Universita' Australiana

Testo teorico della LUCE

8 lezioni

con letture , post-letture e domande

LEZIONI

L1 WHAT IS LIGHT ?

OBJECTIVES

Aims

This chapter is essentially an introduction to the wave theory of light. At this stage you should get a basic understanding of the wave model of light, which involves the idea of light as a complex superposition of many component waves, or elementary waves, each with its own wavelength, frequency, amplitude and phase. You should be able to explain these ideas to yourself and to others.

Minimum learning goals

1. Explain, interpret and use the terms:

   wave, elementary wave, wavelength, frequency, period, wave speed, speed of light, refractive index, amplitude of a wave, irradiance, intensity, monochromatic light, spectrum, continuous spectrum, line spectrum.

2. Describe the basic features of the wave model of light.
3. State and apply the relation among wavelength, frequency and speed of a wave.
4. State and use typical values for the wavelengths in vacuum of the components of visible light and for the speed of light in vacuum.
5. Explain the distinction between coherent and incoherent sources and waves.
6. Name the parts of the electromagnetic spectrum and arrange them in order of wavelength or frequency.
7. State and apply the inverse square law for light intensities.

PRE-LECTURE

1-1 INTRODUCTION

There was an ancient belief, which is regularly reinvented by children, that you see something by sending out some kind of probe from your eyes. A more scientific view is that we see things because light comes from them to our eyes. But only a few things generate their own light. Before the middle of the nineteenth century, practically all light came from a few kinds of luminous object - the sun, the stars and fires. So those were the only objects that could be seen by their own light. To see other things we need a luminous object as a source of light. Light travels from the luminous source to the object and then to our eyes. In the process the character of the light may be changed. Some of the so called "white light" from the sun bounces off grass to become "green" light.

Somehow light must also carry information about the location and shape of the objects that we see. We normally assume that a thing is located in the direction where the light comes from. So it would seem that when it is not actually bouncing off something light must travel in more or less straight lines. This idea that light travels through space along straight lines, although not strictly correct, is the basis of the very useful **ray model** of light, which explains a great deal about how we see things. The elements of the ray theory, called **geometrical optics**, will be explored in chapters L2 and L3.

Until the work of Huygens in the late seventeenth century the accepted idea of the nature of light was that it consisted of a flow of invisible corpuscles, like a stream of minute bullets. All the familiar optical phenomena, such as straight line propagation, reflection and refraction could be explained by that corpuscular hypothesis. Although Huygens showed (around 1678) that these phenomena could also be explained by a wave theory, it was the crucial experiments in the nineteenth century by Young and Fresnel on the interference of light which provided convincing evidence that a wave model of light was necessary. Young measured the wavelength of light and its very small value explained why many of the wave properties were so difficult to investigate.

Even after the work of Young not everyone was convinced; it was still possible to explain most of the behaviour of light using the corpuscular idea. Then Foucault found that the speed of light in water was less than its speed in air. On the other hand, the corpuscular theory could explain the bending of a light beam only by supposing that its speed had to be greater in water. So that was the end of the classical corpuscular theory.

A quite different particle theory of light came with quantum theory in the early part of the twentieth century. The current view is that some questions can be answered using a wave model and others can be understood in terms of particles called photons, but the two pictures are never used simultaneously. In this book we need to use only the wave model, while the modern particle model will be used in the *Atoms and Nuclei* unit.

1-2 WAVES

Many kinds of wave carry energy. For mechanical waves which travel in a material medium, such as sound waves, water waves and earthquakes, the energy is mechanical energy - kinetic energy plus potential energy. The potential energy is associated with the forces between particles and their displacements from their equilibrium positions, while the kinetic energy is associated with their movement. The wave energy is propagated through the continual interchange between potential and kinetic energy as the medium oscillates. Electromagnetic waves, on the other hand, can travel through empty space so there is no material medium involved - the energy oscillates between the electric and magnetic fields. Whatever the kind of wave, there are always at least two physical variables associated with its propagation. In the case of sound waves these variables might be the velocity and displacement of particles and in the case of light they are the electric and magnetic fields.

In a material medium sound waves and other kinds of mechanical waves consist of disturbances in some property of the medium. These disturbances move through the medium but the medium itself does not move along with the wave. For example, in mechanical waves (waves on a string, water waves, sound waves) small sections of the medium (the string, the water, the air) vibrate to and fro, but there is no net flow of material from one end of the medium to the other. For example a wave on a string might look like figure 1.1; the string oscillates up and down and energy flows along with the wave but there is no movement of matter along the string.

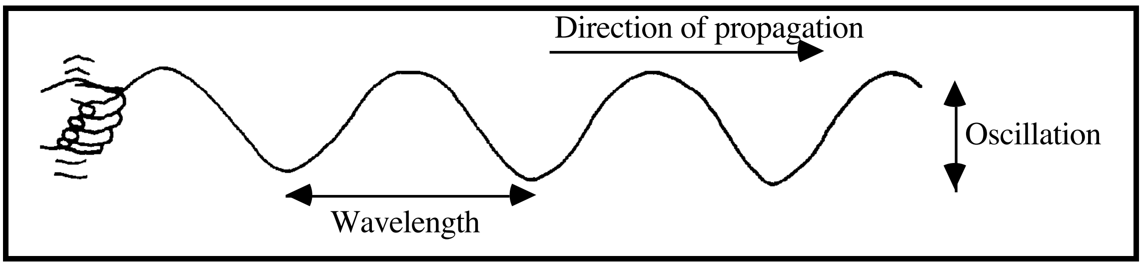


Figure 1.1. A transverse wave on a string

A complete description of an ordinary beam of light using the wave model would be immensely complex. Even water waves on the surface of the sea can be very intricate. But it is not necessary to go into detail about that complexity if all you want to do is understand the underlying principles of wave motion and behaviour. The mathematical theory of waves includes the very useful principle that any complex wave at all can be represented as the sum, or superposition, of simple harmonic waves; so all the fundamental properties of waves are expressed in terms of the behaviour of simple harmonic waves. Figure 1.2 shows an example of a relatively uncomplex wave which can be analysed as a combination of only four elementary waves.

Elementary waves

The simplest kind of wave to describe mathematically is a simple harmonic wave that travels in one direction. The wave property (electric field, pressure or whatever it is that does the waving) is represented here by W and varies with position x in space and with time t. The wave can be described by the equation:

$$W = A\sin(kx - \omega t + \phi) \qquad \text{... (1.1)}$$

in which A, k, ω and ϕ are constants. Their significance is discussed below.

This equation tells us several things about the wave. The expression in parentheses, $(kx - \omega t + \phi)$, which is called the **phase** of the wave, tells what stage the oscillation has reached at any point x and time t. The quantity ϕ is called the initial phase. We can get a kind of snapshot of the wave by making graphs of W plotted against x for particular values of the time t (figure 1.3). The graphs show the familiar sine-curve shape of the wave. The constant A is called the **amplitude** of the wave and the value of the wave property varies between $-A$ and $+A$. As time progresses the wave moves forward, but its shape is the same.

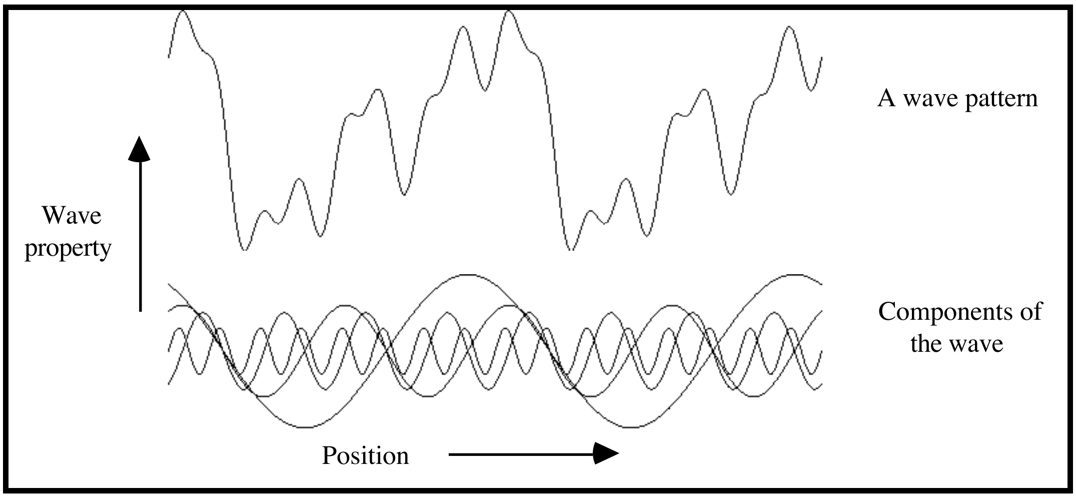


Figure 1.2. Analysis of a wave in terms of elementary sine waves
The complex wave is plotted in the top diagram and the mixture of its four components is shown below.

