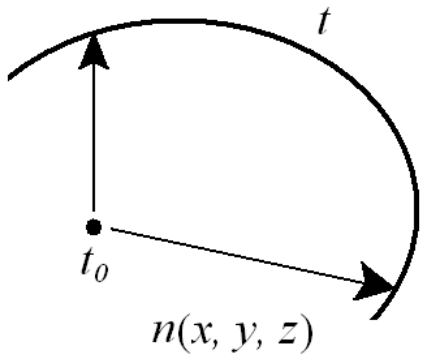


Figure 9.1: The locus of points with equal optical path-length at time t for rays emanating from a single point at time t_0 constitutes a geometric wavefront.

Figure 9.1 shows the locus of points on rays emanating from a single point at time t_0 at a later time t. These points are all in phase with one another and constitute a *geometric wavefront*. Putting $T = t - t_0$, we may then multiply T by c to express the propagation time in dimensions of space

$$\Lambda(r) = c\,(t - t_0)\,. \tag{9.1}$$

The quantity $\Lambda(r)$ is known as the *optical path length* and is a function of distance. Now, if $dS/dt = v = c/n$, then

$$\Lambda(r) = \int_0^r n(x, y, z)\; dS. \tag{9.2}$$

In an isotropic and homogeneous medium, $n(x, y, z) = n$, a constant, so

$$\Lambda(r) = n \int_0^r dS = nr, \tag{9.3}$$

i.e. a sphere centered on the emanation point of the rays.

As $r \to \infty$, over the same distance of arc the wavefronts will tend to a straight line and become plain waves. That is, all the points in a plain orthogonal to the direction of propagation have the same phase.

9.4 Fermat's Principle

In it's original form *Fermat's Principle* may be stated as

> *The path taken between two points by a ray of light is the path that can traversed in the least time*.

So, if $T(\lambda)$ is the time taken to traverse a path that depends in some way on λ, then we require

$$\frac{dT}{d\lambda} = 0. \tag{9.4}$$

More generally, however, T will depend on the *function* that defines the path-length. Suppose we denote such a function by f. We should then express T as

$$T = T[f], \tag{9.5}$$

where the square brackets indicate that T is a *functional* of f (i.e. a quantity that depends on a function). An integral, then, is a common example of a functional.

9.4.1 Rectilinear propagation

In a homogeneous medium, it may be shown that Eq. (9.7) takes a minimum when $y(x)$ is a straight line passing through points A and B. Thus we reproduce the earlier result for rectilinear propagation found by invoking Huygens' Principle.

For a homogeneous medium we may put

$$T[f] = \Lambda[f] = n\int_0^r dS, \tag{9.6}$$

which leads to

$$\Lambda\left[y'\right] = n\int_A^B \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{1/2} dx. \tag{9.7}$$

That is, T is a functional of the derivative of y. Putting $f = y'$, the condition of Eq. (9.4) becomes

$$\frac{\delta T}{\delta f(x)} = 0, \tag{9.8}$$

where $\delta T/\delta f(x)$ is known as the *functional derivative* of $T[f]$ with respect to $f(x)$. Solving Eq. (9.8) is the task of the *calculus of variations*.

9.4.2 Calculus of variations

EulerLagrange equation

Given a functional of the form

$$T[f] = \int_{x_1}^{x_2} L\left[x, f(x), f'(x)\right]\ dx, \tag{9.9}$$

it may be shown that the functional derivative of $T[f]$ is

$$\frac{\delta T}{\delta f(x)} = \frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'}. \tag{9.10}$$

Setting the right hand side of this to zero yields the *Euler-Lagrange equation*.

The shortest optical path length between two points

For an optical path between two points (x_1, y_1) and (x_2, y_2) in the $x-y$ plane, the total length is given by

$$\Lambda[f] = n\int_{x_1}^{x_2} \left[1 + f'^2\right]^{1/2}\ dx. \tag{9.11}$$

Hence

$$L\left[x, f(x), f'(x)\right] = \left[1 + f'^2\right]^{1/2}, \tag{9.12}$$

$$\frac{\partial L}{\partial f} = 0 \tag{9.13}$$

and

$$\frac{d}{dx}\frac{\partial L}{\partial f'} = \frac{d}{dx}\frac{f'}{[1 + f'^2]^{1/2}}. \tag{9.14}$$

Setting this last result to zero and integrating this with respect to x gives, after re-arranging,

$$f'(x) = \left[\frac{A}{1-A}\right]^{1/2} = m, \tag{9.15}$$

where A is the constant of integration and thus m is a constant. Integrating with respect to x again gives

$$f(x) = mx + c, \tag{9.16}$$

which is the equation of a straight line. The value of m and c may then be found by applying the boundary conditions that the line must pass through (x_1, y_1) and (x_2, y_2).

- *Thus we have shown that for an isotropic and homogeneous medium, the shortest optical path length between two points is a straight line.*

9.4.3 Restatement of Fermat's Principle

Fermat's Principle may be re-stated as

- *Light traverses the route between two points for which the optical path length is a minimum*.

As we shall see, this may applied with certain constraints, such as the ray of light must travel via a point in the interface between two media.

9.4.4 Reflection

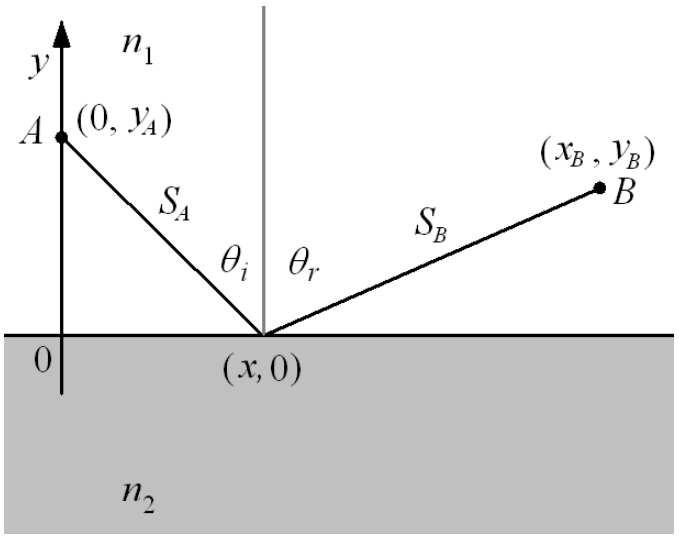


Figure 9.2: Rays between points A and B constrained to touch a point in the boundary plain between media.

Applying Fermat's Principle to the case of reflection, we constrain the minimization problem such that the path between points A and B goes via

a point in the boundary plane between media with refractive indices n_1 and n_2, as shown in Fig. 9.2. The geometric path-lengths between the points A and B and the point at $(x, 0)$ in the plane are given by

$$S_A = \sqrt{x^2 + y_A^2} \tag{9.17}$$

and

$$S_B = \sqrt{(x_B - x)^2 + y_B^2}. \tag{9.18}$$

Note that

$$x = S_A \sin\theta_i \tag{9.19}$$

and

$$x_B - x = S_B \sin\theta_r. \tag{9.20}$$

In a homogeneous medium the optical path-length is then

$$\Lambda = n_1 (S_A + S_B). \tag{9.21}$$

Differentiating this with respect to x and using Eqs. (9.19) and (9.20),

$$\frac{d\Lambda}{dx} = n_1 \left(\frac{x}{S_A} - \frac{x_B - x}{S_B} \right) = n_1 (\sin\theta_i - \sin\theta_r). \tag{9.22}$$

Hence, when $d\Lambda/dx = 0$

$$\sin\theta_i = \sin\theta_r \tag{9.23}$$

or

$$\theta_i = \theta_r, \tag{9.24}$$

reproducing the result of Sec. 4.5.2.

9.4.5 Refraction

In the case of transmitted light (see Fig. 9.3), Eqs. (9.17) to (9.20) still hold (after substituting θ_t for θ_r. The optical path-length is now

$$\Lambda = n_1 S_A + n_2 S_B \tag{9.25}$$

and

$$\frac{d\Lambda}{dx} = n_1 \frac{x}{S_A} - n_2 \frac{x_B - x}{S_B} = n_1 \sin\theta_i - n_2 \sin\theta_t. \tag{9.26}$$

When $d\Lambda/dx = 0$, we now find

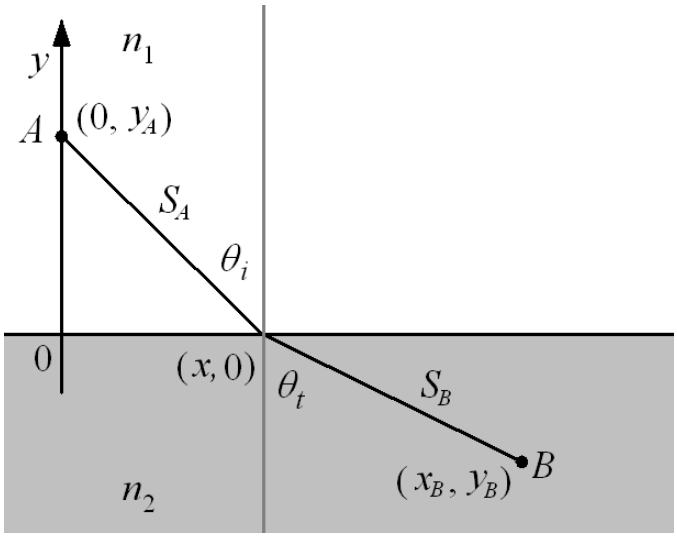


Figure 9.3: Rays between points A and B propagating as straight lines to and from a point in the boundary plain between media.

$$n_1 \sin\theta_i = n_2 \sin\theta_t, \tag{9.27}$$

i.e. Snell's Law, as found earlier.

9.5 Perfect mirrors

9.5.1 Imaging rays from a point onto a plane

Figure 9.4 shows a sketch of a mirror that perfectly images all rays from the origin onto a plane wave, for instance, the planar wavefront at $y = A$ shown. In other words, all such rays must have the same optical path-length Λ. The general form for Λ in term of $y = y(x)$ (the equation of the surface of the mirror) will then be

$$\Lambda = n\,(S + A - y) = B, \tag{9.28}$$

where B is a constant. Since all rays travel in the same refractive index n, we may put $B' = B/n$ and expand S to give

$$\left(x^2 + y^2\right) = B' - A + y = y + C, \tag{9.29}$$

where $C = B' - A$ is a constant. Squaring both sides,

$$x^2 + y^2 = y^2 + 2yC + C^2. \tag{9.30}$$

The y^2 term then cancels and, after re-arranging, we arrive at

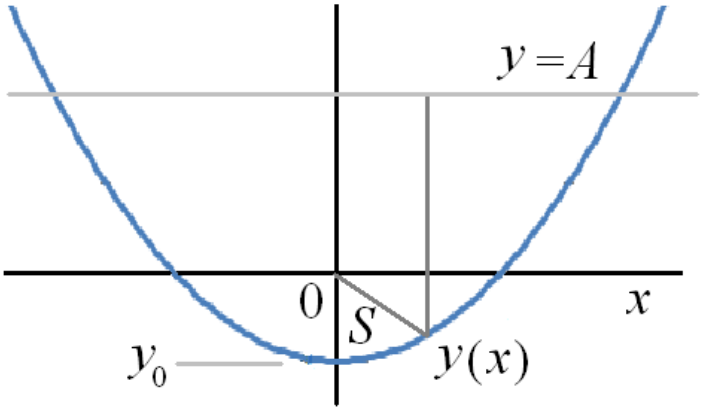


Figure 9.4: Sketch of a perfect mirror that images all rays from a point (at the origin) onto a plane wave.

$$y = \frac{x^2}{2C} - \frac{C}{2}. \tag{9.31}$$

This is the equation of a *parabola*.

Putting $y(0) = y_0$, this becomes

$$\boxed{y = -\frac{x^2}{4y_0} + y_0} \tag{9.32}$$

(note that y_0 is negative).

9.5.2 Imaging rays from a point onto another point

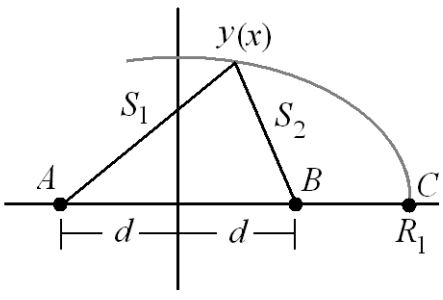


Figure 9.5: Partial sketch of a mirror that perfectly images light from the point A to the point B (and vice versa).

Consider a mirror that perfectly images light from one point onto another point, as implied by Fig. 9.5. We require the equation for the surface of a mirror that perfectly images light from one point onto another point. Hence all optical path-lengths from A to B via $y(x)$, as shown in Fig. **??**, must be equal. Label the point where the curve meets the positive x axis C and define this to be at $x = R_1$. Referring to Fig. **??**, a general optical path length is then given by

$$\Lambda = nS_1 + nS_2 = n\left([d+x]^2 + y^2\right)^{1/2} + n\left([x-d]^2 + y^2\right)^{1/2}. \quad (9.33)$$

We may then equate this with the optical path-length from A to $x = R_1/2$ and back to B along the x axis

$$\Lambda_{ACB} = n\left(d + R_1 + [R_1 - d]\right) = 2nR_1. \quad (9.34)$$

Since the refractive index is the same for all optical path-lengths, this cancels out. Hence, equating Eqs. (9.33) and (9.34) and rearranging, we have

$$\left([d+x]^2 + y^2\right)^{1/2} = 2R_1 - \left([x-d]^2 + y^2\right)^{1/2}. \quad (9.35)$$

Squaring both sides

$$d^2+2xd+x^2+y^2 = 4R_1^2 - 4R_1\left([x-d]^2 + y^2\right)^{1/2} + x^2 - 2xd + d^2 + y^2. \quad (9.36)$$

Subtracting $d^2+x^2+y^2$ from both sides and rearranging to make the square-rooted term the subject,

$$\left([x-d]^2 + y^2\right)^{1/2} = R_1 - \frac{xd}{R_1}. \quad (9.37)$$

Squaring again gives

$$x^2 - 2xd + d^2 + y^2 = R_1^2 - 2xd + \frac{x^2d^2}{R_1^2}. \quad (9.38)$$

Adding $2xd$ to both sides and moving all constant terms to the right-hand-side gives

$$x^2\left(1 - \frac{d^2}{R_1^2}\right) + y^2 = R_1^2\left(1 - \frac{d^2}{R_1^2}\right). \quad (9.39)$$

Finally, defining $R_2^2 = R_1^2\left(1 - d^2/R_1^2\right)$ and dividing through by this gives

$$\frac{x^2}{R_1^2} + \frac{y^2}{R_2^2} = 1. \quad (9.40)$$

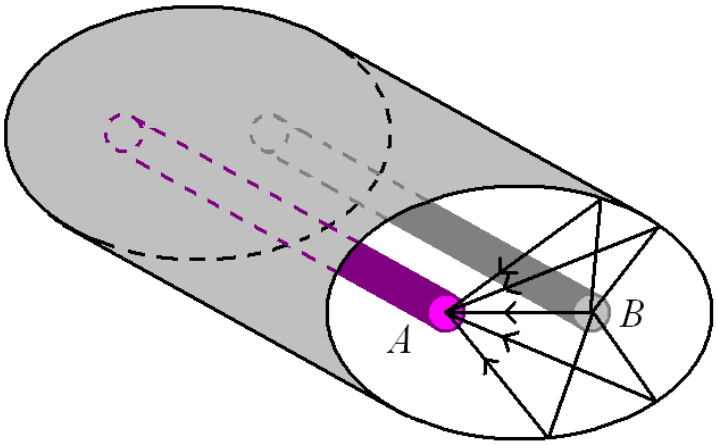


Figure 9.6: A sketch of part of a ruby laser.

This is the equation of an *ellipse*.

An example application of an elliptical mirror is illustrated in Fig. 9.6, which shows a sketch of part of a ruby laser. This consists of a tube of elliptical cross-section with a highly reflective inner coating. A ruby rod is positioned at A aligned with one of the foci of the elliptical mirror all along the tube (the foci are the points that perfectly image onto one another). A flash lamp is similarly aligned with the other focus at B. Hence, all the light emitted from the flash lamp is focused onto the ruby rod, optically pumping the electrons to excited energy states (creating a 'population inversion') as a necessary condition for lasing.

9.6 Perfect lenses

Figure 9.7 shows a sketch of a lens that perfectly images rays from an external point at A onto a plane-wave within the lens. We therefore seek the equation of the surface of the lens that gives the same optical path-length for all such rays.

The optical path-length for the ray propagating from A along the x-axis to a wave-front at $x = C$ is

$$\Lambda_{A0C} = n_1 d_1 + n_2 C. \tag{9.41}$$

For the more general ray shown, we have

$$\Lambda_{ABC} = n_1 S_1 + n_2 \left(C - x\right), \tag{9.42}$$

where

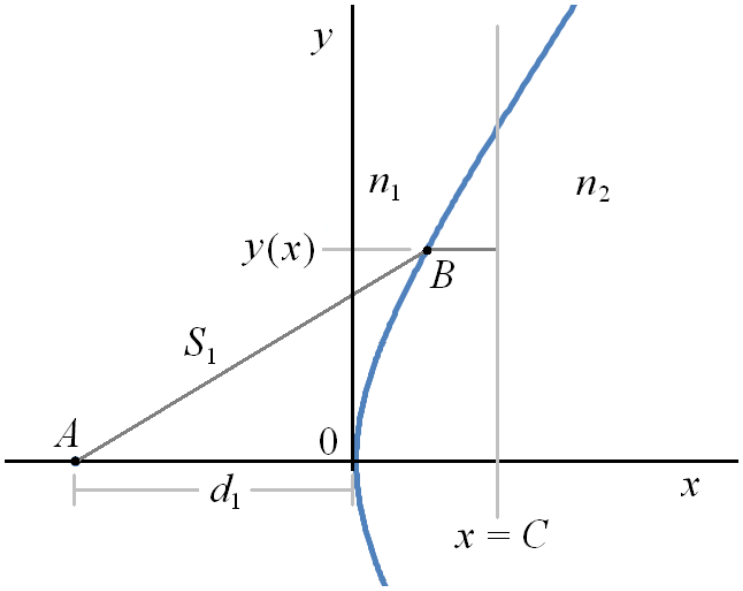


Figure 9.7: Sketch of a perfect lens that images rays from an external point at A onto a plane-wave within the lens.

$$S_1 = \left([x + d_1]^2 + y^2\right)^{1/2}. \tag{9.43}$$

We therefore require

$$\Lambda_{ABC} = \Lambda_{A0C} \tag{9.44}$$

Substituting for both sides of this equation from Eqs. (9.41) and (9.42), we have

$$n_1 \left([x + d_1]^2 + y^2\right)^{1/2} + n_2 (C - x) = n_1 d_1 + n_2 C. \tag{9.45}$$

Now $n_2 C$ cancels from both sides, so after re-arranging we have

$$n_1 \left([x + d_1]^2 + y^2\right)^{1/2} = n_1 d_1 + n_2 x. \tag{9.46}$$

Squaring both sides gives us

$$n_1^2 \left(x^2 + 2x d_1 + d_1^2 + y^2\right) = n_1^2 d_1^2 + 2 n_1 n_2 x d_1 + n_2^2 x^2 \tag{9.47}$$

The $n_1^2 d_1^2$ terms cancel out and we re-arrange to obtain

$$\left(n_2^2 - n_1^2\right) x^2 + 2 n_1 d_1 \left(n_2 - n_1\right) x - n_1^2 y^2 = 0. \tag{9.48}$$

This is the equation of a *hyperbola*. However, it is not in standard form.

The standard form for a hyperbola is

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{9.49}$$

We therefore need to translate Eq. (9.48) to remove the x terms.

Putting $x \to x - \Delta x$, Eq. (9.48) becomes

$$\begin{aligned} \left(n_2^2 - n_1^2\right)\left(x^2 - 2x\Delta x + \Delta x^2\right) &+ 2n_1 d_1 \left(n_2 - n_1\right) x \\ &- 2n_1 d_1 \left(n_2 - n_1\right) \Delta x - n_1^2 y^2 = 0. \end{aligned} \tag{9.50}$$

To eliminate the terms involving x, we therefore require

$$2n_1 d_1 \left(n_2 - n_1\right) = 2\left(n_2^2 - n_1^2\right) \Delta x, \tag{9.51}$$

which gives

$$\Delta x = \frac{n_1 d_1}{n_2 + n_1}. \tag{9.52}$$

Substituting this back into Eq. (9.50) and moving the constant terms to the right-hand-side gives

$$\left(n_2^2 - n_1^2\right) x^2 - n_1^2 y^2 = n_1^2 d_1^2 \frac{n_2 - n_1}{n_2 + n_1}. \tag{9.53}$$

Finally, dividing through by the constant term on the right-hand-side gives

$$\boxed{\frac{\left(n_2 + n_1\right)^2}{n_1^2 d_1^2} x^2 - \frac{\left(n_2 + n_1\right)}{\left(n_2 - n_1\right) d_1^2} y^2 = 1.} \tag{9.54}$$

This is then the equation of the surface of the lens in standard form.

9.7 Curvature

An infinitesimal curve segment dS may be related to the angle ϕ the gradient of the curve at that point makes with x-axis (i.e. $\tan\phi = dy/dx$). This is illustrated in Fig. 9.8. The *curvature* of a surface at a given point may then be defined as

$$\boxed{\kappa \equiv \frac{d\phi}{dS} \equiv \frac{1}{R},} \tag{9.55}$$

where R is the *radius of curvature*.

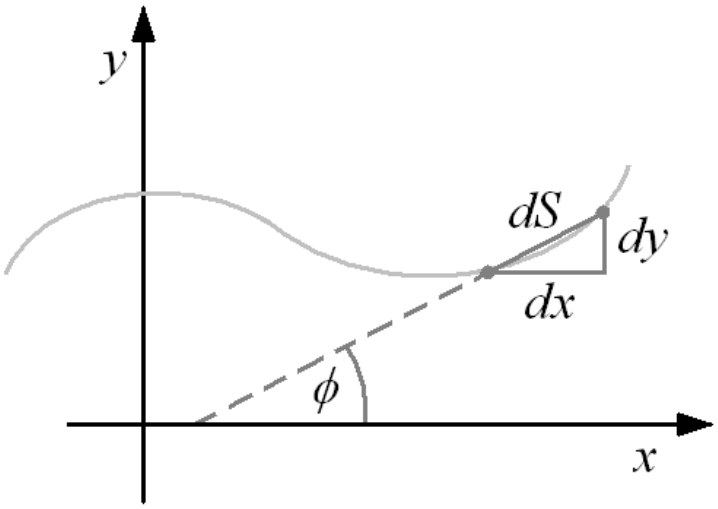


Figure 9.8: Sketch of a general curve illustrating the idea of curvature as the rate of change of the infinitesimal arc-length dS with respect to ϕ.

9.7.1 Curvature of a circle

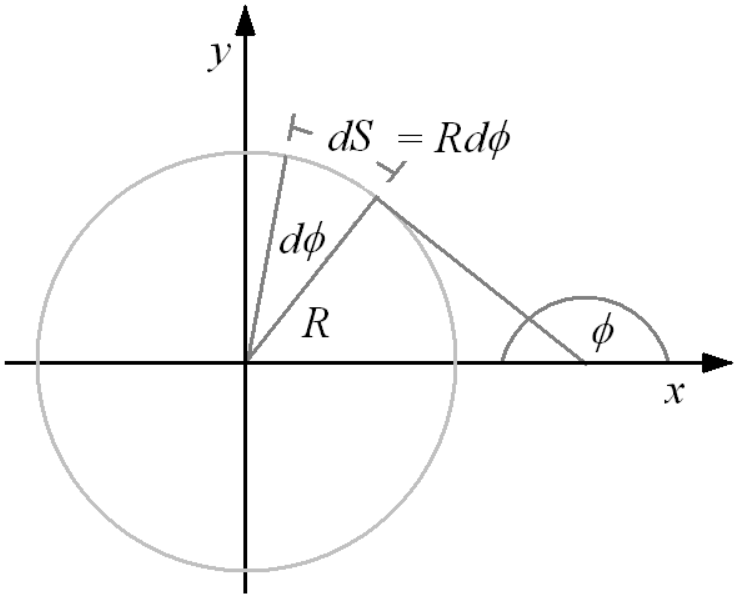


Figure 9.9: Sketch illustrating the curvature of a circle in terms of an infinitesimal arc-length and the angle that it subtends.

In the case of a circle of radius R (see Fig. 9.9), $d\phi$ is just the angle

subtended by an infinitesimal arc-length $dS = Rd\phi$. The is then most easily found using cylindrical polar coordinates

$$\kappa = \lim_{dS\to 0} \frac{d\phi}{dS} = \lim_{dS\to 0} \frac{d\phi}{Rd\phi} = \frac{1}{R}. \tag{9.56}$$

So the curvature of a circle of radius R is just $1/R$, as we might expect.

9.8 Parameterisation of a curve

In Fig. 9.10, we see that the curvature of a general curve $y(x)$ may be matched at a given point (x, y) to that of a circle with radius R. Hence, the curvature of $y(x)$ at (x, y) is $1/R$.

In practice, calculation of the curvature may be achieved by parameterising a curve. For a curve lying in the $x - y$ plane, x and y may be given in terms of a parameter t

$$x = x(t), \quad y = y(t). \tag{9.57}$$

The derivatives with respect to t are then denoted by

$$\frac{dx}{dt} = x', \quad \frac{dx}{dt} = y' \tag{9.58}$$

and similarly for higher derivatives. Using the chain rule, the curvature is then given by

$$\kappa = \frac{d\phi}{dS} = \frac{d\phi}{dt}\frac{dt}{dS}. \tag{9.59}$$

First, we need to find an expression for $d\phi/dt$. Referring to Fig. 9.8, we see that

$$\tan\phi = \frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = \frac{y'}{x'}, \tag{9.60}$$

so

$$\phi = \tan^{-1}\left(\frac{y'}{x'}\right). \tag{9.61}$$

Recalling that

$$\frac{d\tan^{-1}\xi}{d\xi} = \left(1 + \xi^2\right)^{-1} \tag{9.62}$$

and using the chain rule, we have

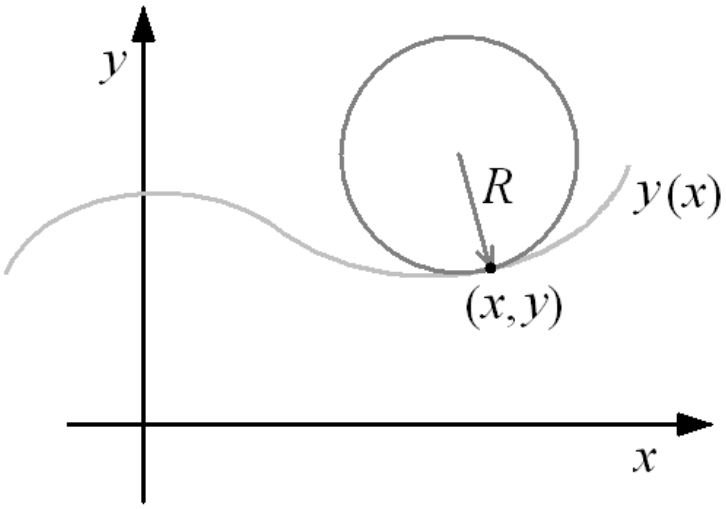


Figure 9.10: Sketch of a general curve showing how the radius of curvature at a given point is related to the radius of a circle sharing a common tangent.

$$\begin{aligned}\frac{d\phi}{dt} &= \frac{d\left(y'/x'\right)}{dt}\left(1+\left[\frac{y'}{x'}\right]^2\right)^{-1} \\ &= \left(\frac{y''}{x'}-\frac{x''y'}{x'^2}\right)\left(1+\left[\frac{y'}{x'}\right]^2\right)^{-1} \\ &= \frac{x'y''-x''y'}{x'^2+y'^2}. \end{aligned} \tag{9.63}$$

Meanwhile, we have

$$dS = \left(dx^2+dy^2\right)^{1/2}, \tag{9.64}$$

so

$$\frac{dS}{dt} = \left(\left[\frac{dx}{dt}\right]^2+\left[\frac{dy}{dt}\right]^2\right)^{1/2} = \left(x'^2+y'^2\right)^{1/2}. \tag{9.65}$$

Hence, combining Eqs. (9.63) and (9.65) according to Eq. (9.59), we arrive at the general expression for the curvature

$$\boxed{\kappa = \frac{x'y''-x''y'}{\left(x'^2+y'^2\right)^{3/2}}.} \tag{9.66}$$

9.8.1 Parameterisation of a circle

The general equation for a circle of radius R centred at (x_C, y_C) is

$$\frac{(x-x_C)^2}{R^2} + \frac{(y-y_C)^2}{R^2} = 1. \tag{9.67}$$

Since

$$\cos^2 t + \sin^2 t = 1, \tag{9.68}$$

we may parameterise Eq. (9.67) by putting

$$\begin{aligned} x(t) &= R\cos t + x_C, \\ y(t) &= R\sin t + y_C \end{aligned} \tag{9.69}$$

giving

$$\begin{aligned} x' &= -R\sin t, \quad y' = R\cos t, \\ x'' &= -R\cos t, \quad y'' = -R\sin t. \end{aligned} \tag{9.70}$$

Using Eq. (9.66) then gives

$$\begin{aligned} \kappa &= \frac{R^2\left(\sin^2 t + \cos^2 t\right)}{R^3\left(\sin^2 t + \cos^2 t\right)^{3/2}} \\ &= \frac{1}{R}, \end{aligned} \tag{9.71}$$

as found earlier.

9.8.2 Parameterisation of a parabola

The equation of a parabola is

$$y = ax^2 + bx + c. \tag{9.72}$$

Hence, a straight-forward parameterisation is

$$\begin{aligned} x &= t, \quad y = at^2 + bt + c, \\ x' &= 1, \quad y' = 2at + b, \\ x'' &= 0, \quad y'' = 2a. \end{aligned} \tag{9.73}$$

Using Eq. (9.66), we have

$$\kappa = \frac{y''}{(x'^2 + y'^2)^{3/2}} = \frac{2a}{\left(1 + [2at+b]^2\right)^{3/2}}. \tag{9.74}$$

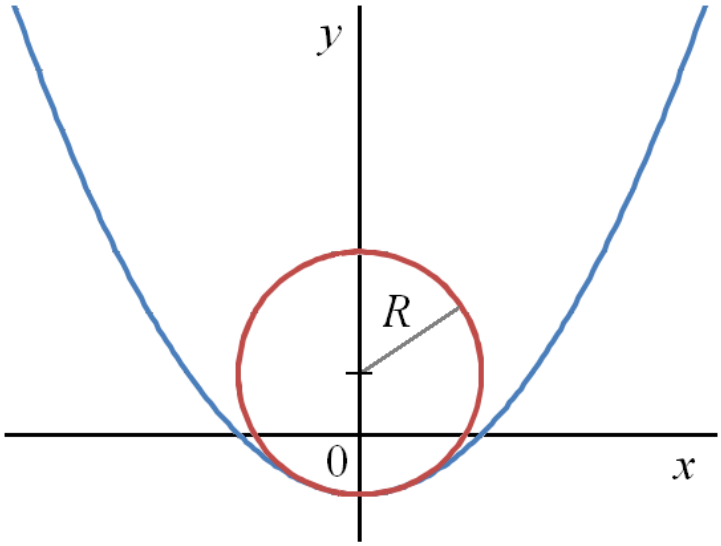


Figure 9.11: Graph of a parabola with a circle of radius R, equal to the radius of curvature of the parabola at $x = 0$ imposed over the top.

At the minimum of the curve, we have $y' = 0$, so

$$\kappa = y'' = 2a = \frac{1}{R}, \tag{9.75}$$

giving

$$R = \frac{1}{2a} \tag{9.76}$$

for the radius of curvature. For the equation of the parabolic mirror we derived earlier, we found

$$a = -\frac{1}{4y_0} \tag{9.77}$$

(recall that y_0 is negative), so the radius of curvature of the mirror is

$$R = -2y_0. \tag{9.78}$$

Figure 9.11 shows a graph of a parabola with a circle of radius R, equal to the radius of curvature of the parabola at $x = 0$ imposed over the top. As can be seen from the figure, the circle matches the parabola very well around the minimum point. This indicates that a spherical mirror would provide a good approximation to a parabolic mirror at its minimum point.

9.8.3 Parameterisation of a hyperbola

Since the standard form for a hyperbola is given by Eq. (9.49), we may use the following parameterisation:

$$\begin{array}{rclrcl} x &=& a\cosh t, & y &=& b\sinh t, \\ x' &=& a\sinh t, & y' &=& b\cosh t, \\ x'' &=& a\cosh t, & y'' &=& b\sinh t. \end{array} \tag{9.79}$$

Substituting these expressions into Eq. (9.66) gives the curvature as

$$\kappa = -\frac{ab}{\left(a^2\sinh^2 t + b^2\cosh^2 t\right)^{3/2}}. \tag{9.80}$$

When $t = 0$, we have

$$x = a, \quad y = 0, \tag{9.81}$$

and the magnitude of the curvature is (using our earlier results)

$$|\kappa| = \frac{a}{b^2} = \frac{n_1}{(n_2 - n_1)\,d_1}. \tag{9.82}$$

Hence, the radius of curvature is

$$R = \frac{(n_2 - n_1)\,d_1}{n_1}. \tag{9.83}$$

Figure 9.12 shows the graph of a hyperbola with a circle of radius R, equal to the radius of curvature of the hyperbola at $y = 0$ imposed over the top. Again, we see that the hyperbolic lens will be well approximated by a circular lens around $y = 0$.

9.8.4 Parameterisation of an elllipse

Equation (9.40) may be parameterised via

$$x(t) = R_1\cos t \tag{9.84}$$

and

$$y(t) = R_2\sin t \tag{9.85}$$

Using these results, the derivatives with respect to t are

$$x' = -R_1\sin t, \quad y' = R_2\cos t \tag{9.86}$$

and

$$x'' = -R_1\cos t, \quad y'' = -R_2\sin t \tag{9.87}$$

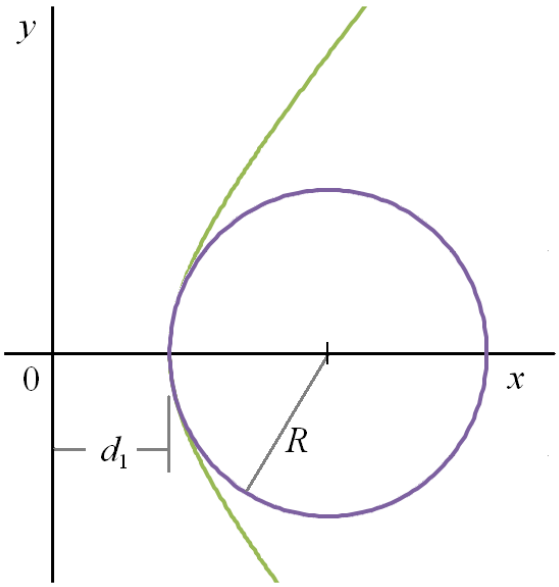


Figure 9.12: Graph of a hyperbola with a circle of radius R, equal to the radius of curvature of the hyperbola at $y = 0$ imposed over the top.

Hence, using Eq. (9.66), we have

$$\kappa = \frac{R_1 R_2 \left(\sin^2 t + \cos^2 t\right)}{\left(R_1^2 \sin^2 t + R_2^2 \cos^2 t\right)^{3/2}} = \frac{R_1 R_2}{\left(R_1^2 \sin^2 t + R_2^2 \cos^2 t\right)^{3/2}}, \tag{9.88}$$

As special cases of this result, we first recognise that $y = 0$ when $t = 0$. The curvature is then

$$\kappa = \frac{R_1 R_2}{\left(R_2^2\right)^{3/2}} = \frac{R_1}{R_2^2} \tag{9.89}$$

and the radius of curvature is

$$R = \frac{R_2^2}{R_1}. \tag{9.90}$$

As another particular case, $x = 0$ when $t = \pi/2$, so the curvature is

$$\kappa = \frac{R_1 R_2}{\left(R_1^2\right)^{3/2}} = \frac{R_2}{R_1^2}, \tag{9.91}$$

and the radius of curvature is

$$R = \frac{R_1^2}{R_2}. \tag{9.92}$$

9.9 Summary

- **Fermat's Principle**
  Light traverses the route between two points for which the optical path length is a minimum.

- **The laws of geometric optics**
  Fermats' Principles may be applied to derive:

  - *The law of rectilinear propagation*
    In a homogeneous medium, light travels in straight lines.
  - *The law of reflection*
    In a homogeneous incident medium, the angle of incidence equals the angle of reflection.
  - *The law of refraction*
    When light passes from a homogeneous medium with refractive index n_i into another homogeneous medium with refractive index n_t , the angles of incidence and refraction, θ_i and θ_t, are given by Snell's Law

$$\boxed{n_i \sin\theta_i = n_t \sin\theta_t.}$$

- **Perfect imaging**
  Imaging from a point or plane to a point or plane such that all rays have the same optical path-length.

- **Perfect mirrors**

  - *Parabolic*
    A parabolic mirror perfectly images rays from a point (at the focus of the parabola) onto a plane (and vice versa).
  - *Elliptical*
    An elliptical mirror perfectly images rays from a point (at one focus of the ellipse) onto another point (at the other focus).

- **Perfect lenses**

  - *Hyperbolic*
    A hyperbolic lens perfectly images rays from a particular point outside the lens onto a plane inside the lens.

- **Curvature**
  A curve may be characterised by its curvature at a given point

$$\boxed{\kappa \equiv \frac{d\phi}{dS} \equiv \frac{1}{R},} \tag{9.93}$$

where R is the radius of curvature.

- **Spherical approximation**
  Using the concept of curvature, it has been shown that spherical mirrors and lenses may be used to approximate perfect mirrors and lenses.

10. Spherical Lenses and Mirrors

10.1 General remarks

Having looked at the special cases of perfect imaging, we now turn our attention to the case of *spherical lenses and mirrors*. We saw, in the previous chapter, that mirrors and lenses with conic section profiles may be approximated by a circle or sphere over a small angle, when we match the curvatures. This approximation means that in the *paraxial* approximation of small angles we may get very *nearly* perfect imaging. As the paraxial approximation begins to fail, we start to encounter *spherical aberrations*, due to the fact that a sphere cannot image perfectly.

First, we shall take a look at spherical lenses generally, before applying the paraxial and 'thin' lens approximations to obtain the *thin lens equation*. This introduces the concept of the *focal length*. We follow this with an analysis of thin lenses in combinations.

We then consider spherical mirrors, where a similar treatment to that used for lenses is employed for finding the analogous equations and focal lengths. Then, using the methods so far developed, the rules of image construction for convex and concave lenses are described.

10.2 Learning objectives

The aims of this section are to understand and be able to apply

- Analysis of spherical lenses
- Analysis of 'thin' lenses
  - Thin lens equation
  - Lens maker's formula
  - Gaussian lens formula
- Lens and mirror sign conventions
- Thin lenses in combination.
- Analysis of spherical mirrors

- Image construction
- Magnification
- Monochromatic aberration
  - Third order aberration
  - Spherical aberration
  - Coma
  - Astigmatism
  - Field curvature
  - Distortion

10.3 Spherical lenses

10.3.1 Lens sign conventions

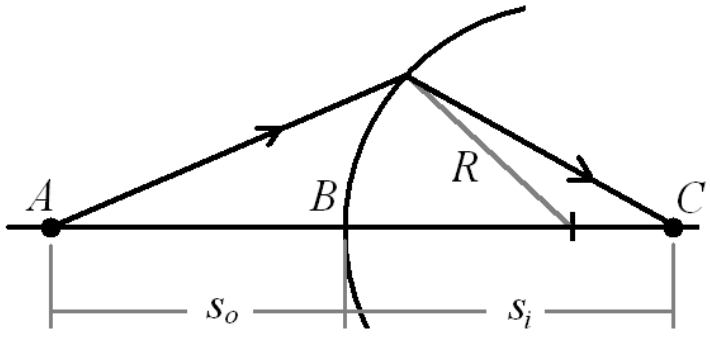


Figure 10.1: Guide for lens conventions.

With reference to Fig. 10.1, the conventions for lens calculations are taken to be

1. Light is always taken to propagate from left to right.
2. If A is to the left of B, then s_o is taken to be positive (and vice versa).
3. If C is to the right of B, then s_i is taken to be positive (and vice versa).
4. If the centre of the sphere is to the right of B, R is taken to be positive. This is a *convex* lens. If the centre of the sphere is to the left of B, R is taken to be negative. This is a *concave* lens.