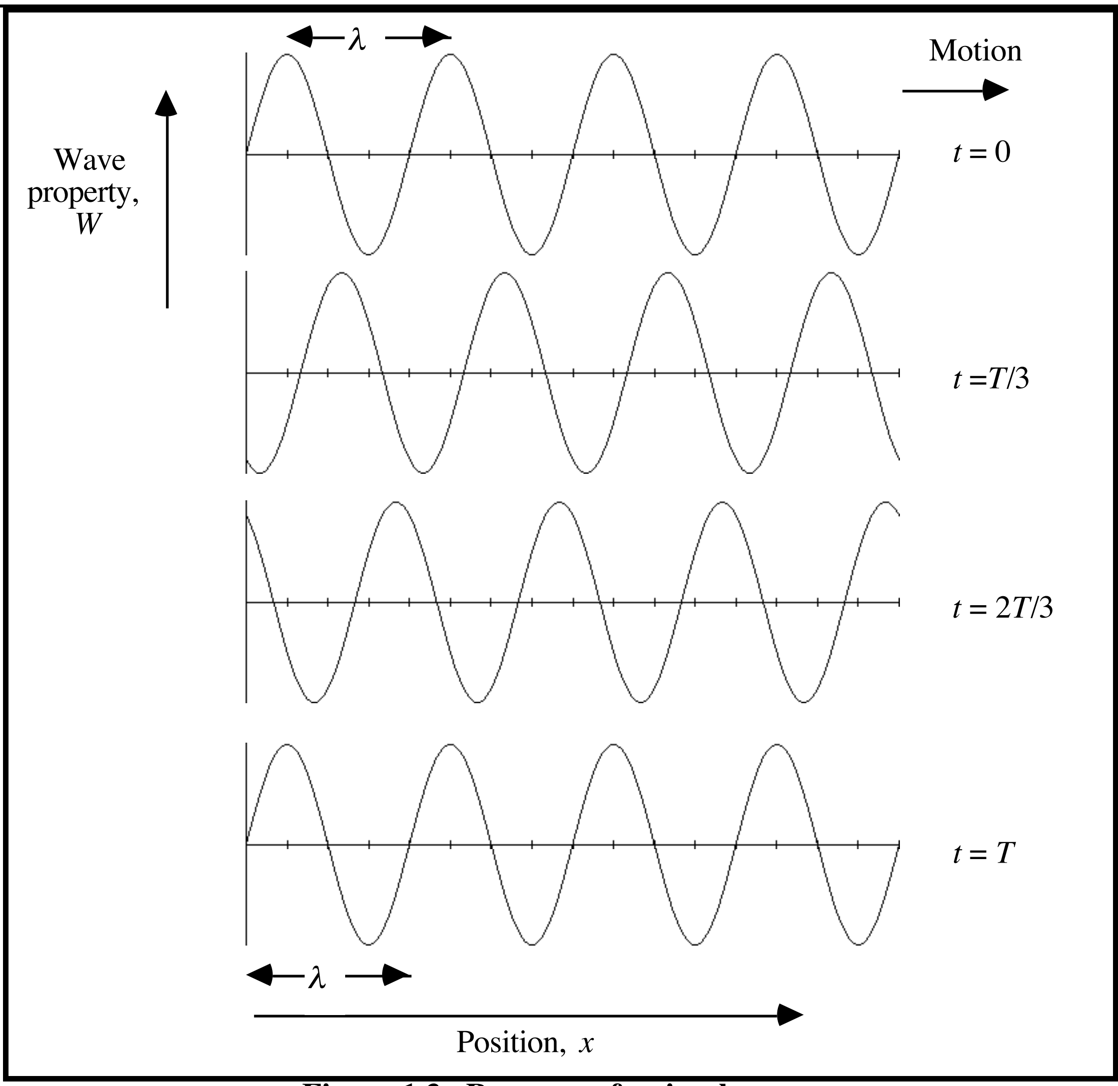


Figure 1.3. Progress of a simple wave
The whole wave pattern moves to the right. In one period (T) it moves one wavelength (λ).

The equation (1.1) and the graphs (figure 1.3) both show that the pattern of the wave is repeated exactly once every time that the position coordinate x increases by a certain amount λ, which is called the **wavelength**. The constant k in equation (1.1) is called the **propagation constant** or the wave number. It is inversely related to the wavelength:

$$\lambda = \frac{2\pi}{k}.$$

By looking at what happens at a fixed point (x) as the wave goes past, we can see that the variation of the wave property with time is also described by a sine function: the variation of W is a simple harmonic oscillation (figure 1.4).

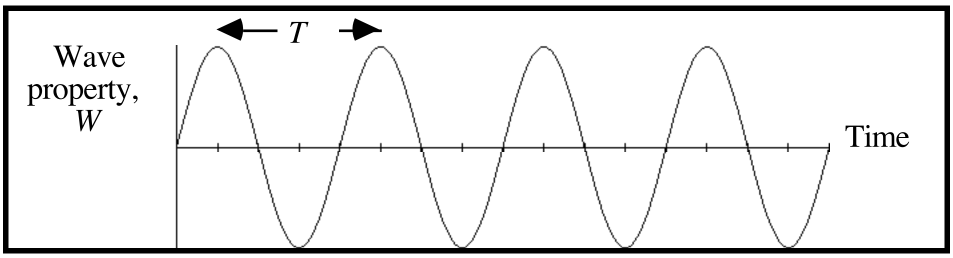


Figure 1.4. The wave oscillation at a fixed location
Although W is now plotted against time, the shape of the graph is just like the shape of the wave shown in figure 1.3 .

The constant ω in the equation is the **angular frequency** of the oscillation and the wave. The wave's **period** T and its **frequency** f are given by the relations:

$$T = \frac{2\pi}{\omega} = \frac{1}{f} . \qquad \text{... (1.2)}$$

By studying the graphs in figure 1.3 you should be able to satisfy yourself that the wave moves forward by one wavelength in one period, so the **wave speed** must be equal to λ/T or $f\lambda$:

$$v = f\lambda . \qquad \text{... (1.3)}$$

Note that the wave equation quoted above describes the progression of an idealised wave in a one-dimensional space. The main differences for real waves in three-dimensional space are that the amplitude A generally decreases as the wave moves further away from its source and that we need some way of describing how the waves spread out as they go.

LECTURE

1-3 LIGHT WAVES

In the wave model light is viewed as electromagnetic waves. Since these waves consist of oscillating electric and magnetic fields which can exist in empty space, light can travel through a vacuum.

Since light can be analysed as a complex mixture of a huge number of individual electromagnetic waves, the important properties of light and other electromagnetic waves can therefore be understood in terms of the properties of these simple **elementary waves**.

At any point on the path of a simple harmonic light wave the strengths of the electric and magnetic fields are continually changing. At each point the two fields always change in step, so that the maximum value of the electric field occurs at the same time as the maximum magnetic field. The electric and magnetic fields point in directions at right angles to each other and also at right angles to the direction in which the wave travels. Since a complete knowledge of the electric field determines the magnetic field, the wave can be described adequately by specifying the electric field only.