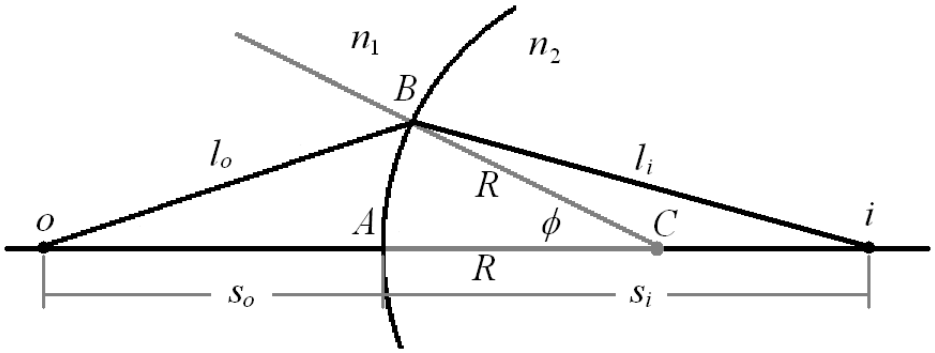


Figure 10.2: Sketch of a ray of light propagating from a point at o to a point i within a spherical lens.

Figure 10.2 shows a ray of light propagating from a point at o in a medium of refractive index n_1 to a point i within a spherical lens of refractive index n_2 via the point B on the lens surface. The geometric distances $l_o = oB$ and $l_i = Bi$ are given using the cosine rule by

$$l_o^2 = (s_o + R)^2 + R^2 - 2\,(s_o + R)\,R\cos\phi \tag{10.1}$$

and

$$l_i^2 = (s_i - R)^2 + R^2 + 2\,(s_i - R)\,R\cos\phi, \tag{10.2}$$

whilst the optical path-length for the ray is

$$\Lambda = n_1 l_0 + n_2 l_i. \tag{10.3}$$

Although we do not expect a spherical lens to image perfectly, reasonable imaging may be possible to a good approximation. The requirement

for *perfect* imaging is that all the rays have equal optical path-length. That is, we require Λ to be a constant

$$\frac{d\Lambda}{d\phi} = 0. \tag{10.4}$$

Applying this to Eqs. (10.1) to (10.2),

$$\frac{d\Lambda}{d\phi} = n_1 \frac{dl_0}{d\phi} + n_2 \frac{dl_i}{d\phi} = 0, \tag{10.5}$$

where

$$\begin{aligned} \frac{dl_0}{d\phi} &= \frac{(s_o + R)\, R \sin\phi}{\left([s_o + R]^2 + R^2 - 2\,[s_o + R]\, R\cos\phi \right)^{1/2}} \\ &= \frac{(s_o + R)\, R \sin\phi}{l_0} \end{aligned} \tag{10.6}$$

and, similarly,

$$\frac{dl_i}{d\phi} = -\frac{(s_i - R)\, R \sin\phi}{l_i}. \tag{10.7}$$

Inserting these results into Eq. (10.5),

$$n_1 \frac{(s_o + R)}{l_0} = n_2 \frac{(s_i - R)}{l_i}, \tag{10.8}$$

which may be re-arranged to give

$$\left(\frac{n_1}{l_0} + \frac{n_2}{l_i} \right) R = \frac{n_2 s_i}{l_i} - \frac{n_1 s_o}{l_0}. \tag{10.9}$$

10.3.2 The paraxial approximation

It is important to note that l_o and l_i appearing in Eq. (10.9) are still functions of ϕ and the associated distances s_0 or s_i. It is therefore not possible to determine either s_0 or s_i analytically. However, we may make progress by employing the *paraxial approximation*. The is a small angle approximation in which the sinusoidal functions are taken to be approximately equal to the first terms of their Taylor series expansions. Hence, for small ϕ, we may put

$$\sin\phi \approx \phi \tag{10.10}$$

and

$$\cos\phi \approx 1. \tag{10.11}$$

With these approximations, we find

$$l_o^2 \to (s_o + R)^2 + R^2 - 2\,(s_o + R)\,R = s_0^2 \tag{10.12}$$

and

$$l_i^2 \to (s_i - R)^2 + R^2 + 2\,(s_i - R) = s_i^2. \tag{10.13}$$

Hence, Eq. (10.9) becomes

$$\boxed{\left(\frac{n_1}{s_0} + \frac{n_2}{s_i}\right) = \frac{1}{R}\,(n_2 - n_1)\,.} \tag{10.14}$$

This is the general equation for a spherical lens surface (note that we could have obtained the same result by applying Snell's law directly to this problem).

Let us now consider the case when $s_o \to \infty$. Equation (10.14) then becomes

$$\frac{n_2}{s_i} = \frac{1}{R}\,(n_2 - n_1)\,, \tag{10.15}$$

from which we find

$$\boxed{s_i = \frac{n_2}{n_2 - n_1} R \equiv f_i,} \tag{10.16}$$

where we have defined the *focal length* inside the lens f_i. Similarly, when $s_i \to \infty$, we get

$$\boxed{s_o = \frac{n_1}{n_2 - n_1} R \equiv f_o,} \tag{10.17}$$

where f_o is defined as the *focal length* outside the lens.

10.3.3 Special cases

Figure 10.3 illustrates some special cases of Eq. (10.14).

Case: $s_i < 0$ In this case n_2/n_1 is not large enough to refract the transmitted ray below the horizontal and s_i marks a virtual image to the left of the lens.

Case: $R \to \infty$ In this case, the curvature of the lens becomes zero, i.e. the lens surface becomes a flat plane, and we have

$$\frac{n_1}{s_o} + \frac{n_2}{s_i} = 0, \tag{10.18}$$

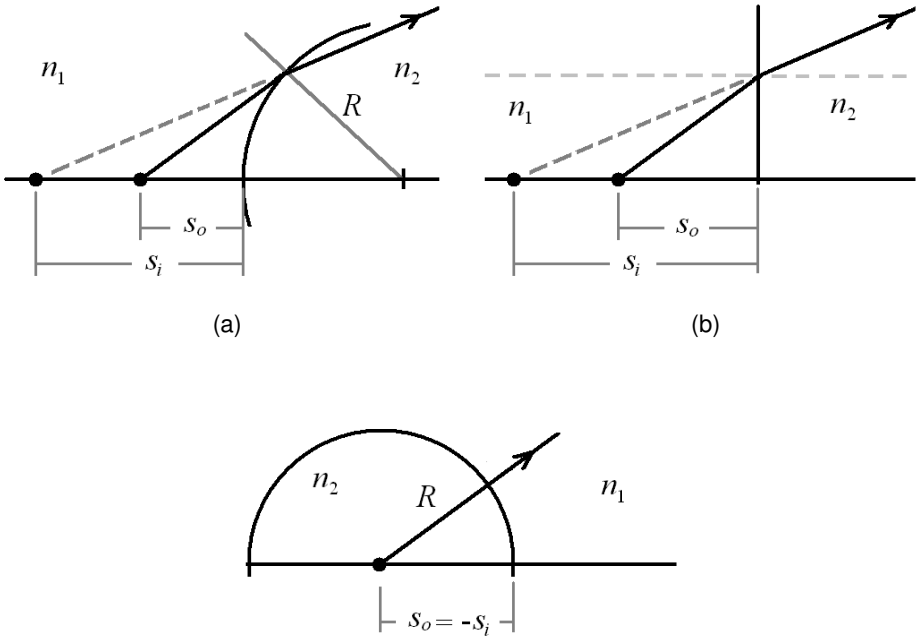


Figure 10.3: Special cases of Eq. (10.14): (a) $s_i < 0$. In this case n_2/n_1 is not large enough to refract the transmitted ray below the horizontal and s_i marks a virtual image to the left of the lens. (b) When $R \to \infty$, the curvature of the lens becomes zero (the lens surface becomes a flat plane). (c) $s_o = s_i$. This can only be the case when the source of the ray is inside the lens and $|s_o| = |R|$.

giving

$$s_i = -\frac{n_2}{n_1} s_o, \tag{10.19}$$

showing that s_i is once again negative.

Case: $s_o = s_i$ This can only be the case when the source of the ray is inside the lens. Equation 10.3 then becomes

$$\frac{n_1}{s_o} - \frac{n_2}{s_0} = \frac{1}{R}\left(n_2 - n_1\right), \tag{10.20}$$

so

$$R = -s_o. \tag{10.21}$$

Note that R is now negative.

10.4 Thin lenses

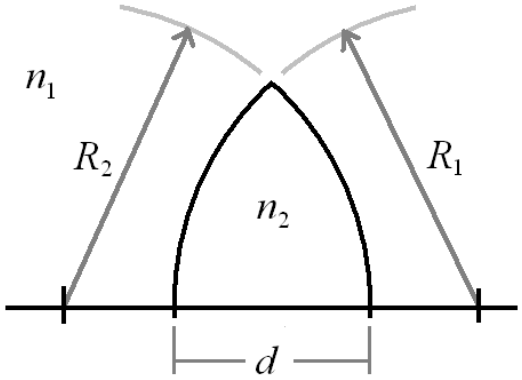


Figure 10.4: A 'thick' lens of thickness d and spherical surfaces with radii of curvature R_1 and R_2.

Figure 10.4 shows a 'thick' lens of thickness d and spherical surfaces with radii of curvature R_1 and R_2. We require an approximate model for which we can let $d \to 0$, in other words, a 'thin' lens. To achieve this, we begin by applying Eq. (10.14) to each surface in turn. For the surface with radius R_1, we have

$$\left(\frac{n_1}{s_{01}} + \frac{n_2}{s_{i1}}\right) = \frac{1}{R_1}\left(n_2 - n_1\right), \tag{10.22}$$

where s_{01} and s_{i1} are shown in Fig. 10.5. Note that s_{i1} is the distance to a virtual image and is negative.

We obtain a similar result for R_2, except that s_{i2} is now in the medium with refractive index n_1 whilst s_{o2} is in the medium with refractive index n_2. Thus, we have

$$\left(\frac{n_1}{s_{i2}} + \frac{n_2}{s_{o2}}\right) = -\frac{1}{R_2}\left(n_2 - n_1\right). \tag{10.23}$$

Combining Eqs. (10.22) and (10.23), we have

$$n_1\left(\frac{1}{s_{i2}} + \frac{1}{s_{o1}}\right) = \left(\frac{1}{R_1} - \frac{1}{R_2}\right)\left(n_2 - n_1\right) - n_2\left(\frac{1}{s_{i1}} + \frac{1}{s_{o2}}\right). \tag{10.24}$$

Inserting $s_{o2} = d - s_{i1}$, this becomes

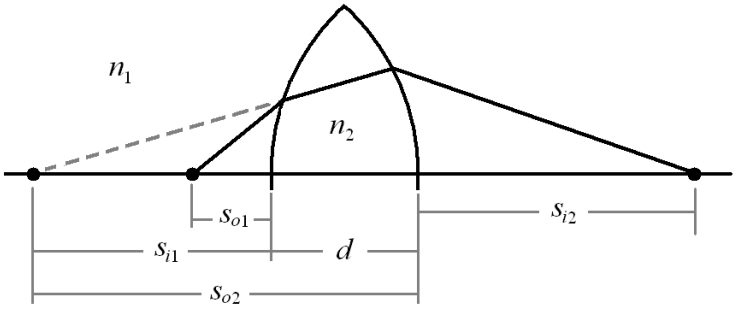


Figure 10.5: The lens of Fig 10.4 showing the imaging distances associated with each surface.

$$n_1\left(\frac{1}{s_{i2}}+\frac{1}{s_{o1}}\right)=\left(\frac{1}{R_1}-\frac{1}{R_2}\right)(n_2-n_1)-\frac{n_2 d}{s_{i1}\,(d-s_{i1})}. \tag{10.25}$$

Finally, taking the limit $d \to 0$ and putting $s_{i2} = s_i$ and $s_{o1} = s_o$, we obtain

$$\boxed{\frac{1}{s_o}+\frac{1}{s_i}=\frac{n_2-n_1}{n_1}\left(\frac{1}{R_1}-\frac{1}{R_2}\right)}. \tag{10.26}$$

This is the *thin lens equation*.

If either $s_o \to \infty$ or $s_i \to \infty$, then, since the right-hand-side of Eq. (10.80) is a constant, the resulting term on the left-hand-side must also be a constant and we may put

$$\boxed{\frac{1}{f}=\frac{n_2-n_1}{n_1}\left(\frac{1}{R_1}-\frac{1}{R_2}\right)}. \tag{10.27}$$

This is known as the *lens maker's formula*. From this it follows that

$$\boxed{\frac{1}{s_o}+\frac{1}{s_i}=\frac{1}{f},} \tag{10.28}$$

which is the *Gaussian lens formula*.

10.5 Combinations of thin lenses

Consider two thin lenses separated by a distance d. Our objective is to obtain find expressions for the effective focal lengths of this combination. We shall refer to these as the *front* and *back focal lengths*, f_f and f_b respectively.

Using the Gaussian lens formula for the first lens, we have

$$\frac{1}{s_{o1}} + \frac{1}{s_{i1}} = \frac{1}{f_1}, \tag{10.29}$$

from which we obtain

$$s_{i1} = \frac{s_{o1} f_1}{s_{o1} - f_1}. \tag{10.30}$$

Similarly, for the second lens, we have

$$s_{i2} = \frac{s_{o2} f_2}{s_{o2} - f_2}. \tag{10.31}$$

Now $s_{o2} = d - s_{i1}$, so

$$s_{i2} = \frac{(d - s_{i1})\, f_2}{d - s_{i1} - f_2}. \tag{10.32}$$

Substituting for s_{i1} from Eq. (10.30), we obtain

$$s_{i2} = \frac{f_2\,(s_{o1}\,[d - f_1] - f_1 d)}{s_{o1}\,(d - f_1 - f_2) - f_1\,(d - f_2)}. \tag{10.33}$$

If we now take the limit of $s_{o1} \to \infty$, s_{i2} becomes the back focal length f_b. We therefore have

$$f_b = \frac{f_2\,(d - f_1)}{(d - f_1 - f_2)}. \tag{10.34}$$

Similarly, s_{o1} becomes f_f as $s_{i2} \to \infty$. This will be the case when the denominator on the right-hand-side of Eq. (10.33) tends to zero. Hence, we have

$$f_f = \frac{f_1\,(d - f_2)}{d - f_1 - f_2}. \tag{10.35}$$

10.5.1 Thin lenses in close combination

If we let the distance between the lenses tend to zero, we have

$$f_b = \frac{f_2 f_1}{f_1 + f_2} = f_f, \tag{10.36}$$

Putting $f_b = f$, this may be re-written as

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}. \tag{10.37}$$

This result may also be generalised to a system of thin lenses in close combination

$$\frac{1}{f} = \sum_i \frac{1}{f_i}. \tag{10.38}$$

10.6 Spherical mirrors

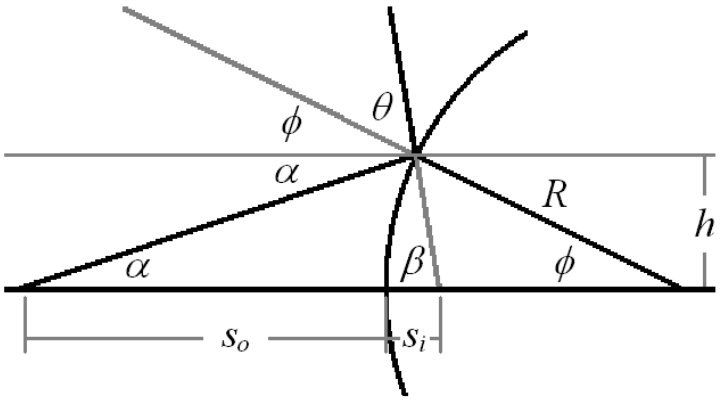


Figure 10.6: A ray of light reflecting off the surface of a spherical mirror.

10.6.1 Mirror sign conventions

With reference to Fig. 10.6, the conventions for mirror calculations are taken to be

1. Light is always taken to propagate from left to right.
2. The object distance s_o is positive when it is to the left of the mirror surface.
3. The image distance s_i is positive when it is to the left of the mirror surface (real image).
4. The image distance s_i is negative when it is to the right of the mirror surface (virtual image).
5. The radius R is positive if the mirror surface is to the right of the centre of the sphere (convex mirror)
6. The radius R is negative if the mirror surface is to the left of the centre of the sphere (concave mirror)

Figure 10.6 shows a ray of light reflecting off the surface of a spherical mirror. Note that, since s_i is the left of the centre of curvature of the mirror, by convention it is taken to be negative. θ is the angle of reflection, so must be equal to the angle of incidence. Hence, from Fig. 10.6,

$$\theta = \phi + \alpha. \tag{10.39}$$

The reflected ray makes an angle β with the horizontal. The incident ray makes an angle α with the horizontal and meets the reflected ray at an angle of 2θ. Thus, from the figure, we must have

$$\pi - 2\theta = \pi - (\alpha + \beta)\,, \tag{10.40}$$

which gives

$$2\theta = \alpha + \beta. \tag{10.41}$$

Multiplying Eq. (10.39) by 2 and subtracting Eq. (10.41), we have

$$0 = 2\phi + \alpha - \beta. \tag{10.42}$$

Hence

$$\alpha - \beta = -2\phi. \tag{10.43}$$

From Fig. 10.6, we see that $h = R\sin\phi$. For small ϕ we have $\phi \approx \sin\phi$, and hence

$$\phi \approx \frac{h}{R}. \tag{10.44}$$

At the same time, the length of the chord subtended by ϕ approaches h and becomes normal to the horizontal. Hence

$$\alpha \approx \frac{h}{s_o} \tag{10.45}$$

and, noting that s_i is negative by convention,

$$\beta \approx -\frac{h}{s_i}. \tag{10.46}$$

Hence, substituting these expressions into Eq. (10.43) gives

$$\boxed{\frac{1}{s_o} + \frac{1}{s_i} = -\frac{2}{R}.} \tag{10.47}$$

As $s_0 \to \infty$ we may put $1/s_i = 1/f$, giving

$$\frac{1}{f} = -\frac{2}{R}, \tag{10.48}$$

where f lies to the right of the mirror surface. Hence, f is negative. Substituting Eq. (10.48) into Eq. (10.47) then gives

$$\boxed{\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}.} \tag{10.49}$$

This is the same expression as the Gaussian lens formula for a spherical lens.

10.7 Image construction

For a spherical lens, the following general guide for image construction should be applied (all rays propagate from left to right)

1. Sketch a ray from the tip of the object parallel to the horizontal (the principle axis) to the centre line of the lens. From there, sketch another ray passing through the focus associated with the left-hand lens surface.
2. Sketch a ray from the tip of the object directly through the centre of the lens without deviation.
3. Sketch a ray from the tip of the object passing through the focus associated with the right-hand lens surface to the centre line of the lens. From there sketch a line parallel to the principle axis towards the image.

The tip of the image will then be the intersection of the rays described above.

10.7.1 Convex lens

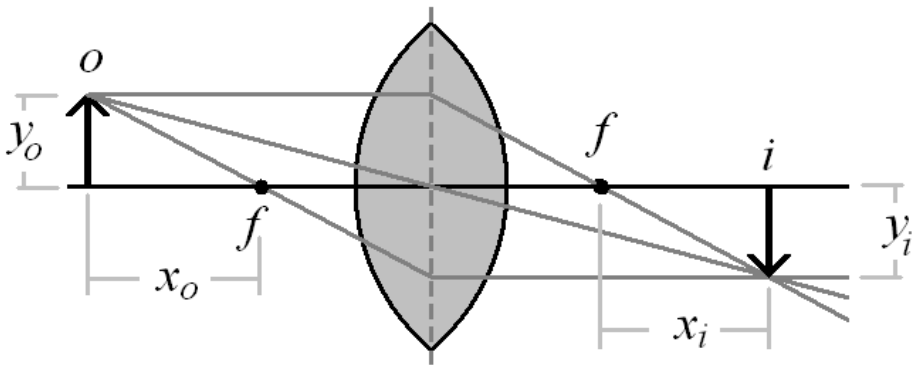


Figure 10.7: Image formation for a convex lens ($f > 0$).