Figure 1.5 is an instantaneous representation of the fields in part of an elementary electromagnetic wave. Notice that the electric and magnetic fields are in phase, their maxima occur at the same place at the same time. Since both fields are perpendicular to the direction of travel of the wave, the wave is said to be **transverse**. (A wave in which the direction of the wave property is parallel to the direction of travel is called a longitudinal wave.)

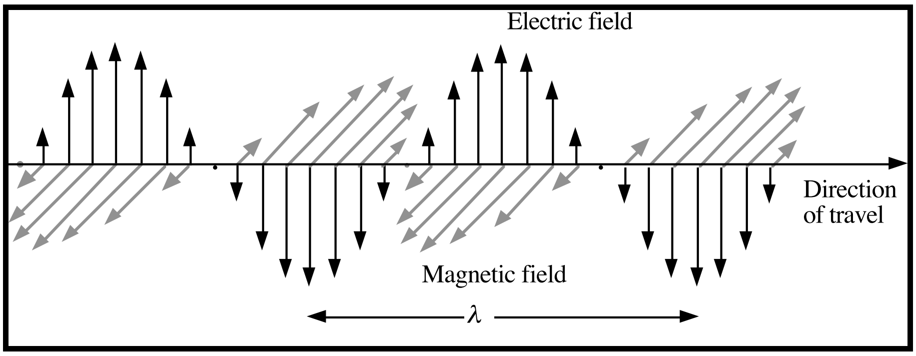


Figure 1.5. Instantaneous plot of part of a simple electromagnetic wave
The direction of each field is shown by the direction of the arrow and its magnitude is represented by the length of the arrow.

Wavelength and frequency

An important property of electromagnetic waves is that in empty space they all travel at exactly the same speed of about 300 000 kilometres per second ($2.997\,924\,58 \times 10^8$ m.s^{-1} to be more precise) quite independently of their wavelength and frequency.

The quantities which characterise each elementary wave are its amplitude, its frequency and its wavelength. Amplitude and frequency are difficult or impossible to measure directly but there are several kinds of experiment which can be used to measure wavelength. Experiments have yielded values for the wavelengths of visible light which lie roughly in the range, 400 nm to 700 nm. The usual unit for light wavelengths, which is consistent with SI, is the nanometre; 1 nm = 1×10^{-9} m. (In older literature you may find reference to two obsolete units. The angstrom, symbol Å, is 10 nm and the micron is equivalent to the micrometre, μm.)

Since the speed of light in vacuum is fixed, each wavelength corresponds to a different frequency. The range of frequencies for visible light is from about 7×10^{14} Hz (at 400 nm wavelength) to about 4×10^{14} Hz (at 700 nm). When the wave theory of light is extended to take account of light's interaction with matter, it turns out that when an elementary light wave goes from one material into another its frequency is unchanged but the speed and the wavelength are altered. So the property which really distinguishes each elementary wave is its frequency, rather than its wavelength. The common practice of describing light in terms of wavelengths is related to the fact that wavelengths can be measured reasonably directly but frequencies are to hard to measure. Since wavelength changes what does it mean to quote values for wavelength? The answer is that unqualified references to wavelength are understood to mean wavelength in vacuum, or possibly air. (Fortunately wavelengths of the same wave in air and vacuum are almost equal.)

Light which contains a relatively narrow range of wavelengths looks coloured. The colours correspond to those in the rainbow, ranging from violet (upwards of 400 nm) through blue, green (around 550 nm) and yellow, to red (up to about 700 nm). Normal sunlight, which contains the whole range, is usually described as white light.

Speed of light and refractive index

The speed of light in a transparent material is always less than the speed, c, in vacuum. The ratio of the speed in a vacuum to the speed in the medium is called the **refractive index** (n) of the medium.

$$n = \frac{c}{v} \quad . \qquad \text{... (1.4)}$$

Medium	**Speed** $\dfrac{v}{10^8 \text{ m.s}^{-1}}$ for $\lambda = 589.3$ nm	**Refractive index** $n = \dfrac{c}{v}$ for $\lambda = 589.3$ nm
Vacuum	2.998	1.0000
Air	2.997	1.0003
Water	2.249	1.333
Glass	1.972	1.520
Diamond	1.239	2.419

When any elementary electromagnetic wave, including light, passes from one medium into another, its frequency remains the same. This can be explained in terms of the interaction between the radiation and the electrons in the material. The electromagnetic waves actually interact with the atoms or unbound electrons which then re-radiate the energy, forming a new wave at the same frequency.

Polarisation

Have another look at figure 1.5 and notice that the directions of the electric fields are all parallel or anti-parallel; they all lie in the same plane. Hence the wave is said to be **plane polarised** or

linearly polarised. (Similarly, note that the magnetic field vectors all lie in a common plane, which is perpendicular to the plane of the electric field.)

In ordinary light, which is a complex mixture of elementary waves, the only restriction on the plane of vibration of the electric field is that it should be at right angles to the direction of travel of the light wave. Otherwise it can have any orientation. Consider radiation from an ordinary light globe. The total electric field at a particular place (due to the radiation from all parts of the filament) changes direction quite randomly but always stays perpendicular to the direction of travel of the light wave. Light waves which behave like this are said to be **randomly polarised** or **unpolarised** (figure 1.6).

However, if by some means the electric field is restricted to one plane only, i.e. if the individual elementary waves all have the same polarisation, then the light beam as a whole is said to be plane polarised or linearly polarised**.**

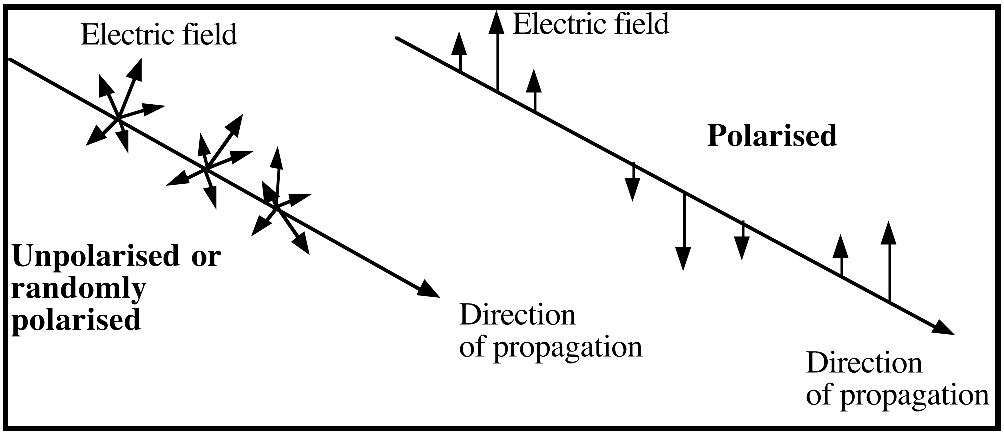


Figure 1.6. Polarised and unpolarised light

1-4 DETECTING LIGHT

Light detectors respond in many different ways. For example light entering a light meter produces an electric current which deflects the pointer of the meter. And light interacting with a photographic plate causes a chemical change in the emulsion which gives a permanent record of the incident light pattern.

Most kinds of continuously operating light detectors respond to the rate at which the light's energy is absorbed by the detector; they indicate the power. How is this response related to the electric field of the light waves? No detector can respond to the instantaneous value of the field because the field changes far too rapidly, so the response must be to some kind of average of the field over time. A detector which responded simply to the time-averaged value of the electric field itself would be useless, because that average value is zero. On the other hand most detectors respond to the time average of the *square* of the field's value, i.e. to E^2. This can be related to the rate at which waves deliver energy by recalling that the energy in a simple harmonic motion is proportional to the square of its amplitude (chapter FE7). In the case of an elementary light wave with amplitude E_0 the rate of energy transfer is proportional $E_0{}^2$, the square of the amplitude, which is also equal to the average value of E^2.

Other factors which affect the response of a detector are its receiving area (the bigger it is the more light it collects) and the spectral composition of the light - i.e. the distribution of the light's power over the various wavelengths or frequencies of the light.