Figure 10.7 shows the image construction for a convex lens ($f > 0$) employing the ray-tracing rules described above. The at i is real and inverted. Using the notation for the distances marked in Fig 10.7, the magnification M_T of this lens is given by

$$M_T = \frac{y_i}{y_o}. \tag{10.50}$$

Note, since y_i is negative, so is M_T (the image is inverted). On the right-hand-side of the lens, we find similar triangles giving the relation

$$\frac{y_o}{f} = -\frac{y_i}{x_i}. \tag{10.51}$$

Hence, in terms of x_i and f,

$$M_T = -\frac{x_i}{f}. \tag{10.52}$$

We also find similar triangles on the left-hand-side of the lens, giving

$$\frac{y_o}{x_o} = -\frac{y_i}{f}, \tag{10.53}$$

which gives

$$M_T = -\frac{f}{x_o}. \tag{10.54}$$

Combining Eqs. (10.50) and (10.54), we find

$$f^2 = x_o x_i. \tag{10.55}$$

10.7.2 Concave lens

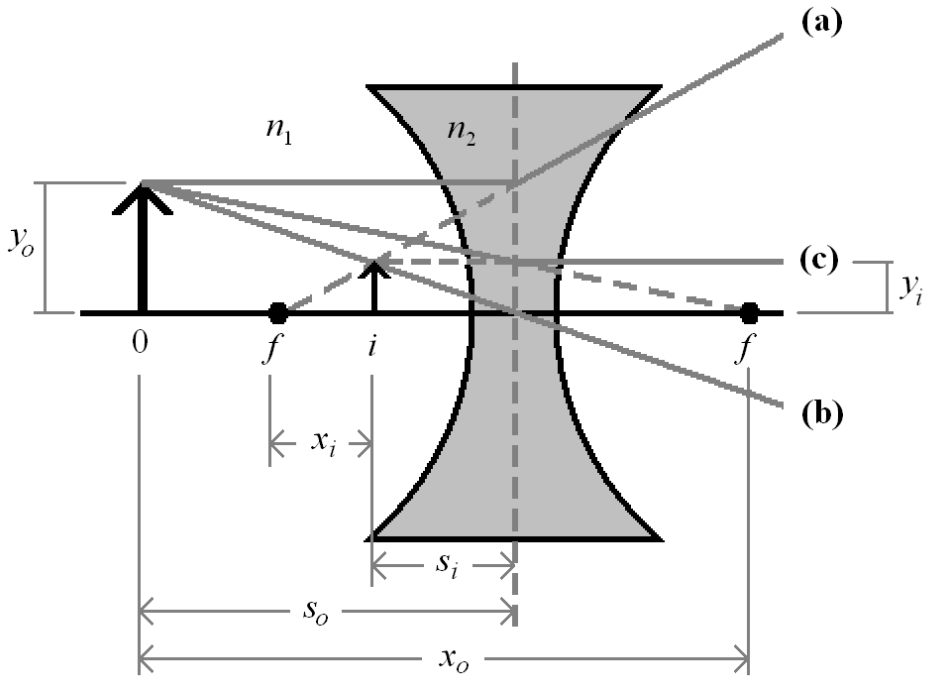


Figure 10.8: Sketch of a concave lens ($f < 0$) with an object to be imaged at 0.

Figure 10.8 shows a sketch of a concave lens with an object to be imaged at o. Referring to this figure, the image is constructed by sketching rays from the tip of the object:

1. parallel to the optical axis of the lens. Since $f < 0$, this must pass through f on the *left* of the lens (as a virtual ray).
2. passing through the centre of the lens without deviation;
3. following the line through f on the *right* of the lens (this extension is virtual on the right) and emerging parallel to the optical axis. The parallel line is then extended to the left as a virtual ray (dashed)

The tip of the image occurs where the rays from the tip of the object intersect on the left-hand-side, shown in the figure at i. In this case, x_o is the distance between the object and f on the *right* of the lens whilst x_i is the distance between the image and f on the *left* of the lens. The distances x_o, y_o and s_o for the object and x_i, y_i and s_i for the image are shown on the graph.

The magnification is given by

$$M = \frac{y_i}{y_o} \tag{10.56}$$

From similar triangles, we have

$$\frac{y_o}{x_o} = -\frac{y_i}{f} \tag{10.57}$$

(note f is negative by convention). Hence, using Eq. (10.56), we have

$$M = \frac{y_i}{y_o} = -\frac{f}{x_o}. \tag{10.58}$$

Comparing similar triangles again, we see

$$-\frac{y_o}{f} = \frac{y_i}{x_i}, \tag{10.59}$$

giving

$$M = \frac{y_i}{y_o} = -\frac{x_i}{f}. \tag{10.60}$$

Combining these results, we have

$$M = -\frac{f}{x_o} = -\frac{x_i}{f}, \tag{10.61}$$

which implies

$$f^2 = x_o x_i. \tag{10.62}$$

This is the same result as for the convex lens ($f > 0$).

10.8 Monochromatic aberration

10.8.1 Third-order aberration

In the earlier analysis of spherical lenses, we applied the *paraxial approximation* to obtain Eq. (10.14) from the exact expression Eq. (10.9). Specifically, this meant assuming ϕ is close to zero and making the approximations

$$\sin\phi \approx \phi \tag{10.63}$$

and

$$\cos\phi \approx 1. \tag{10.64}$$

More generally, the sinusoidal functions may be expanded Taylor power series as

$$\sin\phi = \phi - \frac{\phi^3}{3!} + \ldots \tag{10.65}$$

and

$$\cos\phi = 1 - \frac{\phi^2}{2!} + \ldots. \tag{10.66}$$

These higher-order terms add in correcting terms as ϕ is allowed to increase. Due to the power of 3 in the $\sin$ expansion, the second terms in each expansion is referred to as a *third-order correction*. The effect of these terms is to introduce *third-order aberrations* from the paraxial treatment of spherical lenses.

Employing the third-order terms, a corrected version of Eq. (10.14) may then be obtained. First, we remind ourselves of the correct expressions for l_o and l_i appearing in Eq. (10.9). We have

$$l_o^2 = (s_o + R)^2 + R^2 - 2(s_o + R) R\cos\phi \tag{10.67}$$

and

$$l_i^2 = (s_i - R)^2 + R^2 + 2(s_i - R) R\cos\phi. \tag{10.68}$$

where R is the radius of the spherical lens.

On the basis of the *paraxial approximation*, we had

$$l_o^2 \to s_o^2 \tag{10.69}$$

and

$$l_i^2 \to s_i^2. \tag{10.70}$$

We now use

$$\cos\phi \to 1 - \frac{\phi^2}{2}, \tag{10.71}$$

leading to

$$l_o \to s_o \left[1 + \frac{(s_o + R)}{s_o^2} R\phi^2\right]^{1/2} \tag{10.72}$$

and

$$l_i \to s_i \left[1 + \frac{(R - s_i)}{s_i^2} R\phi^2\right]^{1/2}. \tag{10.73}$$

Substituting these expressions into Eq. (10.9) then gives

$$\begin{aligned}
\left(\frac{n_1}{l_0} + \frac{n_2}{l_i}\right) R &= \frac{n_2 s_i}{l_i} - \frac{n_1 s_o}{l_0}, \\
&\to n_2 \left[1 + \frac{(R - s_i)}{s_i^2} R\phi^2\right]^{-1/2} \\
&- n_1 \left[1 + \frac{(s_o + R)}{s_o^2} R\phi^2\right]^{-1/2}, \\
&\approx (n_2 - n_1) + \frac{R^2\phi^2}{2}\left[\frac{n_1}{s_o}\left(\frac{1}{R} + \frac{1}{s_o}\right) + \frac{n_2}{s_i}\left(\frac{1}{R} - \frac{1}{s_i}\right)\right],
\end{aligned}$$

where n_2 is the refractive index of the lens and n_1 is the refractive index of the surrounding medium. Thus, our new expression is

$$\left(\frac{n_1}{s_0} + \frac{n_2}{s_i}\right) = \frac{1}{R}(n_2 - n_1) + \frac{R\phi^2}{2}\left[\frac{n_1}{s_o}\left(\frac{1}{R} + \frac{1}{s_o}\right) + \frac{n_2}{s_i}\left(\frac{1}{R} - \frac{1}{s_i}\right)\right]. \tag{10.74}$$

Note that the $R\phi^2$ term gives a measure of the displacement of the intersection of the ray with the lens from the optical axis. Thus, in the third-order treatment, the new term increases in proportion with the *square of the angular displacement*.

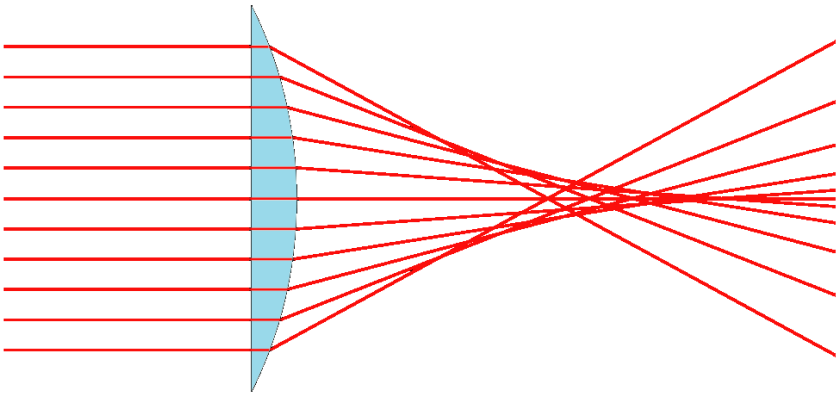


Figure 10.9: Illustration of spherical aberration for a spherical lens of radius R and refractive index n_2 in a medium with refractive index n_1, showing the deviation from the paraxial approximation.

10.8.2 Spherical Aberration

Figure 10.9 illustrates a particular case of third-order aberration in which the rays of light are parallel to the optical axis on the lens-side of the system. In this case, we have $s_i \to \infty$ so that Eq. (10.74) becomes

$$\frac{n_1}{s_0} = \frac{1}{R}(n_2 - n_1) + \frac{R\phi^2}{2}\frac{n_1}{s_o}\left(\frac{1}{R} + \frac{1}{s_o}\right). \tag{10.75}$$

To use this expression, we note that in the paraxial approximation

$$s_o = R\frac{n_1}{(n_2 - n_1)}. \tag{10.76}$$

We use this expression for s_o on the right-hand-side of Eq. (10.75) whilst substituting s'_o for s_o on the left-hand-side. We then solve for s'_o, giving

$$s'_o = \left[\frac{1}{R}\frac{(n_2 - n_1)}{n_1} + \frac{R\phi^2}{2s_o}\left(\frac{1}{R} + \frac{1}{s_o}\right)\right]^{-1}. \tag{10.77}$$

Longitudinal spherical aberration

The *longitudinal spherical aberration* L_{SA} is defined as the distance between the intersection of a ray with the optical axis and the paraxial focus. Using the expressions above, this is therefore

$$L_{SA} = s_o - s'_o. \tag{10.78}$$

Transverse spherical aberration

The *transverse spherical aberration* T_{SA} is defined as the perpendicular distance above (or below) the paraxial focus that a ray actually passes.

The circle of least confusion

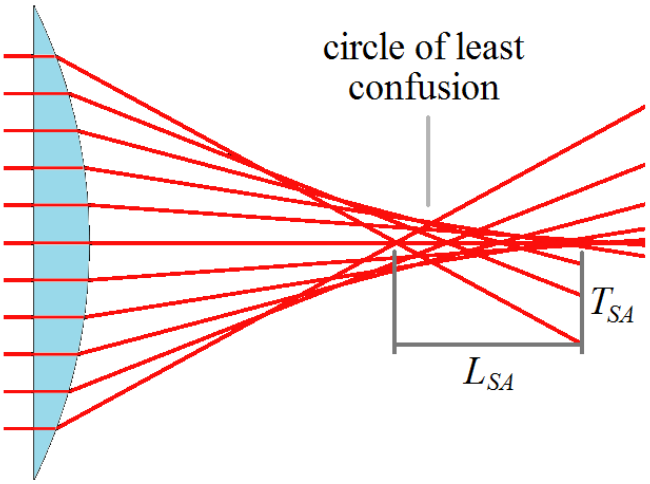


Figure 10.10: Longitudinal, transverse spherical aberration and the circle of least confusion for the lens illustrated in Fig. 10.9.

The *circle of least confusion* is the region between the intersection of all of the rays with the optical axis and the paraxial focus where the rays form their narrowest beam. This is illustrated in Fig. 10.10 for the same lens shown in Fig. 10.9.

Soft focus imaging

Although spherical aberration can be a particular problem for telescopes, most modern astronomical instruments now use reflecting mirrors rather than lenses. However, the deliberate introduction of spherical aberration into camera optics allows the use of *soft focus* imaging in photography. An example is shown in Fig. 10.11.

10.8.3 Coma

Coma (see Fig. 10.12) is a type of aberration in which off-axis points develop a 'comet'-like tail, due to the variable magnification through the optical system. It affects both lenses and mirrors. In particular, Fig. 10.12 illustrates the problem for a parabolic mirror. For on-axis planar wavefronts,

Figure 10.11: The deliberate use of spherical aberration in photography to produce 'soft focus' imaging (left image).

a parabolic mirror images perfectly to a point. However, coma manifests for light at an angle to the optical axis. Newtonian reflectors may be corrected for coma via the incorporation of a dual system of plano-convex and plano-concave lenses fitted to the eyepiece.

10.8.4 Astigmatism

Astigmatism is characterised by rays travelling in perpendicular planes through an optical system seeing different focal lengths. It is a particularly common form of sight defect in the human eye. This phenomenon is illustrated in Fig. 10.13, where the optical axis lies along the intersection of two perpendicular planes

- The vertical plane is the *tangential* plane
- The horizontal plane is the *sagittal* plane

We see in the figure that each plane is associated with a different focal length.

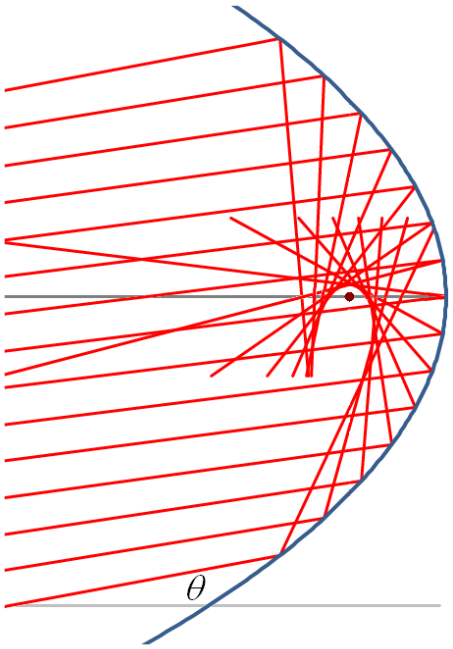


Figure 10.12: The phenomenon of coma for off-axis rays in a parabolic mirror. The dot marked on the image is the focus of the mirror.

10.8.5 Field Curvature

Field curvature is an aberration that occurs when light that is focused onto a curved surface is projected onto a planar screen. This is illustrated in Fig. 10.14.

10.8.6 Distortion

Distortion arises when there is non-uniform magnification of an image with radial distance from the optical axis. Fig. 10.15 illustrates three types of distortion

- *barrel distortion*, in which the magnification decreases with distance from the optical axis
- *pincushion distortion*, in which the magnification increases with distance from the optical axis
- *moustache distortion*, in which initially the magnification decreases with distance from the optical axis, whilst at further distances, the magnification increases with distance.

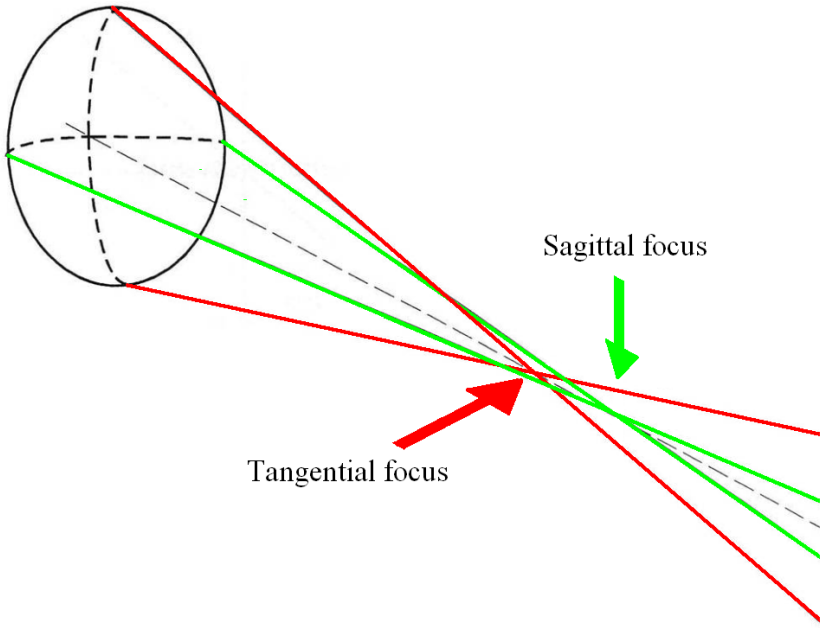


Figure 10.13: Illustration of astigmatism in which a lens has different focal lengths in each perpendicular direction.

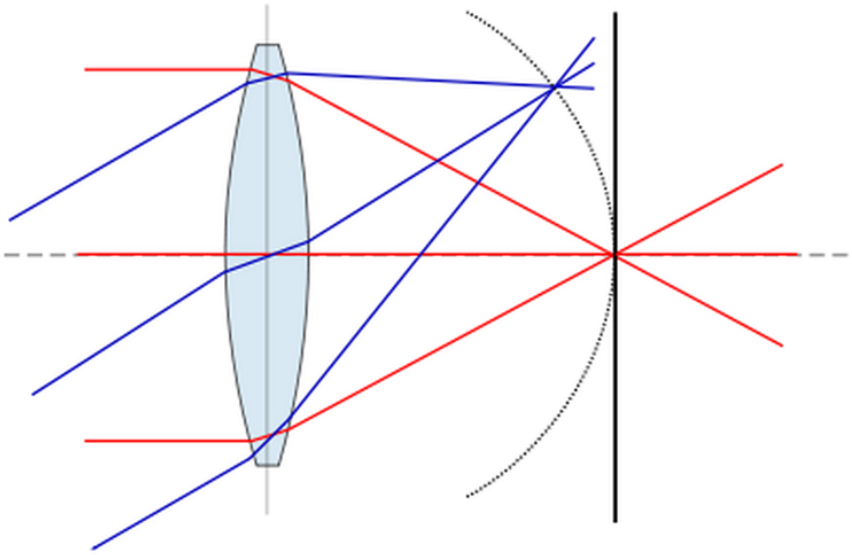


Figure 10.14: Illustration of field curvature, showing the curved image (dotted line). This cannot be projected onto the plane screen without warping the image.

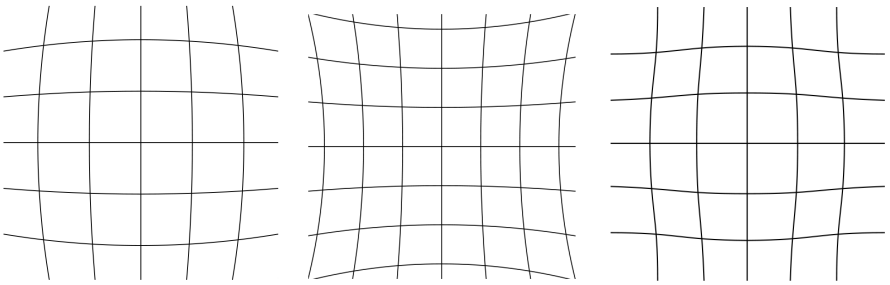


Figure 10.15: (a) Barrel distortion, in which the magnification decreases with distance from the optical axis. (b) Pincushion distortion, in which the magnification increases with distance from the optical axis. (c) Mustache distortion, in which initially the magnification decreases with distance from the optical axis, whilst at further distances, the magnification increases with distance.

10.9 Summary

- The **paraxial (small angle) approximation** may be used to obtain analytical formula for spherical lens surfaces
- The general equation for a spherical lens surface of radius R is

$$\boxed{\left(\frac{n_1}{s_0}+\frac{n_2}{s_i}\right)=\frac{1}{R}\left(n_2-n_1\right).} \tag{10.79}$$

- **Analysis of 'thin' lenses**
  - *Thin lens equation*

$$\boxed{\frac{1}{s_o}+\frac{1}{s_i}=\frac{n_2-n_1}{n_1}\left(\frac{1}{R_1}-\frac{1}{R_2}\right).} \tag{10.80}$$

  - *Lens maker's formula*

$$\boxed{\frac{1}{f}=\frac{n_2-n_1}{n_1}\left(\frac{1}{R_1}-\frac{1}{R_2}\right).} \tag{10.81}$$

  - *Gaussian lens formula*

$$\boxed{\frac{1}{s_o}+\frac{1}{s_i}=\frac{1}{f},} \tag{10.82}$$

- Monochromatic aberration
  - Third order aberration
  - Spherical aberration
  - Coma
  - Astigmatism
  - Field curvature
  - Distortion

Part V

Crystal Optics

11. Crystal Symmetry

11.1 General remarks

Consideration of the symmetries of a crystal allows us to derive physical properties from the formal description. The mathematical apparatus used to formulate crystal symmetry is that of *group theory*. Although appearing to be a somewhat abstract notion, this belies the analytical power of the tools it provides.