The eye

The human eye is sensitive to light with wavelengths between about 400 nm and 700 nm. It is this sensitivity that makes this part of the electromagnetic spectrum so special to us. The eye is more sensitive to some frequencies than to others (figure 1.7). For example the eye is about seven times more sensitive to green light at 550 nm than it is to blue light at 480 nm. So a beam of blue light would need to be seven times as powerful as a similar green beam for the two beams to appear equally bright.

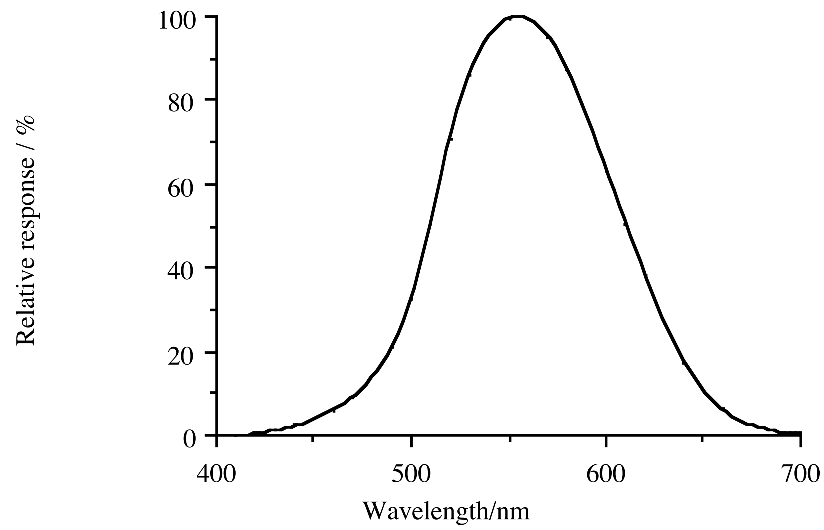


Figure 1.7. Sensitivity of the eye

Irradiance

Since light carries energy we need a way of describing that. Imagine a small flat area of space which is perpendicular to the direction of travel of the light. For a given light beam the power (rate of energy transfer) passing through this small region is proportional to its area; the larger the area the more energy it receives. The relevant property of the light is then the power divided by the perpendicular area; or more precisely the limiting value of that quotient as the area is made smaller. Strictly this quantity should be called **irradiance** but it is commonly known as the **intensity** of the light.

For a harmonic electromagnetic wave the irradiance is proportional to the time-averaged value of E^2.

1-5 IRRADIANCE OF LIGHT FROM A NUMBER OF SOURCES

We now consider how to model the resultant irradiance when light from different sources arrives at the same place. The result depends on the relationship, or lack of relationship, between the phases of the elementary waves in each complex wave. In principle there is only one correct way of doing the calculation: at each instant of time find the total electric field by adding all the individual fields in both beams, taking proper account of their many different directions. Then the square of the total field will be proportional to the instantaneous intensity. In reality however, we are more interested in values averaged over reasonable time intervals (a few milliseconds for example) rather than instantaneous values and in such cases a simpler procedure will give accurate answers.

In most cases the irradiance of light produced at some place by several different independent sources can be found by adding the irradiances from the individual sources. As an example consider two light globes. For each globe the total light output is made up of many small contributions from the large number of atoms in the hot lamp filament. Each atom emits radiation in short bursts which occur at random times; excited atoms emit light quite independently of one another. The light from each globe therefore consists of a complex mixture of many elementary waves with different frequencies, phases and polarisations. Although both light beams contain much the same mixture of frequencies, the phases and polarisations of the elementary waves in the two beams do not match up. Even if we consider a specific frequency, the phases of the elementary waves from one globe are quite random, so they are not related in any way to the phases of the elementary waves from the other globe. The two light sources and the waves which come from them are said to be **incoherent**.

For two incoherent sources A and B, the total irradiance at some place, due to both sources together, can be found from the sum of the irradiances due to each source alone:

$$I_T = I_A + I_B .$$

On the other hand, if there is a definite fixed relationship between the phases and polarisations of the waves from the two sources this procedure gives the wrong answers. For a somewhat artificial example think of two pure, very long, harmonic waves with exactly the same frequency. Suppose that we look at a place in space where these two waves meet with their polarisations parallel. If the two waves are exactly in phase (in step) the amplitude of the total field will be the sum of the two individual amplitudes and if they are half a cycle out of step the resultant amplitude will be the equal to the difference in the individual amplitudes. If they are in step the irradiance will be given by

$$I_T \propto E_{0T}^{\ 2} = (E_{0A} + E_{0B})^2 ,$$

but if they are exactly out of step the irradiance will be

$$I_T \propto E_{0T}^{\ 2} = (E_{0A} - E_{0B})^2$$

This is certainly not the same result as we would get by adding the separate irradiances that would have been produced by each each wave in the absence of the other; i.e. $I_T \neq I_A + I_B$.

Now think of two sources of light which emit a complex mixture of elementary waves, but this time suppose that the mixture of light emitted by one of them is an exact copy of the collection of elementary waves emitted by the other. We can pair off the elementary waves and apply the argument about adding the fields. Once again the irradiances that the beams would have produced individually do not add; we must add the fields and then take the appropriate time-averages if we want to know the irradiance. In this case the two sources and the waves from them are said to be **coherent**. There is a definite relationship between the phases of the elementary waves in the two complex waves.

Example

Light from four identical incoherent sources arrives at the same place, having travelled the same distance from each. The total irradiance is the sum of the four individual irradiances:

$$I_T = I + I + I + I = 4\,I$$

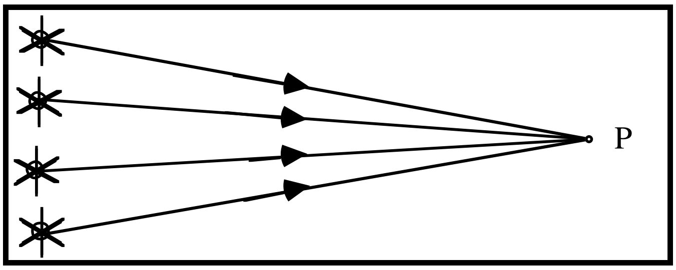


Figure 1.8. Irradiance due to four identical sources
The point P at which the irradiance is measured is equidistant from the sources.

If on the other hand the four identical sources are coherent and if the polarisations match up we could get all matching components from the two sources to arrive exactly in phase, so the amplitudes would add up and the resulting irradiance would be

$$I_T = (E_0 + E_0 + E_0 + E_0)^2 = 16\,I .$$

You may think that this result violates the law of conservation of energy. That is not so, because there are other places where the contribution to the total irradiance from the same elementary waves is quite small. The total energy is the same in both examples, it is just distributed differently.

The combination of coherent waves is called **interference**, a topic which will be discussed further in chapter L4.

1-6 SPECTRA FROM SOURCES OF VISIBLE LIGHT

Because of the short wavelengths (about 10^{-7} m) and high frequencies (about 10^{14} Hz) of light waves we can infer that light radiation must be emitted by something small such as the atoms and electrons of the material that forms the source of the light wave. Quantum theory describes how isolated atoms can radiate only at those frequencies which correspond to a particular change from one well-defined atomic energy level to another. The frequency of the emitted wave is given by the formula

$$f = \frac{\Delta E}{h} \qquad \text{... (1.5)}$$

where ΔE is the energy change and h is Planck's constant. For more about this topic see the *Atoms and Nuclei* unit.

However, in a solid the atoms are packed so closely together than there is considerable interaction among them. This leads to a blurring out of the energy levels into a continuous band of energies. A **continuous spectrum** of light frequencies results. More atoms are excited as the temperature of the material increases. Thus a hotter object emits a greater total intensity of electromagnetic waves. Also, as the temperature of a body is increased, it emits a greater proportion of its radiation at higher frequencies (shorter wavelengths). So, as the temperature is increased, the peak irradiance in the spectrum moves to shorter wavelengths.

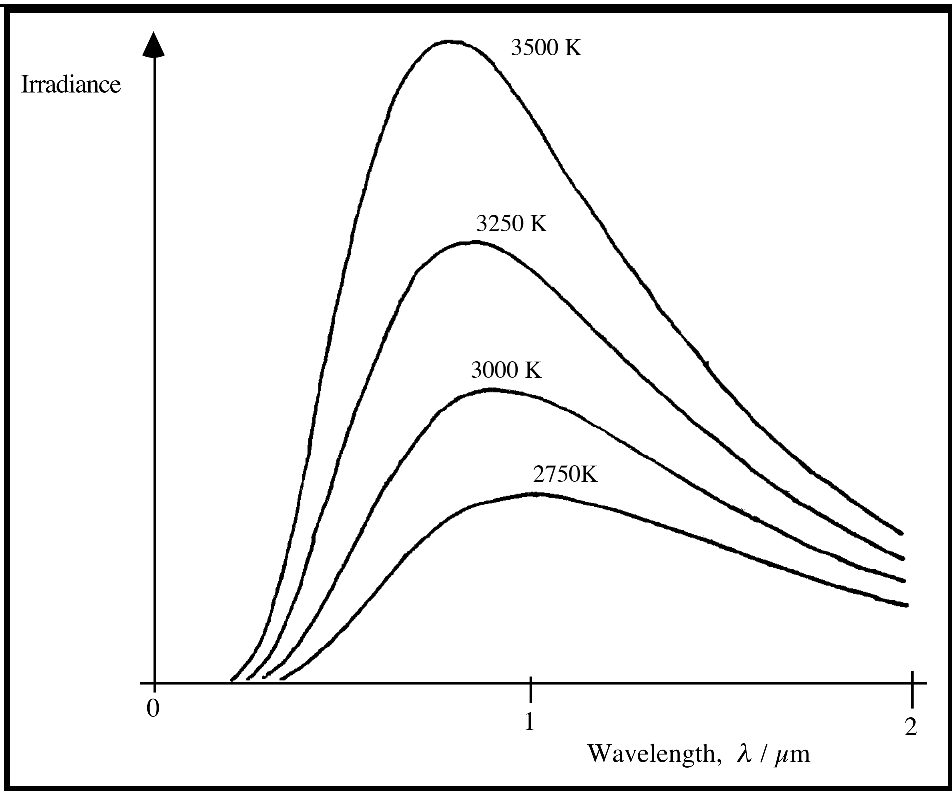


Figure 1.9. Spectra of light from a hot solid
The peak of the continuous spectrum shifts to shorter wavelengths as the temperature is increased. The intensity of radiation also increases with increasing temperature.

In general, gases and vapours, in which the atoms or molecules are well separated, emit line spectra. Every atom or molecule has a characteristic **line spectrum** corresponding to its energy level structure so the spectrum observed depends on the types and numbers of different atoms and molecules present.

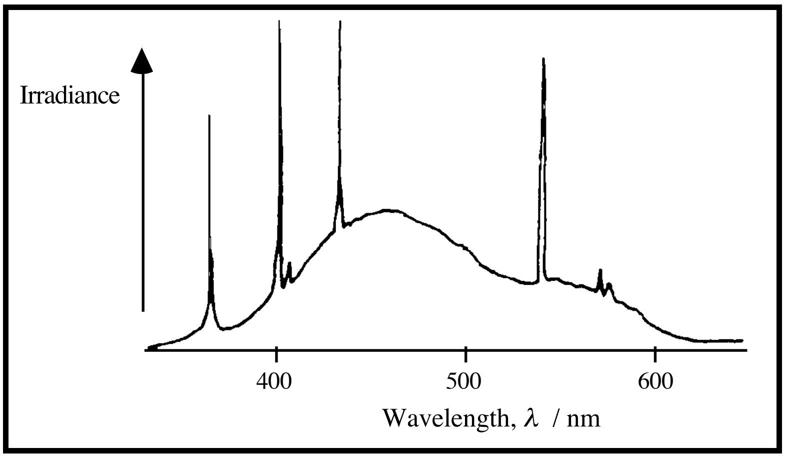


Figure 1.10. Spectrum of light from a fluorescent tube
Note the bright spectral lines from the gas superimposed on the continuous spectrum from the fluorescent solid coating on the inside of the tube.

A laser emits radiation in a very narrow range of wavelengths. Such light is called **monochromatic**.

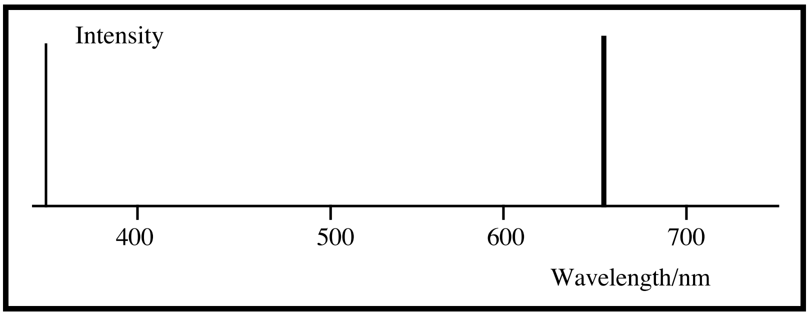


Figure 1.11. The spectrum of radiation from a laser

POST-LECTURE

1-7 THE ELECTROMAGNETIC SPECTRUM

The spectrum of electromagnetic waves is divided up into a number of arbitrarily named sections. The dividing lines between these sections are determined by the detailed properties of a particular range of wavelengths. But there is considerable overlap and the divisions are to some extent arbitrary.

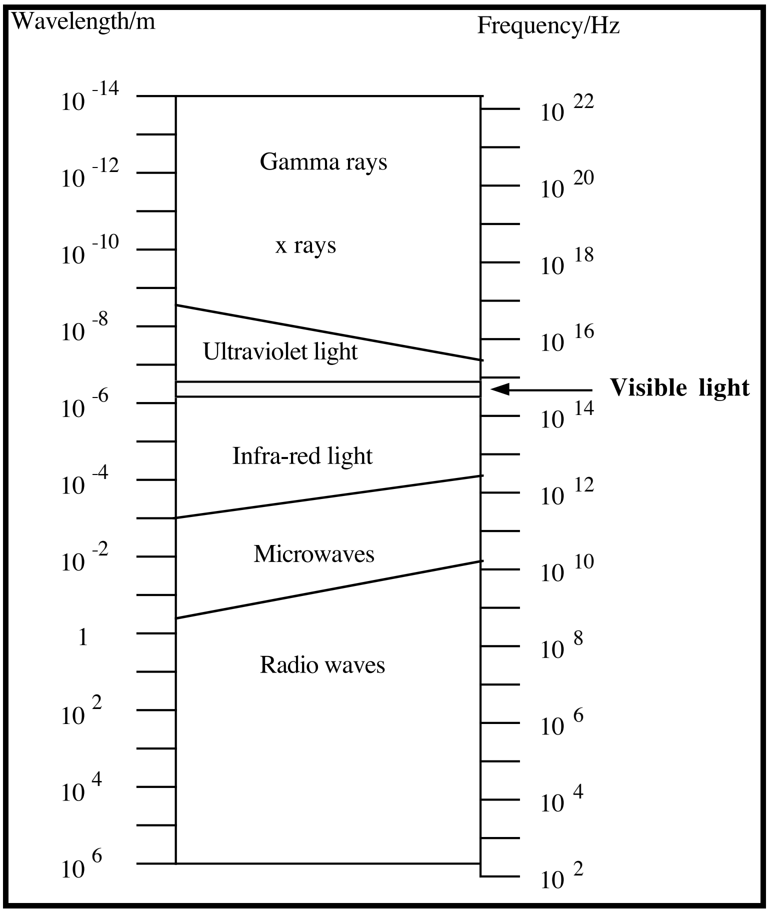


Figure 1.12. The spectrum of electromagnetic waves
Note the logarithmic scales.

Radio waves

Radio waves have wavelengths from about 1 m upwards. They are produced by connecting an electronic oscillator to an antenna. The oscillating electrons in the antenna then lose energy in the form of electromagnetic waves. Radio waves are used for radio and television broadcasting and long-distance communications.