For our purposes, after working through the formal machinery, we shall look at how crystal symmetry may be used to extract information about the electric susceptibility tensor. We shall find that the properties of the tensor fall into three optical classes associated with certain crystal systems. These are

- **isotropic**

  for which the refractive index is the same in all directions.

- **uniaxial**

  in which we have an *optic axis* associated with its own refractive index.

- **biaxial**

  in which we have *two* optical axes associated with different refractive indices.

Note that we only consider the *linear* susceptibility.

11.2 Learning objectives

- Group theory
- Symmetry of a square
- Point groups in 2D
- Point groups in 3D
- Symmetry of the electric susceptibility

- Principal crystal axes
- Symmetry operations

11.3 Group theory

11.3.1 Definition of a group

A *group* is a set of elements which can be combined by some operation, subject to certain conditions.

For example, let us suppose some set G containing elements $x_1, x_2, ..., x_N$ (note that we may let $N \to \infty$), on which a binary operator $\otimes$ acts.

There are then four conditions that must be met for G and $\otimes$ to define a group.

- **Identity**

  There exists an *identity element* I (or E) such that for any element of the group

  $$I \otimes x_i = x_i \otimes I = x_i.$$

- **Invertibility**

  For each element x_i the group contains an *inverse element* x_i^{-1} such that

  $$x_i \otimes x_i^{-1} = x_i^{-1} \otimes x_i = I.$$

- **Closure**

  For any two elements of G, x_i and x_j, if

  $$x_i \otimes x_j = x_k,$$

  then x_k is also in G.

- **Associativity**

  For all elements of G,

  $$(x_i \otimes x_j) \otimes x_k = x_i \otimes (x_j \otimes x_k).$$

Abelian and non-abelian groups

If the operation of the group is commutative, i.e.

$$x_i \otimes x_j = x_j \otimes x_i,$$

the group is said to be abelian. Otherwise, it is a non-abelian group.

As an example of an abelian group, let us consider the operation of addition defined on the integers. We may use the familiar symbol '+' to denote the operation of addition. Now, for any two integers

$$a + b = c,$$

where c is also an integer. Hence the condition of *closure* is met.

We also know that

$$(a + b) + c = a + (b + c),$$

so the condition of *associativity* is met.

The integers include the number 0. Since

$$a + 0 = 0 + a = a,$$

0 is the identity element. Moreover, for every integer a there exists another $-a$, where

$$a + (-a) = (-a) + a = 0.$$

Hence, every element has an *inverse*. Since addition of integers is also commutative, this is an *abelian* group.

11.4 Symmetry of a square

A *symmetry operation* is a transformation of some kind that leaves an entity unchanged. As a simple example, we will consider the geometrical symmetry operations of a *square*. In the course of our investigation, we will uncover two important facts about such operations

- The symmetry operations may be represented by *matrices*.
- The set of symmetry operations forms a *group*.

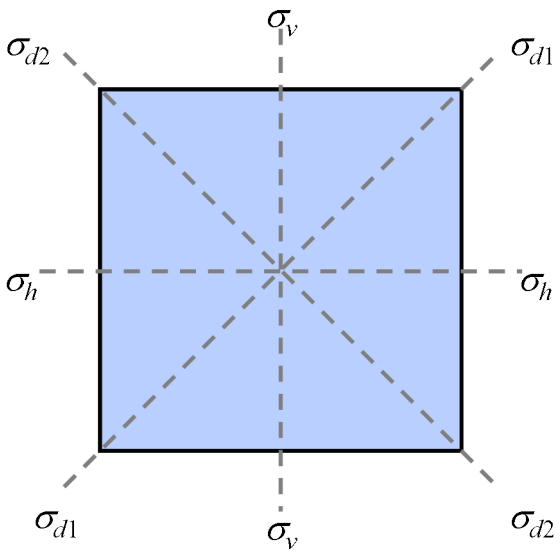


Figure 11.1: Lines of reflection symmetry for a square. Note that reflections are often indicated by the symbol σ by convention.

11.4.1 Reflection lines

Figure 11.1 shows the reflection lines of symmetry for a square. In any particular case, the symmetry operation σ_i is then just a reflection in the corresponding axis. As can be seen, there are four possible reflection operations that leave the square unchanged. These are detailed in Fig. 11.2, where the coloured dots have been added for reference only.

11.4.2 Rotations

In addition to reflections, we also have rotation operations that leave the square unchanged. These are shown in Fig. 11.3, where the sense of the rotations is anti-clockwise. Again. the coloured dots have been added for reference only. Note now, however, that the rotation by 2π may be thought of as equivalent to 'doing nothing'. It is therefore equivalent to the *identity* operation, which just maps the square to itself.

11.4.3 Combining operations

Symmetry operations may be combined. Consider the operation of a rotation by $\pi/2$ followed by a reflection in the horizontal line. Letting R be the combined operation, we would write this as

$$R = \sigma_h C_{\pi/2}. \tag{11.1}$$

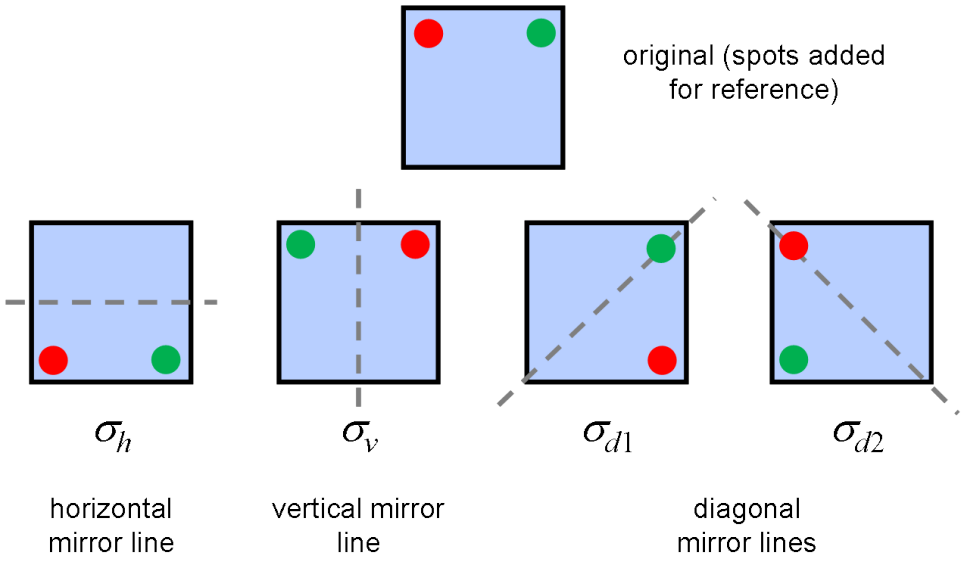


Figure 11.2: Reflection symmetry operations for a square. Note that the dots are added for reference only.

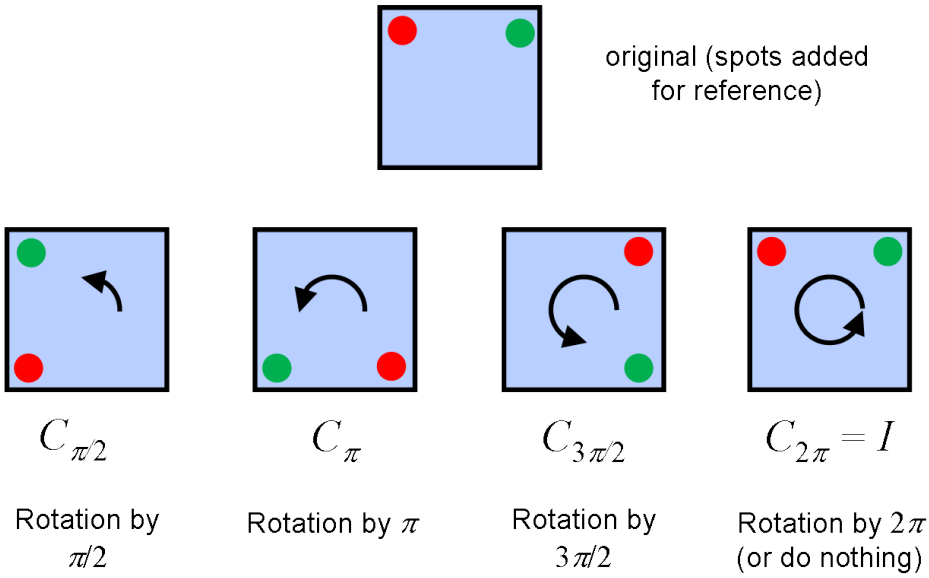


Figure 11.3: Rotation symmetry operations for a square. Note that the dots are added for reference only.

That is, we write the *first* operation to be performed on the right-hand-side. A little consideration should convince one that this is equivalent to a reflection in the $d2$ axis. In other words

$$\sigma_h C_{\pi/2} = \sigma_{d2}. \tag{11.2}$$

Combining all operations in this way, we can build a *multiplication table* for the symmetries of the square, as shown in Table 11.1.

Table 11.1: Multiplication table for the symmetry operations of a square. The operations in the first row are performed first followed by the operations in the first column.

	I	$C_{\pi/2}$	C_π	$C_{3\pi/2}$	σ_h	σ_v	σ_{d1}	σ_{d2}
I	I	$C_{\pi/2}$	C_π	$C_{3\pi/2}$	σ_h	σ_v	σ_{d1}	σ_{d2}
$C_{3\pi/2}$	$C_{3\pi/2}$	I	$C_{\pi/2}$	C_π	σ_{d2}	σ_{d1}	σ_h	σ_v
C_π	C_π	$C_{3\pi/2}$	I	$C_{\pi/2}$	σ_v	σ_h	σ_{d2}	σ_{d1}
$C_{\pi/2}$	$C_{\pi/2}$	C_π	$C_{3\pi/2}$	I	σ_{d1}	σ_{d2}	σ_v	σ_h
σ_h	σ_h	σ_{d2}	σ_v	σ_{d1}	I	C_π	$C_{3\pi/2}$	$C_{\pi/2}$
σ_v	σ_v	σ_{d1}	σ_h	σ_{d2}	C_π	I	$C_{\pi/2}$	$C_{3\pi/2}$
σ_{d1}	σ_{d1}	σ_h	σ_{d2}	σ_v	$C_{\pi/2}$	$C_{3\pi/2}$	I	C_π
σ_{d2}	σ_{d2}	σ_v	σ_{d1}	σ_h	$C_{3\pi/2}$	$C_{\pi/2}$	C_π	I

11.4.4 Matrix representations

We may represent the symmetry operations of a geometrical object in terms of the matrices that reproduce the operations. In 2D we may limit our consideration to the action of a 2×2 matrix on a point with coordinated x and y. Representing this point by a column vector, we may put

$$\mathbf{r} = \begin{bmatrix} x \\ y \end{bmatrix}. \tag{11.3}$$

The σ_h operation may be represented by a reflection in the x-axis as shown in Fig. 11.4 (a). In this case $y \to -y$ and x remains unchanged. Hence our transformed points are

$$\mathbf{r}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ -y \end{bmatrix}. \tag{11.4}$$

Representing this as a matrix operation of the form

$$\mathbf{r}' = \sigma_h \mathbf{r}, \tag{11.5}$$

we see that

$$\boxed{\sigma_h = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.} \tag{11.6}$$

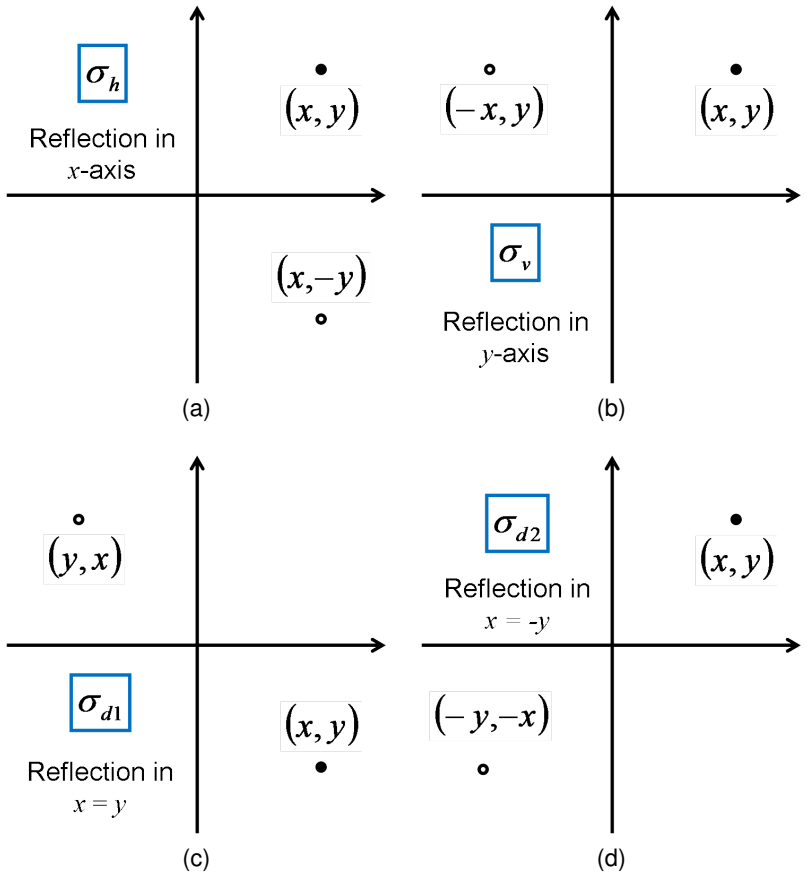


Figure 11.4: .

From Fig. 11.4 (b) for σ_v we have $x \to -x$ and y remains unchanged. Hence, we deduce

$$\boxed{\sigma_v = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.} \tag{11.7}$$

Considering Figs. 11.4 (c) and (d), we also find

$$\boxed{\sigma_{d1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}} \tag{11.8}$$

and

$$\boxed{\sigma_{d2} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.} \tag{11.9}$$

For an arbitrary rotation (see Appendix A.2), we may define the *rotation matrix*

$$\boxed{C_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.} \tag{11.10}$$

For the particular symmetry rotations of the square, we then just substitute for θ

$$\boxed{C_{\pi/2} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},} \tag{11.11}$$

$$\boxed{C_\pi = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}} \tag{11.12}$$

and

$$\boxed{C_{3\pi/2} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.} \tag{11.13}$$

Note that for $\theta = 2\pi$ we just get the identity matrix $C_{2\pi} = I$, where

$$\boxed{I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.} \tag{11.14}$$

Combining operators

We saw previously that, for instance,

$$\sigma_h C_{\pi/2} = \sigma_{d2}. \tag{11.15}$$

We can perform this using the matrix representation via matrix multiplication

$$\sigma_h C_{\pi/2} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} = \sigma_{d2}. \tag{11.16}$$

It is straightforward to then show that the multiplication table of Table 11.1 is replicated by the appropriate combinations of matrices.

We see that the set of symmetry operations *forms a group*

- **Identity** - the set of operations includes the identity operation.
- **Invertibility** - every element has an inverse element (every row or every column contains the identity element as a product).
- **Closure** - every product of a symmetry operation is in the set of symmetry operations.

- **Associativity** - since matrix multiplication is associative, so are the symmetry operations.

Note that since, for instance,

$$\sigma_h C_{\pi/2} \neq C_{\pi/2}\sigma_h, \tag{11.17}$$

this group is *non-abelian*.

11.5 Point groups in 2D

Table 11.2: Multiplication table for C_4.

	I	$C_{\pi/2}$	C_π	$C_{3\pi/2}$
I	I	$C_{\pi/2}$	C_π	$C_{3\pi/2}$
$C_{3\pi/2}$	$C_{3\pi/2}$	I	$C_{\pi/2}$	C_π
C_π	C_π	$C_{3\pi/2}$	I	$C_{\pi/2}$
$C_{\pi/2}$	$C_{\pi/2}$	C_π	$C_{3\pi/2}$	I

A *point group* is a group of symmetries (known as isometries as distance is preserved) that keep at least one point fixed.

Consideration of the symmetries of the square just found, we see that they constitute a point group, since

- reflections keep all points along the mirror line fixed
- rotations keep the point at the centre of the square fixed
- the identity keeps *all* points fixed

11.5.1 Families of groups

In 2D, the types of point groups occur in two families of groups. Using *Schönflies notation*

- **Cyclic groups**

  C_n - groups of n-fold rotation

- **Dihedral groups**

  D_n - groups of n-fold rotation and reflection

The point group of the square is therefore known as D_4. It is said to have *order* 8 corresponding to the 8 symmetry operations. The 4-fold rotation group C_4 is then said to be a *subgroup* of D_4 (note from the multiplication table that C_4 does indeed form a group).

11.5.2 Group generators

Generators of C_4

Consider the repeated operation of $C_{\pi/2}$

$$\begin{aligned} C_{\pi/2}I &= C_{\pi/2}, \\ C_{\pi/2}C_{\pi/2} &= C_{\pi/2}^2 = C_\pi, \\ C_{\pi/2}C_\pi &= C_{\pi/2}^3 = C_{3\pi/2}, \\ C_{\pi/2}C_{3\pi/2} &= C_{\pi/2}^4 = I. \end{aligned}$$

We see that on each operation, each member of C_4 is produced. Thus $C_{\pi/2}$ is said to be a *generator* of the group. It is straightforward to confirm that $C_{3\pi/2}$ is also a generator of C_4. Note that I and C_π are *not* generators of the group.

A group that may be generated from powers of a single element in this way is said to be a *cyclic group*. Another example of a cyclic group would be the group of elements $\{I, \sigma_h\}$ (σ_h is its own inverse).

Generators of D_4

Consider the repeated operation of $C_{\pi/2}$ on σ_h

$$\begin{aligned} C_{\pi/2}\sigma_h &= \sigma_{d1}, \\ C_{\pi/2}\sigma_{d1} &= C_{\pi/2}^2\sigma_h = \sigma_v, \\ C_{\pi/2}\sigma_v &= C_{\pi/2}^3\sigma_h = \sigma_{d2}, \\ C_{\pi/2}\sigma_{d2} &= C_{\pi/2}^4\sigma_h = \sigma_h. \end{aligned}$$

Since we also have

$$\sigma_h\sigma_h = \sigma_h^2 = I,$$

these two elements generate the elements of D_4 between them. Thus $\{C_{\pi/2}, \sigma_h\}$ is a *generating set* of D_4.

11.6 Point groups in 3D

In 3D, there are 7 *axial groups*

$$C_n, S_{2n}, C_{nh}, C_{nv}, D_n, D_{nd}, D_{nv} \tag{11.18}$$

and 7 *polyhedral groups*

$$T, T_d, T_h, O, O_h, I, I_h. \tag{11.19}$$

C_n is the same as the family of rotation groups in 2D, where the n-fold rotation is now around a rotation axis. Of the other groups, we shall only be concerned with O_h, having the *full octahedral symmetry*. This is the symmetry of the cube and has order 48 (i.e. there are 48 symmetry operations).

As a few examples, the simple cubic lattice has the same symmetry. The number of symmetries may be reduced by the configuration of the basis atoms. *Zinc blende* has the point group T_d. This group has order 24. *Diamond structure* has the point group O_h - the symmetry of the cube.

11.7 Symmetry of the electric susceptibility

From thermodynamic considerations (time symmetry), it was shown in Chapter (6) that the electric susceptibility tensor must be symmetric, i.e. that the elements satisfy

$$\chi_{ij} = \chi_{ji}. \tag{11.20}$$

Another way describing this symmetry is to say that $\boldsymbol{\chi}_{\mathbf{E}}$ is equal to its *transpose* (see Appendix A.2). Thus, we may re-write Eq. (11.20) as

$$\boldsymbol{\chi}_{\mathbf{E}} = \boldsymbol{\chi}_{\mathbf{E}}^{T}. \tag{11.21}$$

We shall now explore the implications of this symmetry, making use of the following two theorems:

- **Theorem 1**

  A symmetric $n \times n$ matrix has n real eigenvalues associated with n eigenvectors.

- **Theorem 2**

  The eigenvectors of a symmetric $n \times n$ matrix are orthogonal to one another.