Microwaves

Microwaves are short radio waves with wavelengths down to about 1 mm. They can be produced electronically by methods analogous to the production of sound waves when you blow across the top of a resonating cavity such as a bottle. Because microwaves are not absorbed very strongly by the atmosphere, but are reflected well off solid objects such as buildings and aircraft, they can be used for radar location of distant objects. Microwaves are also used extensively for communications but they require direct line-of-sight paths from transmitter to receiver so that microwave stations are located on top of hills and tall structures.

Infrared radiation

The infrared part of the spectrum comprises wavelengths from 0.1 mm (far-infrared) down to about 700 nm. Infrared radiation is emitted by excited molecules and hot solids. Much of the energy released by the element of an electric oven is in the form of infrared radiation. The radiation is very easily absorbed by most materials so the energy becomes internal energy of the absorbing body. When you warm your hands by a fire you are absorbing infrared radiation.

Visible light

Light is that part of the electromagnetic spectrum which we can see. Visible light is emitted by excited atoms and molecules and by very hot solids.

Ultraviolet radiation

Ultraviolet 'light' has wavelengths less than 400 nm. It is emitted by excited atoms. The 'black light' used to produce fluorescence in light shows is ultraviolet. Much of the ultraviolet radiation from the sun is absorbed by the atmosphere but that which gets through can cause sunburn and skin cancers. Ultraviolet light can also be harmful to the eyes. The irradiance of ultraviolet light increases at high altitudes where the atmosphere is thinner. Part of the concern about the depletion of the atmosphere's ozone layer is based on the fact that the ozone layer absorbs ultraviolet radiation from the sun.

X rays and gamma rays

The wavelengths of x rays and gamma rays overlap, but the different names indicate different ways of producing the radiation. X rays are produced in processes involving atoms and electrons. For example they can be produced by bombarding a metal target with high energy electrons. They are also emitted in some high-energy atomic energy level transitions. X rays usually have wavelengths less than 10 nm. On the other hand the term gamma rays is reserved for electromagnetic radiation emitted in sub-atomic processes such as the decay of excited nuclei or collisions between sub-nuclear particles. Gamma radiation generally has wavelengths less than 0.1 nm. It is emitted by excited nuclei of atoms.

1-8 THE INVERSE SQUARE LAW FOR LIGHT

Take a point source of light which is radiating uniformly in all directions and consider a sphere of radius r centred on the source. The total light power, P, radiated by the source must pass through this sphere. Irradiance of radiation is defined as the power per area, which strikes (or passes through) a surface which is perpendicular to the direction of propagation.

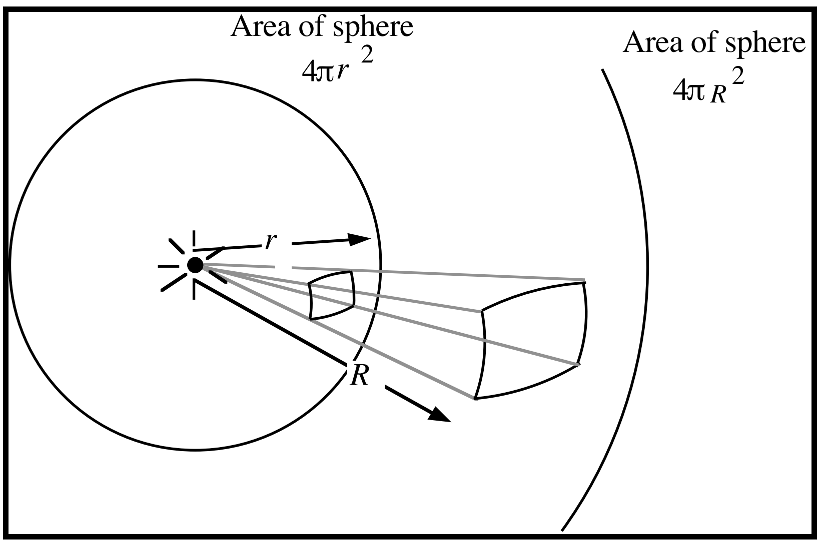


Figure 1.13. Inverse square law for light

In this case, since the energy is distributed uniformly over the surface of a sphere, so

$$I = \frac{\text{total power}}{\text{total area}} .$$

At distance r

$$I = \frac{P}{4\pi r^2} . \qquad \text{... (1.6)}$$

At a larger distance R,

$$I = \frac{P}{4\pi R^2} \text{, which is smaller.}$$

The irradiance is inversely proportional to the square of the distance from the point source.

QUESTIONS

Exercises

Q1.1 Calculate the distance travelled by light in 1.0 μs.

Q1.2 A typical wavelength for visible light is 500 nm.

a) Calculate the frequency of this light.

b) Calculate the wavelength, frequency and speed of this light in a glass with refractive index 1.50.

Q1.3 Calculate the irradiance of the light coming from three identical sources all at the same distance

a) when the three sources are incoherent;

b) when the three sources are coherent and the fields have the same polarisation and phase.

Q1.4 On the large diagram of the electromagnetic spectrum mark the wavelengths of the following sources. You may have to do some searching for the answers.

a) Radio station 2GB.

b) TV Channel 2

c) A green spectral line.

d) X rays used by a radiographer.

e) The range of wavelengths of an electric radiator as it warms up to red heat.

f) A gamma ray.

Q1.5 Suppose that a point source is radiating light waves at a rate of 10 W. Calculate the irradiance at a distance of 20 m from the source.

Q1.6 Refer to the sensitivity curve for the eye, figure 1.7. At what wavelength does a normal human eye have maximum sensitivity? At what wavelengths does it have half its maximum sensitivity? At what wavelengths does it have only 1% of its maximum sensitivity?

Discussion questions

Q1.7 Give some scientific arguments against the view that we see things by sending some kind of probe out from our eyes.

Q1.8 How could you measure the sensitivity curve for the human eye ?

Q1.9 The eye detects the visible part of the electromagnetic spectrum. The human body is also affected by radiation in other parts of the electromagnetic spectrum. How?

Q 1.10 People used to do experiments to measure the speed of light. But the metre is now defined in terms of the speed of light. Does this mean that those experiments are no longer useful? Discuss.

Q1.11 Which of the following affect the speed of light in vacuum: (a) speed of the source, (b) speed of the observer, (c) intensity of the light (d) wavelength, (e) frequency ?

Q1.12 Why does a microwave oven cook the chicken but not the plate?

Q1.13 A photographic plate and a radio set both operate as detectors of electromagnetic waves. Yet they are not interchangeable. Comment.

L2 REFLECTION AND REFRACTION

OBJECTIVES

General aims

When you have finished studying this chapter you should understand the nature of reflection and refraction of light and the simple laws which govern those processes. You will learn how to use the ray model for describing the behaviour of light and you should be able to apply the model to simple examples. Also, you will learn to describe dispersion, the process responsible for rainbows.

Minimum learning goals

1. Explain, interpret and use the terms:

   wavefront, spherical wavefront, plane wavefront, ray, point source, scattering, reflection, reflectivity, specular reflection, diffuse reflection, refraction, refractive index, Snell's law, internal reflection, total internal reflection, critical angle, grazing incidence, dispersion, spectrum, optical fibre, light pipe.
2. State the laws of reflection and refraction, describe examples and apply the laws to simple examples involving plane boundaries.
3. Describe partial and total reflection. Derive, recall and apply the relation between critical angle and refractive indices.
4. Describe what happens to speed, frequency and wavelength when monochromatic light goes from one medium to another. Apply these descriptions to simple quantitative problems.
5. Describe the phenomenon of dispersion and its explanation in terms of refractive index and the wave model of light. Describe examples which illustrate dispersion by refraction.
6. Remember that the speed of light in air is practically equal to its speed in vacuum.
7. Describe and explain the operation of optical fibres and other examples of total internal reflection.

Extra Goals

8. Describe and explain the formation of mirages and rainbows.

TEXT

2-1 WAVEFRONTS AND RAYS

Imagine a wave moving outwards from a source, like the expanding ripples that appear when the surface of a pond is disturbed by dropping a stone into it. Those ripples constitute a wave. All the points on the crest of a particular ripple are at the same stage, or phase, of the wave's vibration.

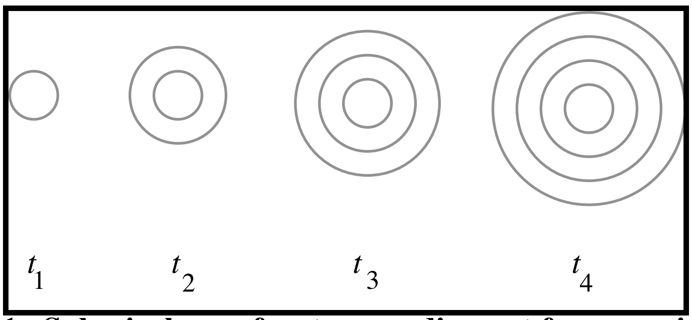


Figure 2.1. Spherical wavefronts spreading out from a point source

A curved line, or a surface for a three dimensional wave, that connects all adjacent points that have the same phase is called a **wavefront**. For the water waves on the pond a wavefront could be one of the expanding circles corresponding to a particular wave crest or trough. For sound waves the wavefront would be a surface containing all adjacent points where the wave pressure is in step. For light the wavefronts are surfaces connecting adjacent points where the oscillating electric fields are in step. Note that for any given wave we can define any number of wavefronts. It is often useful, however, to focus attention on a set of wavefronts separated from one another by one wavelength.

If the light comes from a point source, then the wavefronts are concentric spheres, centred on the source and expanding away from the source at the speed of light; light from a point source has **spherical wavefronts** (see figure 2.2). At a large distance from the source the curvature of a small section of a spherical wavefront is so small that the wavefront is nearly flat and is a good approximation to a **plane wave**.

The ray model of light

If we select a small section on a wavefront and follow its progress as it moves away from the source, the path traced out by this section is called a **ray**. A ray by its nature is always an imaginary directed line perpendicular to the wavefronts.

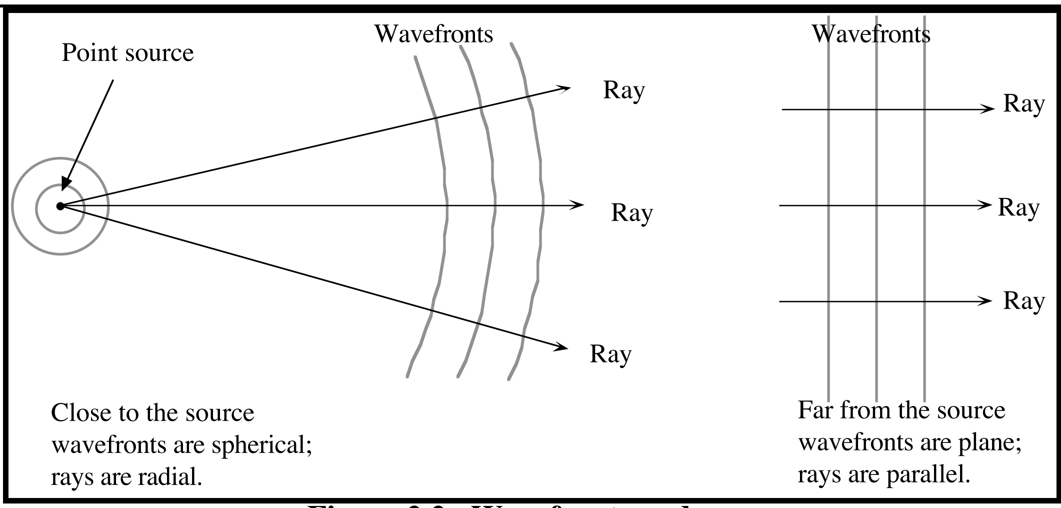


Figure 2.2. Wavefronts and rays

In very general terms rays are lines along which light travels. The direction of a ray at a point in space shows the direction in which the wave's energy is travelling at that place. We can talk about rays even without using the wave model of light.

A **beam** of light is like a tube; unlike a ray it has a non-zero width. In principle we can imagine an infinite number rays within a beam, but in practice we use only a few rays to describe the progress of the light. A narrow beam of light is often called a **pencil**.

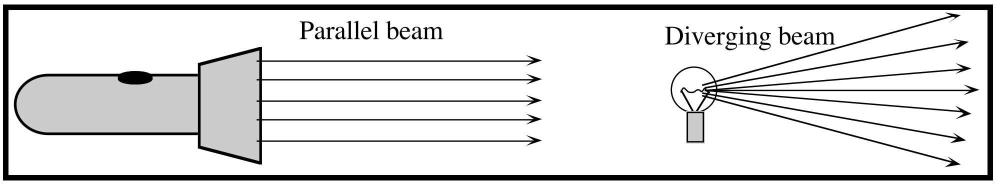


Figure 2.3. Beams of light represented as bundles of rays

2-2 INTERACTIONS OF LIGHT WITH MATTER

This chapter is concerned mostly with what happens to light when it encounters the boundary between two different materials. Before going into details of reflection and refraction we start with an overview of the processes that can happen.

We can represent light travelling through empty space or air using rays which continue straight ahead until the light meets some material object. However when light travels through a material medium the description may not be so simple. Some portion of the light in each beam may be scattered away from its original direction (figure 2.4). This **scattering** is caused by the interaction of light with small particles, even atoms or molecules, within the material. The scattered light goes off in many different directions, and may be scattered again and again before it is finally absorbed somewhere. For monochromatic light the probability of scattering depends on the relative sizes of the particle and the wavelength of the light. So some wavelengths are more susceptible to scattering than others.

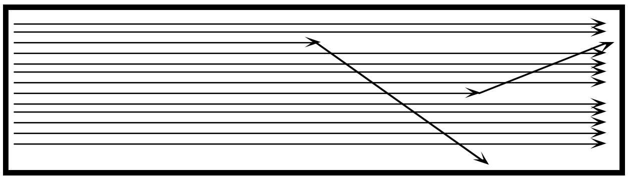


Figure 2.4. Scattering from a beam of light

Scattering is the basis of explanations of why the sky is blue and why the setting sun looks reddish. Light coming through the atmosphere from the sun is scattered by individual air molecules. Since scattering is more likely for shorter wavelengths, some fraction of the short wavelength part of sunlight - blue light - gets scattered out of the direct path from the sun. Multiple scattering spreads the scattered blue light over the whole sky. Since some of the blue light is removed from the direct white-light beam from the sun, the light that still comes through without scattering is somewhat redder than it would be if there were no atmosphere. This explanation is supported by the fact that the sun looks redder at sunset, when the light has to traverse a greater thickness of the atmosphere, than it does at midday.

On a smaller scale, the scattering of a small fraction of the light in a beam by dust or smoke particles in the atmosphere can help in tracing the path of the main beam. This effect is often used in demonstrations which allow us to see the paths of beams of light.

Transmission and absorption of light

The main interest in this chapter is in what happens to light when it comes to the boundary between two different materials. Briefly, several things can happen there: some of the light may be **reflected** back into the material where it came from while some of it may continue to travel through the second medium. You can see an example of this partial reflection when you look obliquely at a window. You can usually see a reflected image of the scene nearby, but most of the light from outside goes in through the window. Light which goes through is said to be **transmitted**. Transmitted light may or may not be **absorbed** significantly along the way. Window glass, for example absorbs very little light but a brown bottle glass absorbs quite strongly.

Light penetrates some materials better than it does others. If light penetrates without much scattering the material is said to be **transparent**. If there is a significant amount of scattering as the light goes through, the material is **translucent**. You can see things clearly through transparent materials but not so well through translucent materials. Materials which let no light through are said to be **opaque**. Light can be gradually absorbed even as it travels through a transparent material, so that a thick piece of a transparent material may appear to be opaque. Furthermore, the rate at which light is absorbed as it travels through the material can depend on the spectral composition of the light, i.e. on the mixture of different frequency components. For example white light, after passing through a slab of coloured glass, will emerge from the other side with a different mixture of frequencies, i.e. it will have a different colour.