  Hence, if $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are eigenvectors of a symmetric $n \times n$ matrix, then

  $$\mathbf{x}^{(i)T}\mathbf{x}^{(j)} = \delta_{ij}. \tag{11.22}$$

  where δ_{ij} is the *Kronecker delta* defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

If $\mathbf{A}$ is such a symmetric matrix, then we have

$$\mathbf{A}\mathbf{x}^{(i)} = \lambda^{(i)}\mathbf{x}^{(i)}, \tag{11.23}$$

where $\lambda^{(i)}$ is a real eigenvalue and $\mathbf{x}^{(i)}$ is its eigenvector. Putting

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & x_1^{(3)} \\ x_2^{(1)} & x_2^{(2)} & x_2^{(3)} \\ x_3^{(1)} & x_2^{(2)} & x_2^{(3)} \end{bmatrix} \tag{11.24}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda^{(1)} & 0 & 0 \\ 0 & \lambda^{(2)} & 0 \\ 0 & 0 & \lambda^{(3)} \end{bmatrix}, \tag{11.25}$$

the eigenvalue relation can be written out in full as

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}. \tag{11.26}$$

Multiplying on the left by the inverse of $\mathbf{X}$, we have

$$\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{\Lambda}. \tag{11.27}$$

The transformation $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ therefore diagonalises the matrix $\mathbf{A}$.

11.8 Principal crystal axes

We have seen previously that the electric displacement $\mathbf{D}$ may be given in terms of the electric field $\mathbf{E}$ by

$$\mathbf{D} = \varepsilon_0 \left(\mathbf{I} + \boldsymbol{\chi}_E\right) \mathbf{E}, \tag{11.28}$$

where $\boldsymbol{\chi}_E$ gives the *frequency response* of a material medium to $\mathbf{E}$. Let us diagonalise $\boldsymbol{\chi}_E$ according to the prescription given above

$$\boldsymbol{\chi}'_E = \mathbf{U}^{-1}\boldsymbol{\chi}_E\mathbf{U}, \tag{11.29}$$

where $\mathbf{U}$ is the matrix formed from the eigenvectors of $\boldsymbol{\chi}_E$. We may obtain the diagonalised matrix $\boldsymbol{\chi}'_E$ from Eq. (11.28) by multiplying on the left by $\mathbf{U}^{-1}$

$$\mathbf{U}^{-1}\mathbf{D} = \varepsilon_0\mathbf{U}^{-1}\left(\mathbf{I} + \boldsymbol{\chi}_E\right)\mathbf{E}. \tag{11.30}$$

Now, since $\mathbf{U}\mathbf{U}^{-1} = \mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$, we may insert this before $\mathbf{E}$. Then, since matrix multiplication is distributive over addition, we have

$$\begin{aligned} \mathbf{U}^{-1}\mathbf{D} &= \varepsilon_0 \mathbf{U}^{-1}\left(\mathbf{I} + \boldsymbol{\chi}_E\right)\mathbf{U}\mathbf{U}^{-1}\mathbf{E}, \\ &= \varepsilon_0 \left(\mathbf{U}^{-1}\mathbf{I}\mathbf{U} + \mathbf{U}^{-1}\boldsymbol{\chi}_E\mathbf{U}\right)\mathbf{U}^{-1}\mathbf{E}, \\ &= \varepsilon_0 \left(\mathbf{I} + \mathbf{U}^{-1}\boldsymbol{\chi}_E\mathbf{U}\right)\mathbf{U}^{-1}\mathbf{E}. \end{aligned} \tag{11.31}$$

Putting $\mathbf{D}' = \mathbf{U}^{-1}\mathbf{D}$ and $\mathbf{E}' = \mathbf{U}^{-1}\mathbf{E}$, we may write this as

$$\mathbf{D}' = \varepsilon_0 \left(\mathbf{I} + \boldsymbol{\chi}'_E\right)\mathbf{E}', \tag{11.32}$$

which is exactly the same form as Eq. (11.28). Now, however, $\boldsymbol{\chi}'_E$ is diagonal and we have the simplified relation between the components of $\mathbf{D}'$ and $\mathbf{E}'$

$$D'_i = \varepsilon_0 \left(1 + \chi'_{ii}\right) E'_i. \tag{11.33}$$

Essentially what we have done is to transform from an original coordinate system O to a new system O' in which the elements of the electric susceptibility tensor are diagonal. The Cartesian axes of this new coordinate system then correspond to the *principal axes* of the crystal.

11.8.1 Diagonalisation in terms of rotations

The transformation to the new coordinate system was obtained via the matrix operation $\mathbf{U}^{-1}$. According to *Euler's rotation theorem* we may change from one Cartesian system to any other by some sequence of (up to three) rotations. Therefore, we may interpret $\mathbf{U}^{-1}$ as the product of these rotations. There are different possible choices for the rotations we might apply but we shall use the following:

(1) A rotation around the z-axis by an angle α

As a matrix operation, this rotation is given by

$$\mathbf{R}^{(z)}_{\alpha} = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11.34}$$

(2) A rotation around the x-axis by an angle β

This is given by

$$\mathbf{R}^{(x)}_{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\beta & \sin\beta \\ 0 & -\sin\beta & \cos\beta \end{bmatrix}. \tag{11.35}$$

(3) A rotation around the z-axis by an angle γ

This is given by

$$\mathbf{R}_{\gamma}^{(z)} = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11.36}$$

Thus, the total transformation is given by

$$\mathbf{U}^{-1} = \mathbf{R}_{\gamma}^{(z)}\mathbf{R}_{\beta}^{(x)}\mathbf{R}_{\alpha}^{(z)}. \tag{11.37}$$

11.9 Symmetry operations

11.9.1 Symmetry and physics

So far, we have just considered symmetry operations in the abstract. However, consideration of these operations yields physical information about the crystal structure. In particular

- *A physical property of a system must reflect the symmetry of the system,*

The physical properties we shall be concerned with are those of the electric susceptibility tensor.

11.9.2 The symmetry of the susceptibility

Let us now consider further rotations applied to the diagonalised susceptibility tensor corresponding to *symmetry operations* of the particular crystal structure. It turns out that there are *seven crystal systems* (see Table 11.3) associated with particular sets of symmetry operations.

Let us assume that we have already diagonalised $\boldsymbol{\chi}_E$ and consider some transformation represented by the matrix $\mathbf{T}$. Applying this to Eq. (11.28) gives

$$\mathbf{TD} = \varepsilon_0\mathbf{T}\left(\mathbf{I} + \boldsymbol{\chi}_E\right)\mathbf{T}^{-1}\mathbf{TE}. \tag{11.38}$$

Thus, $\boldsymbol{\chi}_E$ is transformed according to

$$\boldsymbol{\chi}_E' = \mathbf{T}\boldsymbol{\chi}_E\mathbf{T}^{-1}. \tag{11.39}$$

In other words, dropping the prime notation for the diagonalised matrix, $\boldsymbol{\chi}_E$ has the form

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix}. \tag{11.40}$$

It is straightforward to show that reflections and rotations by π impose no constraints on the form of $\boldsymbol{\chi}_E$. On the other hand, we shall see that the C_4 and C_6 symmetries *do* impose constraints on $\boldsymbol{\chi}_E$, yielding three *optical classes*.

11.9.3 Isotropic systems

Cubic symmetry

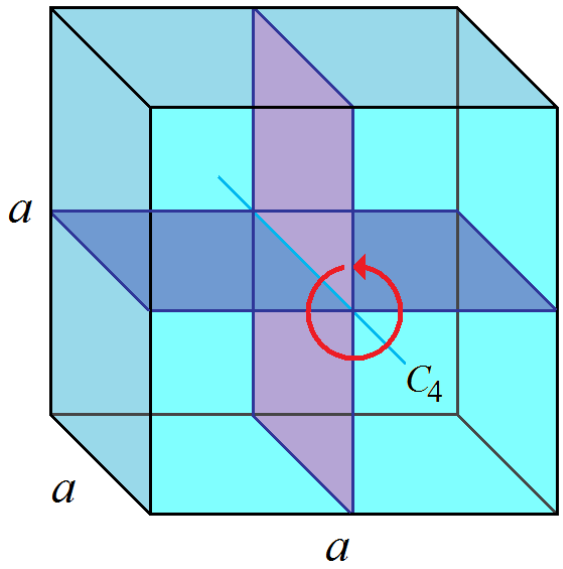


Figure 11.5: Illustration of the C_4 symmetry for a cubic system. Note that the cube has three axes with this symmetry.

As a particular example of how we may make use of point group symmetries, we shall investigate the optical properties of materials with *cubic symmetry*. Let us consider the C_4 group associated with some axis of rotational symmetry (a subgroup of O_h.

For definiteness, let us take the z-axis to be the axis of symmetry for definiteness. Earlier we saw that a $C_{\pi/2}$ is a generator of C_4. From Eq. (11.34) we then see that this rotation is given by

$$C_{\pi/2} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{11.41}$$

whilst its inverse is

$$\left(C_{\pi/2}\right)^{-1} = C_{3\pi/2} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11.42}$$

Substituting $C_{\pi/2}$ for $\mathbf{T}$ in Eq. (11.39) gives

$$\begin{aligned} \boldsymbol{\chi}'_E &= \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\chi_x & 0 \\ \chi_y & 0 & 0 \\ 0 & 0 & \chi_z \end{bmatrix} \\ &= \begin{bmatrix} \chi_y & 0 & 0 \\ 0 & \chi_x & 0 \\ 0 & 0 & \chi_z \end{bmatrix}. \end{aligned} \tag{11.43}$$

Comparing this with the untransformed matrix, we see that χ_x and χ_y have changed places. For a crystal with this symmetry, the physical properties of the system must not change under this operation. Therefore, it must *at least* be the case that

$$\chi_x = \chi_y. \tag{11.44}$$

Similar conclusions would be obtained had the x or y axes been axes of symmetry.

Now, since a cube has the C_4 symmetry associated with the x, y and z axes, we may conclude that for a crystal with *cubic symmetry*, the diagonal elements of $\boldsymbol{\chi}_E$ *must all be equal*. In other words, the cubic symmetry is *isotropic*. The susceptibility tensor therefore has the form

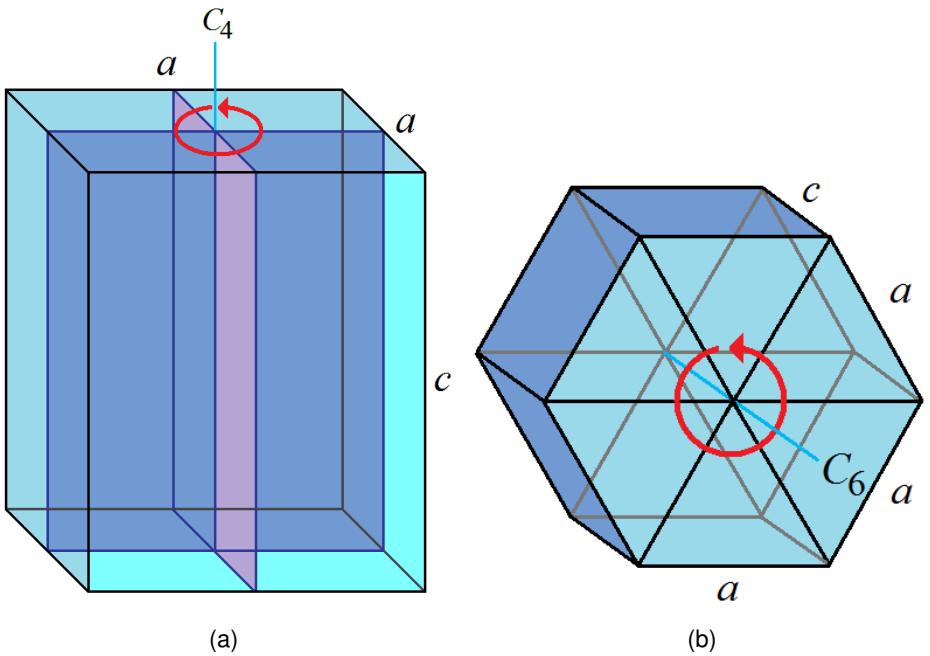
$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_o & 0 & 0 \\ 0 & \chi_o & 0 \\ 0 & 0 & \chi_o \end{bmatrix}. \tag{11.45}$$

Here we have used the 'o' subscript to stand for *ordinary*, in line with convention.

11.9.4 Uniaxial systems

Tetragonal symmetry

The *tetragonal* crystal system unit cell has just one square face and only one axis to which the C_4 symmetry symmetry applies (see Fig. 11.6 (a)).



(a) (b)

Figure 11.6: Illustration of the (a) single C_4 symmetry for a tetragonal system. Note that only the square face has this line of symmetry perpendicular to it. (b) The C_6 symmetry for the hexagonal system.

We can therefore only apply the transformation of Eq. (11.43) once and thus only enforce the condition that *two* of the diagonal elements are equal. If we take these to be the x and y elements, the susceptibility tensor takes the form

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_o & 0 & 0 \\ 0 & \chi_o & 0 \\ 0 & 0 & \chi_e \end{bmatrix}, \tag{11.46}$$

where the 'e' subscript stands for *extraordinary*. Crystals of this type are given the optical classification *uniaxial*, as the direction associated with χ_e is taken to be the *optical axis* of the system.

Hexagonal symmetry

The hexagonal symmetry has an axis with C_6 symmetry, as shown in Fig. 11.6 (b) (i.e. a 6-fold rotation symmetry). In this case, the generator of the group is the rotation matrix

$$C_{\pi/3} = \frac{1}{2}\begin{bmatrix} 1 & \sqrt{3} & 0 \\ -\sqrt{3} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{11.47}$$

Substituting this into Eq. (11.43) gives

$$\begin{aligned} \boldsymbol{\chi}'_E &= \frac{1}{4}\begin{bmatrix} 1 & \sqrt{3} & 0 \\ -\sqrt{3} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix}\begin{bmatrix} 1 & -\sqrt{3} & 0 \\ \sqrt{3} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix} 1 & \sqrt{3} & 0 \\ -\sqrt{3} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \chi_x & -\sqrt{3}\chi_x & 0 \\ \sqrt{3}\chi_y & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix} \\ &= \frac{1}{4}\begin{bmatrix} \chi_x + 3\chi_y & \sqrt{3}(\chi_y - \chi_x) & 0 \\ \sqrt{3}(\chi_y - \chi_x) & 3\chi_x + \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix}. \end{aligned} \tag{11.48}$$

The transformed tensor $\boldsymbol{\chi}'_E$ will therefore retain the same form as $\boldsymbol{\chi}_E$ if $\chi_x = \chi_y$. Thus, $\boldsymbol{\chi}_E$ must have the same form as Eq. (11.46).

Trigonal symmetry

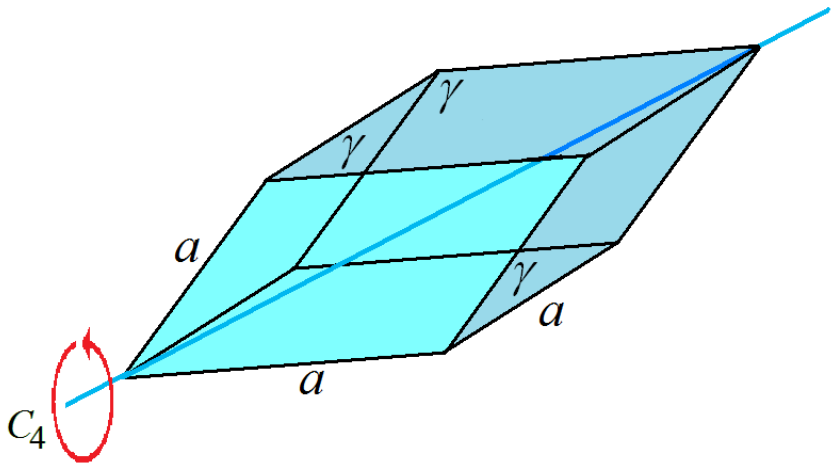


Figure 11.7: The C_4 symmetry for the trigonal system

Figure shows an illustration of the trigonal crystal system. Note that the angles γ and side lengths a are all equal. This gives a single axis with C_4 symmetry along one of the body diagonals. Note that the other body diagonal axes *do not have the* C_4 *symmetry*. This means that the number

Table 11.3: Crystal systems and their optical classifications.

Crystal system	Optical classification
Triclinic	
Monoclinic	Biaxial
Orthorhombic	
Trigonal	
Tetragonal	Uniaxial
Hexagonal	
Cubic	Isotropic

and types of rotation symmetries are the same as for the tetragonal system. Thus, the susceptibility tensor takes the form

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_o & 0 & 0 \\ 0 & \chi_o & 0 \\ 0 & 0 & \chi_e \end{bmatrix}, \tag{11.49}$$

11.9.5 Biaxial systems

In the case of triclinic, monoclinic and orthorhombic systems, there are no rotational symmetries constraining the form of the susceptibility tensor. On the basis of this set of symmetries at least, the susceptibilities must be of the form

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix}, \tag{11.50}$$

11.10 Summary

- **Group theory**
  - **Definition of a group**
    A *group* is a set of elements which can be combined by some operation.
    There are then four conditions that must be met to define a group.
    - **Identity**
    - **Invertibility**
    - **Closure**
    - **Associativity**
- **Symmetry of a square**
  - The symmetry operations may be represented by *matrices*.
  - The set of symmetry operations forms a *group*.
- There are 8 symmetry operations for the square
  - **Reflection**
    There are four axes of reflection: two diagonal σ_{d1} and σ_{d2}, horizontal σ_h and vertical σ_v.

$$\boxed{\sigma_{d1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.} \tag{11.51}$$

$$\boxed{\sigma_{d2} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.} \tag{11.52}$$

$$\boxed{\sigma_h = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.} \tag{11.53}$$

$$\boxed{\sigma_v = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.} \tag{11.54}$$

  - **Rotations**
    There are four rotations around the centre point of the square of $n\pi/2$, $n \in \{1, 2, 3, 4\}$. Note that the rotation of 2π 'does nothing'.

$$\boxed{C_{\pi/2} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.} \tag{11.55}$$

$$\boxed{C_\pi = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.} \tag{11.56}$$

$$\boxed{C_{3\pi/2} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.} \tag{11.57}$$

$$\boxed{C_{2\pi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.} \tag{11.58}$$

- **Point groups in 2D**

  A *point group* is a group of symmetries that keep at least one point fixed.

  - **Families of groups**

    In 2D, the types of point groups occur in two families of groups.

    * **Cyclic groups**
      C_n - groups of n-fold rotation

    * **Dihedral groups**
      D_n - groups of n-fold rotation and reflection

  - **Group generators**

    * **Generators of C_4**

$$\begin{aligned} C_{\pi/2}I &= C_{\pi/2}, \\ C_{\pi/2}C_{\pi/2} &= C_{\pi/2}^2 = C_\pi, \\ C_{\pi/2}C_\pi &= C_{\pi/2}^3 = C_{3\pi/2}, \\ C_{\pi/2}C_{3\pi/2} &= C_{\pi/2}^4 = I. \end{aligned}$$

    * **Generators of D_4**

$$\begin{aligned} C_{\pi/2}\sigma_h &= \sigma_{d1}, \\ C_{\pi/2}\sigma_{d1} &= C_{\pi/2}^2\sigma_h = \sigma_v, \\ C_{\pi/2}\sigma_v &= C_{\pi/2}^3\sigma_h = \sigma_{d2}, \\ C_{\pi/2}\sigma_{d2} &= C_{\pi/2}^4\sigma_h = \sigma_h. \end{aligned}$$

- **Point groups in 3D**

  In 3D, there are 7 *axial groups*

  $$C_n, S_{2n}, C_{nh}, C_{nv}, D_n, D_{nd}, D_{nv} \tag{11.59}$$

  and 7 *polyhedral groups*

  $$T, T_d, T_h, O, O_h, I, I_h. \tag{11.60}$$

  C_n is the same as the family of rotation groups in 2D, where the n-fold rotation is now around a rotation axis.

  O_h, has the *full octahedral symmetry*. This is the symmetry of the cube and has order 48 symmetry operations.

  Zinc blende has the point group T_d. This group has 24 symmetry operations.

  Diamond structure has the point group O_h.

- **Symmetry of the electric susceptibility**

  From time symmetry, the electric susceptibility tensor must be symmetric, i.e. that the elements satisfy

  $$\chi_{ij} = \chi_{ji}. \tag{11.61}$$

  A symmetric matrix $\mathbf{A}$ may be diagonalised via the transformation

  $$\mathbf{A}' = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}, \tag{11.62}$$

  where $\mathbf{X}$ is the matrix formed from the eigenvectors of $\mathbf{A}$.

- **Principal crystal axes**

  The electric displacement $\mathbf{D}$ is given by

  $$\mathbf{D} = \varepsilon_0 \left(\mathbf{I} + \boldsymbol{\chi}_E\right) \mathbf{E}. \tag{11.63}$$

  Diagonalising $\boldsymbol{\chi}_E$ to give $\boldsymbol{\chi}'_E$, we may deduce

  $$\mathbf{D}' = \varepsilon_0 \left(\mathbf{I} + \boldsymbol{\chi}'_E\right) \mathbf{E}'. \tag{11.64}$$

  Here, the eigenvectors of $\boldsymbol{\chi}'_E$ are the *principal axes* of the crystal.

- **Symmetry operations**

  - **Symmetry and physics**

* *A physical property of a system must reflect the symmetry of the system,*

The physical properties we shall be concerned with are those of the electric susceptibility tensor.

- **Cubic symmetry**

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_o & 0 & 0 \\ 0 & \chi_o & 0 \\ 0 & 0 & \chi_o \end{bmatrix}. \tag{11.65}$$

This is an *isotropic* system.

- **Trigonal, Tetragonal and Hexagonal symmetry**

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_o & 0 & 0 \\ 0 & \chi_o & 0 \\ 0 & 0 & \chi_e \end{bmatrix}. \tag{11.66}$$

This is a *uniaxial* system.

- **triclinic, monoclinic and orthorhombic symmetry**

$$\boldsymbol{\chi}_E = \begin{bmatrix} \chi_x & 0 & 0 \\ 0 & \chi_y & 0 \\ 0 & 0 & \chi_z \end{bmatrix}. \tag{11.67}$$

This is a *biaxial* system.

12. The Index Ellipsoid

12.1 General remarks

So far, we have concentrated on isotropic media. Some mention of *uniaxial* anisotropic media previously proved necessary in the context of wave plates and the phenomenon of birefringence. Here, however, we shall approach the problem with a more rigorous and comprehensive treatment.

The starting point for our analysis is with the *electric susceptibility tensor*. Firstly, we shall consider EM wave propagation in an anisotropic media, how the optical vibrations are resolved into modes. We shall see that there are up to two possible optic axes, encompassing the three cases of

- isotropic media
- uniaxial anisotropic media (one optic axis)
- biaxial anisotropic media (two optic axes)

During the course of this analysis, we shall meet the *index ellipsoid*, which enables us to picture the modes of vibration. We shall also provide a second, and somewhat easier, derivation of the index ellipsoid in terms of the electric displacement vector.

A somewhat non-intuitive consequence of anisotropy is that of *Poynting walk-off*. This is the phenomenon in which the energy of the EM radiation, given in terms of the Poynting vector $\mathbf{S}$, travels in a different direction to the wavevector $\mathbf{k}$. An analysis of this effect followed, giving the angular displacement between $\mathbf{S}$ and $\mathbf{k}$.

12.2 Learning objectives

- Wave propagation in anisotropic media
- Poynting walk-off
- The index ellipsoid
- Birefringence

12.3 Wave propagation in anisotropic media

12.3.1 The wave equation

In the following derivation, we shall find it convenient to write expressions in terms of the *relative permittivity* $\boldsymbol{\varepsilon}$, given by

$$\boldsymbol{\varepsilon} = (\mathbf{I} + \boldsymbol{\chi}_E)\,, \tag{12.1}$$

rather than the susceptibility $\boldsymbol{\chi}_E$. As usual, we confine our consideration to the linear regime.

Assuming that there are no free charges or currents, Maxwell's equations gives

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu_0 \mu \frac{\partial^2 \mathbf{D}}{\partial t^2}, \tag{12.2}$$

where we have made the common assumption that the magnetic permeability may be modelled as a scalar. Since $\boldsymbol{\varepsilon}$ does not depend on t, we may employ the usual vector identity to write Eq. (12.2) as

$$\nabla\left(\nabla \cdot \mathbf{E}\right) - \nabla^2 \mathbf{E} = -\mu_0 \epsilon_0 \mu \boldsymbol{\varepsilon} \frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{12.3}$$

Assuming that we have aligned our Cartesian coordinates with the principal axes of the crystal (see Chapter 11), we may write this out in component form as

$$\frac{\partial}{\partial x_i}\left(\nabla \cdot \mathbf{E}\right) - \nabla^2 E_i = -\mu_0 \epsilon_0 \mu \varepsilon_i \frac{\partial^2 E_i}{\partial t^2}. \tag{12.4}$$

Since we are tacitly in the frequency domain, we may write $\mathbf{E}$ as a plane wave with a complex phase $e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$. Then, noting that $\mu_0\epsilon_0 = 1/c^2$ and putting $\mu\varepsilon_i = n_i^2$, Eq. (12.4) becomes

$$-k_i\left(\mathbf{k} \cdot \mathbf{E}\right) + k^2 E_i = n_i^2 \frac{\omega^2}{c^2} E_i. \tag{12.5}$$

Now $\omega^2/k^2 = v^2$ is the squared magnitude of the phase velocity. We may then use $v^2 = c^2/n^2$ to obtain

$$\left(n^2 - n_i^2\right) E_i - \frac{c^2}{\omega^2} k_i \left(\mathbf{k} \cdot \mathbf{E}\right) = 0. \tag{12.6}$$

This, then, is an eigenvalue problem with the characteristic equation

$$\begin{vmatrix} n^2 - n_x^2 - \kappa_x^2 & -\kappa_x\kappa_y & -\kappa_x\kappa_z \\ -\kappa_x\kappa_y & n^2 - n_y^2 - \kappa_y^2 & -\kappa_y\kappa_z \\ -\kappa_x\kappa_z & -\kappa_y\kappa_z & n^2 - n_z^2 - \kappa_z^2 \end{vmatrix} = 0, \tag{12.7}$$

where we have defined $\kappa_i = ck_i/\omega$.

Although Eq. (12.7) appears to yield a polynomial of order n^6, in fact the n^6 terms drop out due to the relation

$$\kappa_i^2 = n^2 a_i^2, \tag{12.8}$$

where a_i is the ith Cartesian component of the unit vector in the $\mathbf{k}$ direction. Moreover, since all powers of n are even, we obtain a quadratic in n^2

$$an^4 + bn^2 + c = 0, \tag{12.9}$$

where, in the general case,

$$\boxed{a = -\left(n_x^2 a_x^2 + n_y^2 a_y^2 + n_z^2 a_z^2\right),} \tag{12.10}$$

$$\boxed{b = n_x^2 n_y^2 \left(1 - a_z^2\right) + n_x^2 n_z^2 \left(1 - a_y^2\right) + n_y^2 n_z^2 \left(1 - a_x^2\right)} \tag{12.11}$$

and

$$\boxed{c = -n_x^2 n_y^2 n_z^2.} \tag{12.12}$$

12.3.2 Optic axes

The solutions for n^2 correspond to *two modes of vibration*, whereby different components of the polarisation see different refractive indices. These modes of vibration may be visualised by means of the *index ellipsoid*, illustrated in Fig. 12.1. Each principal axis of the crystal is associated with a refractive index n_i. We may then construct an ellipsoid in index space with semi-axes n_x, n_y and n_z.

Let us now consider some arbitrary wavevector $\mathbf{k}$. Taking the intersection of the plane perpendicular to $\mathbf{k}$ with the index ellipsoid defines an ellipse, as shown in Fig. 12.1. The semi-axes of this ellipse give refractive indices n' and n'', which correspond to the two modes of vibration D' and D''.

In general, an ellipsoid has *two* circular cross-sections (see Fig. 12.2). In the case of just two distinct semi-axes, we have a *spheroid* and there is just *one* circular cross-section. The normals to these cross-sections are known as the *optic axes* of the crystal. This then explains the nomenclature of the optical classes

- For *biaxial crystals*, there are *two* optic axes (these are shown as N_1 and N_2 in Fig. 12.2).
- For *uniaxial crystals* have *only one optic axis* (taken, by convention, to be along the z-axis).

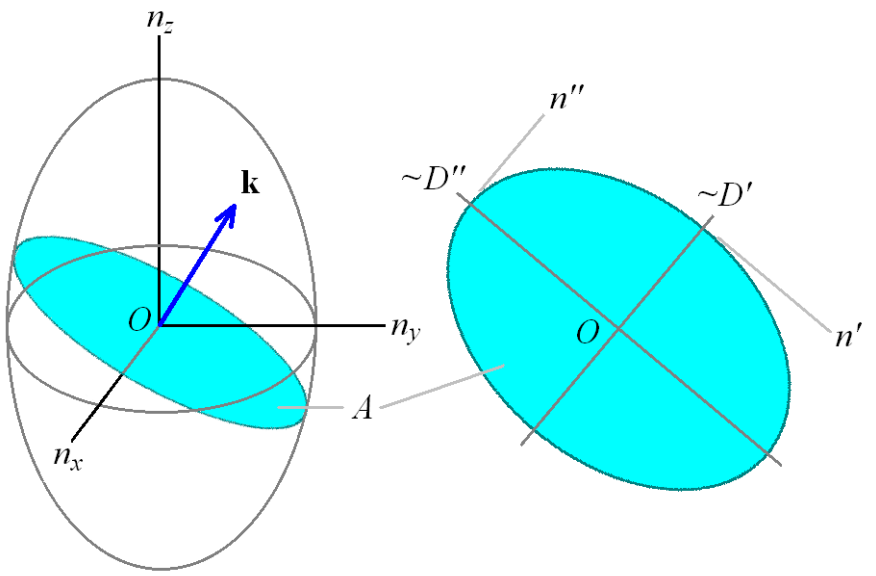


Figure 12.1: The index ellipsoid for an anisotropic material, aligned to the principal axes of the crystal. The plane perpendicular to the $\mathbf{k}$ direction intersects the index ellipsoid in an ellipse. The half-length of the principal axes of this ellipse correspond to the refractive indices, n'' and n', of the two modes of vibration for the optical field.

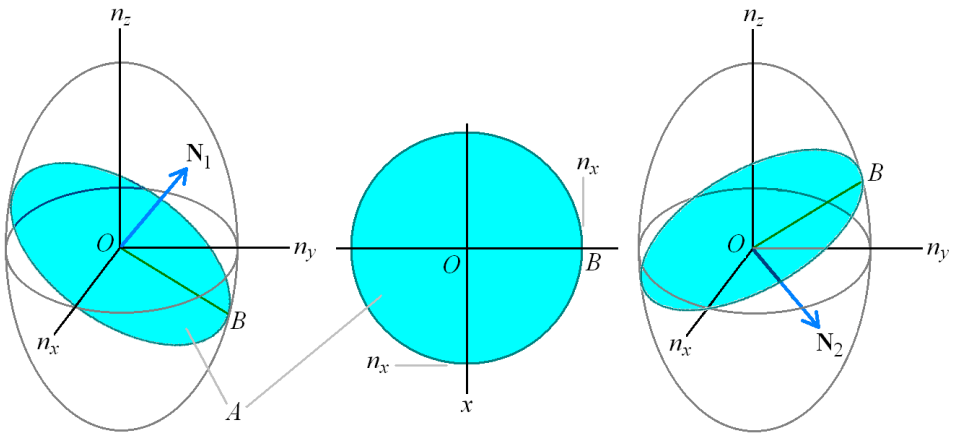


Figure 12.2: The normal to the plane intersecting the index ellipsoid in a circular cross-section is an optic axis of the crystal. In such a direction, the optical field sees only one refractive index and hence there is only one mode of vibration. Illustrated is the general case with *two* optic axes N_1 and N_2.

Uniaxial crystals

For certain special $\mathbf{k}$ directions, Eq. (12.9) will have repeated roots. In these cases, the optical field will only see *one* refractive index and, hence, there will only be one mode of vibration. These directions are known as the *optic axes* of the crystal and are determined by the condition

$$b^2 - 4ac = 0. \tag{12.13}$$

In a uniaxial crystal, we have (by convention) $n_x = n_y = n_o$ and $n_z = n_e$. These are known as the *ordinary* and *extraordinary* refractive indices respectively. The coefficients of the quadratic equation for n^2 given by Eqs.(12.10) to (12.12) are then

$$\boxed{a = -\left[n_o^2 + \left(n_e^2 - n_o^2\right) a_z^2\right],} \tag{12.14}$$

$$\boxed{b = n_o^2 \left[n_o^2 \left(1 - a_z^2\right) + n_e^2 \left(1 + a_z^2\right)\right],} \tag{12.15}$$

$$\boxed{c = -n_o^4 n_e^2} \tag{12.16}$$

and the discriminant is

$$b^2 - 4ac = n_o^4 \left(1 - a_z^2\right)^2 \left[n_o^2 - n_e^2\right]^2. \tag{12.17}$$

For $n_o^2 \neq n_e^2$, the discriminant is zero when $a_z = 1$, i.e. when $\mathbf{k}$ is parallel with the z-axis. Thus, this is the optic axis of the crystal (see Fig. 12.3).

The solutions for n^2 are then

$$\boxed{n^2 = n_e^2 \left(1 + \left[\left(n_e/n_o\right)^2 - 1\right] a_z^2\right)^{-1}} \tag{12.18}$$

and

$$\boxed{n^2 = n_o^2.} \tag{12.19}$$

Hence, we see that one refractive index depends on a_z whilst the other is constant. If $\mathbf{k}$ makes an angle θ with the z-axis, then we have $a_z^2 = \cos^2\theta$ and Eq. (12.18) may be re-written making the θ-dependence explicit as

$$n^2 = n_e^2 \left(1 + \left[\left(n_e/n_o\right)^2 - 1\right] \cos^2\theta\right)^{-1}. \tag{12.20}$$

When $a_z = 1$, Eq. (12.18) gives $n^2 = n_o^2$, so we have only one mode of vibration. On the other hand, when $a_z = 0$, $\mathbf{k}$ lies in the $x - y$ plane and we have the two solutions $n^2 = n_o^2$ and $n^2 = n_e^2$.

These relations are much easier to derive using the *index ellipsoid* discussed in the Section 12.5.

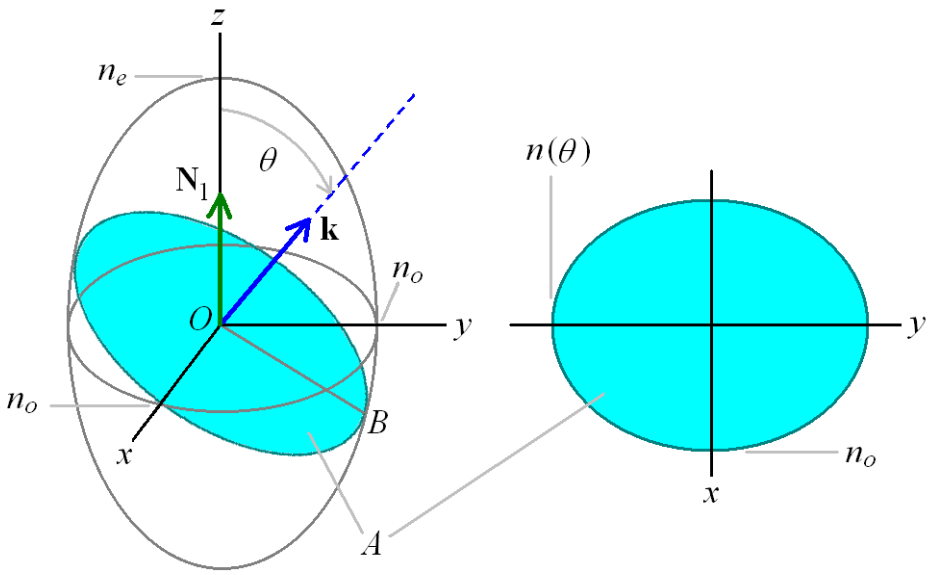


Figure 12.3: The index ellipsoid of a uniaxial crystal. Note that the circular cross-section lies in the $z = 0$ plane. Hence, the optic axis is along the z-axis. For a general $\mathbf{k}$ direction, we find that one of the vibrational modes has a constant refractive index whilst the other has a θ-dependent refractive index.

12.4 Poynting walk-off

We now derive an interesting result, characteristic of wave propagation in anisotropic media. Firstly, we note that the vectors $\mathbf{D}$, $\mathbf{D}$ and $\mathbf{D}$ are *constant* over the wavefront of the propagation. Each must therefore share the same phase factor ϕ. Earlier, we took this to be

$$\phi = \mathbf{k} \cdot \mathbf{r} - \omega t. \tag{12.21}$$

If we apply the effect of the differential operators in terms of wavevector and angular frequency, we have

$$\nabla \cdot \to i\mathbf{k} \cdot, \tag{12.22}$$

$$\frac{\partial}{\partial t} \to -i\omega, \tag{12.23}$$

and

$$\nabla \times \to i\mathbf{k} \times . \tag{12.24}$$

Therefore, in the absence of free charges, applying Eq. (12.22) we have

$$\nabla \cdot \mathbf{D} = i\mathbf{k} \cdot \mathbf{D} = 0. \tag{12.25}$$

This then implies that $\mathbf{k}$ and $\mathbf{D}$ are *perpendicular*.

Now, making the usual assumption that the response of the magnetic susceptibility is negligibly small, the magnetic field $\mathbf{B}$ is given by

$$\mathbf{B} = \mu_0 \mathbf{H} \tag{12.26}$$

and Faraday's Law becomes

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}. \tag{12.27}$$

Using Eqs. (12.24) and (12.23), this becomes

$$i\mathbf{k} \times \mathbf{E} = i\mu_0 \omega \mathbf{H}. \tag{12.28}$$

This means that $\mathbf{H}$ must be *perpendicular* to both $\mathbf{k}$ and $\mathbf{E}$.

Similarly, with no free currents, the Ampere-Maxwell equation becomes

$$i\mathbf{k} \times \mathbf{H} = -i\omega \mathbf{D}. \tag{12.29}$$

Hence, $\mathbf{D}$ is *perpendicular* to both $\mathbf{k}$ and $\mathbf{H}$. Now, since $\mathbf{H}$ is normal to $\mathbf{k}$, $\mathbf{E}$ and $\mathbf{D}$, the latter three *must be coplanar*. However, although $\mathbf{k}$ and $\mathbf{D}$ are normal to one another, in general $\mathbf{E}$ will be at an angle to $\mathbf{D}$ given by

$$\alpha = \cos^{-1}\left[\frac{\mathbf{E} \cdot \mathbf{D}}{|\mathbf{E}|\,|\mathbf{D}|}\right]. \tag{12.30}$$

In the meantime, the energy flow will still be given by the *Poynting vector*

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \tag{12.31}$$

However, since $\mathbf{E}$ and $\mathbf{k}$ are *not*, in general, perpendicular to one another,

- *the Poynting vector is no longer in the* $\mathbf{k}$*-direction*

This phenomenon is known as *Poynting walk-off*, illustrated in Fig. 12.4. In the anisotropic case, it is the direction of $\mathbf{S}$ that gives the direction of a 'ray'.

12.5 The index ellipsoid

The energy density due to the electric field is given by

$$u_E = \tfrac{1}{2}\mathbf{D} \cdot \mathbf{E} = \tfrac{1}{2}\left(D_x E_x + D_y E_y + D_z E_z\right). \tag{12.32}$$

Using Eq. (**??**), this may be re-written as

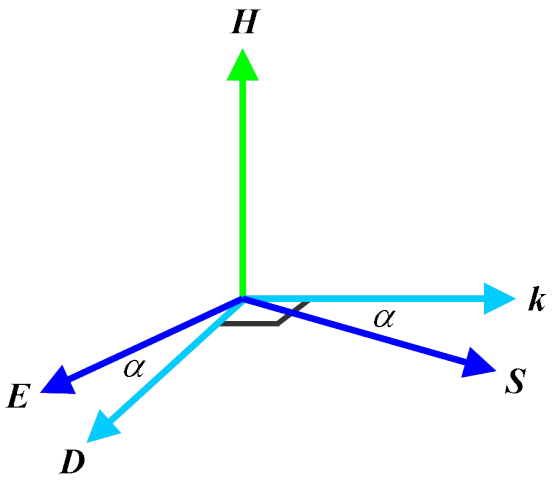


Figure 12.4: Illustration of Poynting walk-off in an anisotropic crystal. Note that the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ is *not* in the same direction as the wavevector $\mathbf{k}$. The angle α between $\mathbf{E}$ and $\mathbf{D}$ is given by Eq. (12.30).

$$u_E = \tfrac{1}{2}\left(\frac{D_x^2}{\varepsilon_0\varepsilon_x} + \frac{D_y^2}{\varepsilon_0\varepsilon_y} + \frac{D_z^2}{\varepsilon_0\varepsilon_z}\right). \tag{12.33}$$

This is the equation of an *ellipsoid* with surfaces of constant energy density. Taking the relative permeability to be unity, Eq. (12.33) may be normalised, putting $x^2 = D_x^2/\left(2u_E\mu_0\right)$ etc., to obtain

$$\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1, \tag{12.34}$$

in terms of the refractive indices associated with each axis. This is known as the *index ellipsoid* and can be used to determine the refractive index that a component of a wave actually sees.

Given the propagation direction $\hat{\mathbf{k}}$, we take the intersection of the plane perpendicular to $\hat{\mathbf{k}}$ and the index ellipsoid. This gives us an ellipse that determines the possible refractive indices that the components of the wave see.

12.5.1 Uniaxial crystals

In the case that two of the axes of the index ellipsoid are the same, we have a uniaxial crystal. By convention, we usually take the equivalent axes to be in the x and y directions. So, putting $n_x = n_y = n_o$ and $n_z = n_e$, we have

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1. \tag{12.35}$$

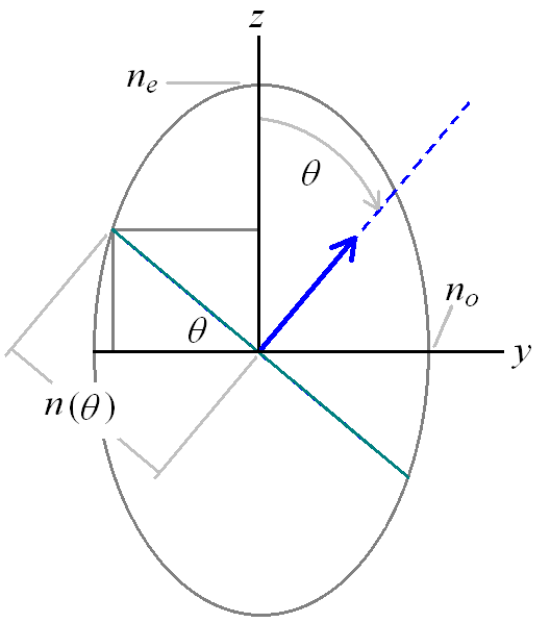


Figure 12.5: The index ellipsoid for a uniaxial crystal projected in 2D. From this, we may find the θ-dependent refractive index from elementary trigonometry and the equation of the index ellipsoid.

The plane perpendicular to $\hat{\mathbf{k}}$ intersects the index ellipsoid in an ellipse. The half length of the axis of this ellipse lying in the x-y plane is n_o, regardless of the direction of $\hat{\mathbf{k}}$. The half length of the other axis is $n_e(\theta)$, which does depend on $\hat{\mathbf{k}}$ (see Fig. 12.3).

Now for any given propagation direction, we may have a component of $\mathbf{D}$ that lies in the x-y plane and thus sees a refractive index of n_o. We call any wave polarised in this direction the *ordinary wave*. Waves polarised in the orthogonal direction within the plane perpendicular to $\hat{\mathbf{k}}$ will see a refractive index $n_e(\theta)$ that depends on the angle θ between the z-axis and $\hat{\mathbf{k}}$. This component is called the *extraordinary* wave.

Since the index ellipsoid has rotational symmetry about the z-axis, we can take $x = 0$ with no loss of generality. We can then find value of $n_e(\theta)$ from the y and z coordinates of the intersection of the ellipse in the z-y plane (see Fig. 12.5).

$$\begin{aligned} x &= 0, \\ y &= -n(\theta)\cos\theta, \\ z &= n(\theta)\sin\theta. \end{aligned} \tag{12.36}$$

Substituting these into Eq. (12.35), we have

$$\frac{n^2(\theta)}{n_o^2}\cos^2\theta + \frac{n^2(\theta)}{n_e^2}\sin^2\theta = 1, \tag{12.37}$$

giving

$$n^2(\theta) = \left(\frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2}\right)^{-1}. \tag{12.38}$$

Using $\sin^2\theta = 1 - \cos^2\theta$, this may be re-written as

$$n^2(\theta) = n_e^2\left(1 + \left[(n_e/n_o)^2 - 1\right]\cos^2\theta\right)^{-1}, \tag{12.39}$$

as found earlier.

12.6 Birefringence

12.6.1 Uniaxial crystal

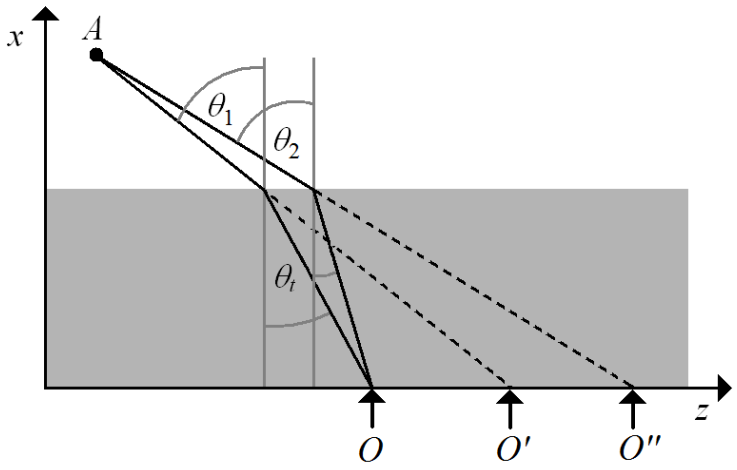


Figure 12.6: Illustration of double refraction in an anisotropic material. The different polarisations of light are refracted by different angles.

The birefringence is defined as

$$\Delta n(\theta) = n(\theta) - n_0. \tag{12.40}$$

For a wave with extraordinary and ordinary components, E_e and E_o respectively, propagating in a direction r, we may write

$$E_e = E_0 \exp\left(i\omega\left[t - \frac{n(\theta)r}{c}\right]\right) \tag{12.41}$$

and

$$E_o = E_0 \exp\left(i\omega\left[t - \frac{n_o r}{c}\right]\right) \tag{12.42}$$

where we have used $k = \omega/v = n\omega/c$. The second of these equations may be re-written

$$E_o = E_0 \exp\left(i\omega\left[t - \frac{n(\theta) r}{c}\right]\right) \exp\left(i\frac{\omega}{c}\left[n(\theta) - n_o\right] r\right). \tag{12.43}$$

Hence, after a distance r, the ordinary wave acquires a retardation

$$\Gamma\left(r\right) = \frac{\omega \Delta n(\theta)}{c} r. \tag{12.44}$$

This then provides the physical mechanism for retardation plates, discussed earlier in Chapter 7.

12.6.2 Double refraction

One consequence of birefringence is the phenomenon of *double refraction*. This occurs where light being transmitted into an anisotropic material at an angle to the normal is refracted by different angles depending on the polarisation. Thus an object underneath the sample will appear to be doubled as shown in Fig. 12.6. An example of this is shown in Fig. 12.7 for calcite, a negative uniaxial crystal.

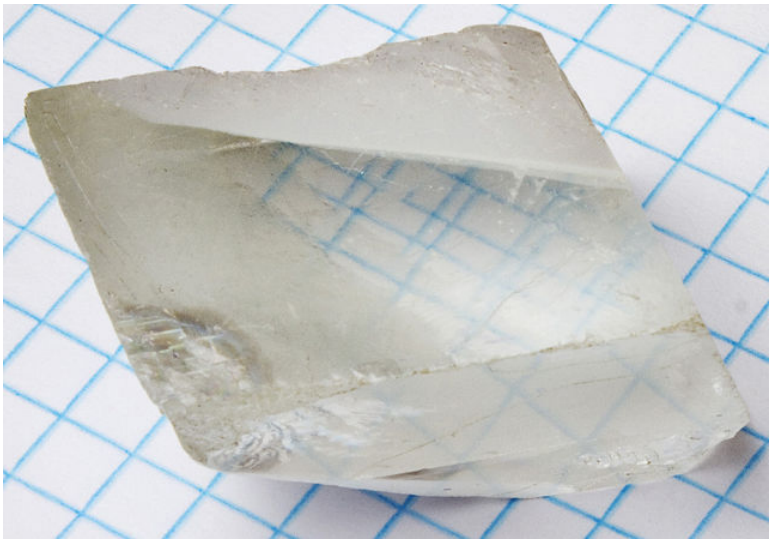


Figure 12.7: Example of double refraction in a sample of calcite, a negative uniaxial crystal.

12.7 Summary

- **Modes of vibration**

  In an anisotropic media there are, in general, *two modes of vibration*. These are orthogonal modes that see two different refractive indices. For wave-vector

$$\mathbf{k} = k\left(a_x \mathbf{e}_x + a_y \mathbf{e}_y + a_z \mathbf{e}_z\right), \tag{12.45}$$

  the refractive indices of these modes are found from the solutions of the quadratic for n^2

$$an^4 + bn^2 + c = 0, \tag{12.46}$$

  where

$$\boxed{a = -\left(n_x^2 a_x^2 + n_y^2 a_y^2 + n_z^2 a_z^2\right),} \tag{12.47}$$

$$\boxed{b = n_x^2 n_y^2\left(1 - a_z^2\right) + n_x^2 n_z^2\left(1 - a_y^2\right) + n_y^2 n_z^2\left(1 - a_x^2\right)} \tag{12.48}$$

  and

$$\boxed{c = -n_x^2 n_y^2 n_z^2.} \tag{12.49}$$

- **Optic axes**

  In any crystal, there is at least one $\mathbf{k}$ direction for which the optical field sees *only one* refractive index (and there is therefore only one mode). Such a direction is called an **optic axis** of the crystal.

  - *Biaxial crystals*
    In biaxial crystals, there are **two** optic axes.
  - *Uniaxial crystals*
    In uniaxial crystals, there is **only one** optic axis.

- **Modes of vibration in a uniaxial crystal**

  In a uniaxial crystal, we have $n_x = n_y = n_o$ and $n_z = n_e$ (note that, by convention, the optic axis is taken to be along the z-axis). The refractive indices of the modes may be found from

$$\boxed{n^2 = n_e^2\left(1 + \left[\left(n_e/n_o\right)^2 - 1\right] a_z^2\right)^{-1}} \tag{12.50}$$

and

$$\boxed{n^2 = n_o^2.} \tag{12.51}$$

Note that

(1) One refractive index is constant, whilst the other depends only on a_z (i.e. the angle between $\mathbf{k}$ and the z-axis)

(2) When $a_z = 1$, we have $n^2 = n_o^2$ for both modes (z is the optic axis)

(3) These expressions may be found far more easily via consideration of the *index ellipsoid*.

- **The index ellipsoid**

  From the energy density for the electric field

$$\boxed{u_E = \tfrac{1}{2}\mathbf{D}\cdot\mathbf{E} = \tfrac{1}{2}\left(D_x E_x + D_y E_y + D_z E_z\right)} \tag{12.52}$$

  we may derive the equation of the **index ellipsoid**

$$\boxed{\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1.} \tag{12.53}$$

  (1) The intersection of the plane perpendicular to $\mathbf{k}$ and the index ellipsoid is an ellipse

  (2) The half-length of the principal axes of this ellipse give the refractive indices of the two modes of vibration

  (3) When $\mathbf{k}$ is directed along an optic axis, the intersection of the perpendicular plane and the index ellipsoid is a circle (i.e. there is only one refractive index)

- **The index ellipsoid for a uniaxial crystal**

  In a uniaxial crystal, the index ellipsoid becomes

$$\boxed{\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1.} \tag{12.54}$$

  Using the index ellipsoid, we may more easily find the refractive indices of the extraordinary and ordinary waves given above.

- **Birefringence**

  The birefringence is defined in terms of the retardation or phase shift acquired between two orthogonal components of light seeing different refractive indices in an anisotropic crystal. Hence, if the retardation is

$$\Gamma(r) = \frac{\omega \Delta n}{c} r, \tag{12.55}$$

  then *the birefringence is the difference in refractive indices* Δn*.*

A. Useful Mathematical Results

A.1 Geometric progression

Consider the series

$$S_n = \sum_{i=0}^{n-1} ar^i. \tag{A.1}$$

Note that each term in the series is increased by a factor r. Such a series is known as a *geometric progression*.

Expanding the summation, we have

$$S_n = a + ar + ar^2 + \ldots + ar^{n-1}. \tag{A.2}$$

Multiplying this by r and subtracting from Eq. (A.2), we have

$$S_n\,(1-r) = a\,(1-r^n) \tag{A.3}$$

or

$$\boxed{S_n = \frac{a\,(1-r^n)}{1-r}.} \tag{A.4}$$

In the case $r \ll 1$ and $n \to \infty$, the r^n term disappears and we are left with

$$\boxed{S_n \to \frac{a}{1-r}.} \tag{A.5}$$

A.2 Matrices

A.2.1 Transpose of a matrix

The operation of *transposition* is usually denoted by a T superscript and involves swapping the columns for the rows of a matrix. That is, for a matrix with elements A_{ij}

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^T = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}. \tag{A.6}$$

More concisely, we can express this as

$$\left(\mathbf{A}^T\right)_{ij} = (\mathbf{A})_{ji}\,. \tag{A.7}$$

A.2.2 Complex conjugation

The operation of *complex conjugaton* of a matrix, usually denoted by a † superscript is similar to that of transposition, except that after taking the transpose, each element is replaced with its complex conjugate. Thus, for a matrix with elements A_{ij}

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^\dagger = \begin{bmatrix} A_{11}^* & A_{21}^* & A_{31}^* \\ A_{12}^* & A_{22}^* & A_{32}^* \\ A_{13}^* & A_{23}^* & A_{33}^* \end{bmatrix}. \tag{A.8}$$

We may write this

$$\left(\mathbf{A}^\dagger\right)_{ij} = (\mathbf{A})^*_{ji}\,. \tag{A.9}$$

A.2.3 Symmetric matrix

A *symmetric matrix* is equal to its transpose. That is, if $\mathbf{A}$ is symmetric, then

$$(\mathbf{A})_{ij} = \left(\mathbf{A}^T\right)_{ij}\,. \tag{A.10}$$

A.2.4 Hermitian matrix

A *Hermitian matrix* is a matrix equal to its complex conjugation. That is,

$$(\mathbf{A})_{ij} = \left(\mathbf{A}^\dagger\right)_{ij}\,. \tag{A.11}$$

Note that if all the elements of this matrix are *real*, then the matrix is *symmetric*.

A.2.5 Rotation matrices

A general rotation in 3D space may be achieved with a combination of three individual rotations. There are several possibilities but we shall consider here the three shown in Fig. A.1. For definiteness, we will take the rotations to be performed in the order $1, 2, 3$. In other words, a general rotation $\mathbf{R}$ is equal to

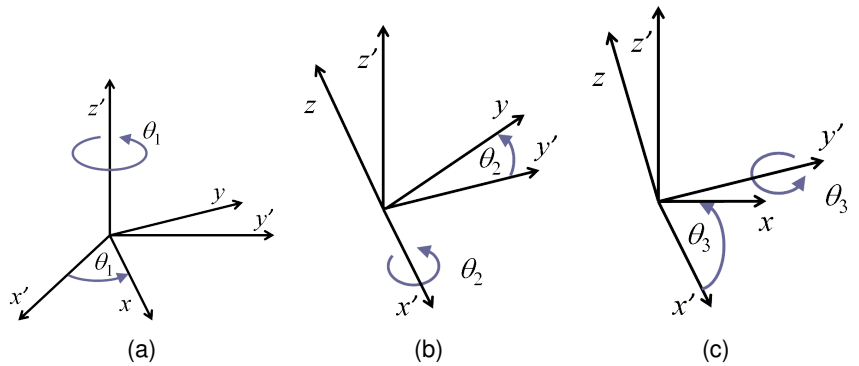


Figure A.1: General rotation matrices in 3D for (a) a rotation around the z axis; (b) a rotation around the x axis and (c) a rotation around the y axis.

$$\mathbf{R} = \mathbf{R}_3\mathbf{R}_2\mathbf{R}_1, \tag{A.12}$$

where

- $\mathbf{R}_1$ is the rotation by θ_1 around the z-axis shown in Fig. A.1 (a)
- $\mathbf{R}_2$ is the rotation by θ_2 around the x-axis shown in Fig. A.1 (b)
- $\mathbf{R}_3$ is the rotation by θ_3 around the y-axis shown in Fig. A.1 (c)

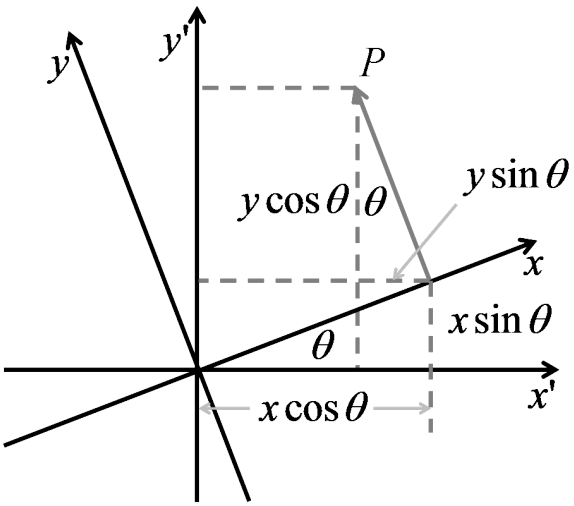


Figure A.2: A general rotation around the z axis of the point P from orginal coordinates (x, y) to new coordinates (x, y)

Considering $\mathbf{R}_1$, it is clear that the z-coordinate is left unchanged. This situation is depicted in Fig. A.2, viewed from along the z-axis. Here we see

the rotation of the point P from orginal coordinates (x, y) to new coordinates (x, y). Note that if this rotation is performed continually as θ goes from 0 to 2π, we would see P rotated around the origin in a complete circle.

From Fig. A.2, we see that the new coordinates are given by

$$\begin{aligned} x' &= x\cos\theta - y\sin\theta, \\ y' &= x\sin\theta + y\cos\theta, \\ z' &= z. \end{aligned}$$

Thus, in matrix form we have

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{A.13}$$

Reintroducing the 1 subscript, we have defined the *rotation matrix*

$$\mathbf{R}_1 = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{A.14}$$

Similarly, we have

$$\mathbf{R}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & -\sin\theta_2 \\ 0 & \sin\theta_2 & \cos\theta_2 \end{bmatrix} \tag{A.15}$$

and

$$\mathbf{R}_3 = \begin{bmatrix} \cos\theta_3 & 0 & -\sin\theta_3 \\ 0 & 1 & 0 \\ \sin\theta_3 & 0 & \cos\theta_3 \end{bmatrix} \tag{A.16}$$

A.3 Vector calculus

A.3.1 Double curl identity

The expression

$$\nabla \times \nabla \times \mathbf{E} \tag{A.17}$$

may be written in component form as

$$(\nabla \times \nabla \times \mathbf{E})_i = \frac{\partial}{\partial x_j} (\nabla \times \mathbf{E})_k - \frac{\partial}{\partial x_k} (\nabla \times \mathbf{E})_j . \tag{A.18}$$

Expanding the $(\nabla \times \mathbf{E})_i$ components, this becomes

$$(\nabla \times \nabla \times \mathbf{E})_i = \frac{\partial}{\partial x_j}\left(\frac{\partial E_j}{\partial x_i} - \frac{\partial E_i}{\partial x_j}\right) - \frac{\partial}{\partial x_k}\left(\frac{\partial E_i}{\partial x_k} - \frac{\partial E_k}{\partial x_i}\right). \tag{A.19}$$

Expanding this and adding and subtracting $\partial^2 E_i/\partial x_i^2$,

$$\begin{aligned}(\nabla \times \nabla \times \mathbf{E})_i &= \frac{\partial^2 E_i}{\partial x_i \partial x_i} + \frac{\partial^2 E_j}{\partial x_i \partial x_j} + \frac{\partial^2 E_k}{\partial x_i \partial x_k} - \frac{\partial^2 E_i}{\partial x_i^2} - \frac{\partial^2 E_i}{\partial x_j^2} - \frac{\partial^2 E_i}{\partial x_k^2}, \\ &= \frac{\partial}{\partial x_i}\left(\frac{\partial E_i}{\partial x_i} + \frac{\partial E_j}{\partial x_j} + \frac{\partial E_k}{\partial x_k}\right) - \frac{\partial^2 E_i}{\partial x_i^2} - \frac{\partial^2 E_i}{\partial x_j^2} - \frac{\partial^2 E_i}{\partial x_k^2}, \\ &= \frac{\partial}{\partial x_i}(\nabla \cdot \mathbf{E}) - \nabla^2 E_i. \end{aligned} \tag{A.20}$$

Hence

$$\boxed{\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}.} \tag{A.21}$$

Index