When light comes from a transparent medium, or empty space, to the boundary of an opaque material, there may be some reflection but there is no significant transmission; all the absorption takes place in a very thin layer of material near the surface.

An important effect on transmitted light is that its direction of travel can change as it crosses the boundary between materials. This effect is called **refraction** and the light is said to be **refracted**. Refraction will be considered in §2-4.

The **speed of light** in a material is also important. In empty space, a vacuum, all light travels at the same constant speed of 3.0×10^8 m.s^{-1}, which we always denote by the symbol c. However when light travels through a material its speed is always less than c. The actual value of the speed can now depend on a number of factors such as the chemical composition and the density of the material. It also depends on the frequency of the light, so that normal light, which contains a complex mixture of components with different frequencies, travels with a range of different speeds. As you will see at various stages in this course, the dependence of speed on frequency has a number of important consequences. For example some parts of a flash of light can be delayed or left behind when the light goes through a material medium.

2-3 REFLECTION

Diffuse reflection

We see objects when light from them enters our eyes. Apart from self-luminous objects, such as the sun, lamps, flames and television screens, all other objects are seen only because they reflect light. Hence the apparent shape, texture and colour of objects depend upon the light which falls on them, called the **incident light**, and the way it is reflected. Even when the incident light comes mostly from one direction, the reflecting surface can scatter the light so that it travels in many different directions. This scattering process, which occurs at a well-defined boundary, is usually called **diffuse reflection**. The diagram shows what happens to a parallel beam of light when it is reflected diffusely. Although all the incident rays are parallel, the reflected rays go all over the place - in many different directions. This model explains why you can see an object in reflected light by looking at it from many different directions - you don't have to be in a particular place to see it.

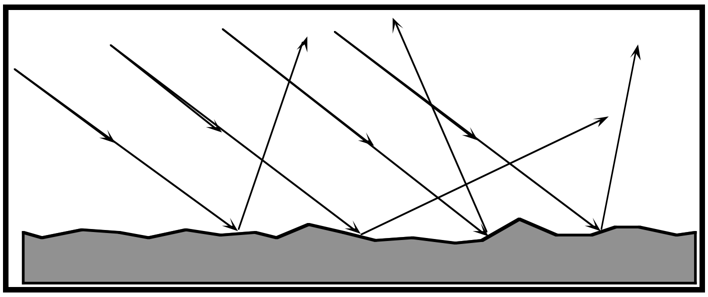


Figure 2.5 Diffuse reflection
The sketch is greatly magnified. On a microscopic scale the reflecting surface is rough,even though it may look smooth to the naked eye.

Reflectivity

When light falls on a surface some of it is absorbed or transmitted and the rest is reflected. The **reflectivity** of the surface is defined as

$$\text{reflectivity} = \frac{\text{total intensity of reflected light}}{\text{total intensity of incident light}} .$$

In this definition the incident and reflected light are each summed over all directions. Reflectivities range from less than 0.5% for black velvet and surfaces covered with powdered carbon to more than 95% for freshly prepared magnesium oxide and polished silver surfaces. White paper has a reflectivity of about 80%.

Colour

Colours of objects can be explained by supposing that their surfaces reflect different proportions of the various frequency (or wavelength) components of the incident light. Different mixes of these components produce the different visual sensations that we call colour.

It is worth noting in passing that there is no one-to-one correspondence between frequency and colour. Although some narrow ranges of light frequencies produce colour sensations such as the colours of the rainbow, red through to violet, there are many colours, such as purple and brown, which do not correspond to any one band of frequencies.

Mirror reflection

Although most examples of reflection in nature are diffuse reflection, the special, regular, kind of reflection exhibited by mirrors and very smooth surfaces plays an important role in the science of optics. This kind of reflection is called **specular reflection** (from the Latin, *speculum*, a mirror) which can be described as reflection without scattering. Some examples of specular reflectors are the surfaces of many types of glass, polished metals and the undisturbed surfaces of liquids. Some of these, such as glass and many liquids, also transmit light, whereas light does not penetrate beyond the surface of a metal. The fact that light is not transmitted through metals can be explained in terms of the interaction between the light and electrons within the metal. An example of a metal reflector is an ordinary mirror - a thin coating of metal is placed on the back surface of a piece of glass and most of the reflection takes place there. In fact the weak reflections at the front surface of the glass are usually a nuisance.

The laws which govern specular reflection can be described most simply in terms of rays. We imagine some incident light, travelling in a well-defined direction, which strikes a flat reflecting surface such as a mirror or a piece of glass. The incident light can be represented by a bundle of parallel rays. The reflected light will also travel in a well-defined direction which can be represented using another bundle of parallel rays. Since there is no scattering, for each incident ray there is only one reflected ray.

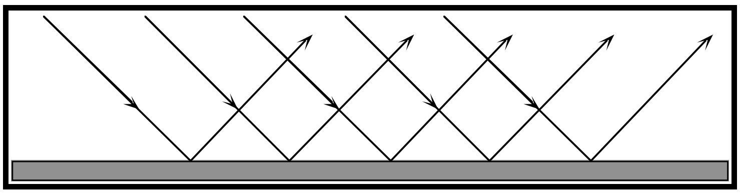


Figure 2.6. Specular or mirror-like reflection

In order to describe the relation between reflected and incident rays we need to look at the point where the incident ray meets the reflecting surface. At that point we imagine a line constructed perpendicular to the surface, in geometrical language called the **normal** to the surface. The reflected ray also departs from the same point. The angle between the incident ray and the normal is called the **angle of incidence** and the angle between the normal and the reflected ray is called the **angle of reflection**. The behaviour of the rays in specular reflection can be described completely by two laws, illustrated in figure 2.7.

- The incident ray, the normal and the reflected ray all lie in one plane.
- The angle of incidence is equal to the angle of reflection.

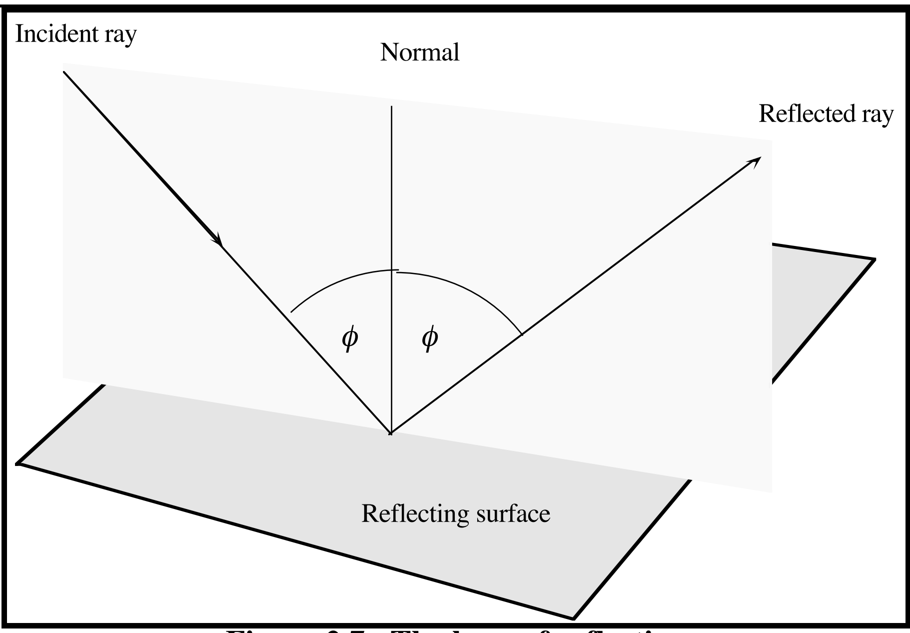


Figure 2.7. The laws of reflection

Notes

- Since any two intersecting lines define a plane, we can draw a plane diagram, like figure 2.8 below, containing the incident ray and the normal. The first of the two laws says that the reflected ray will lie in the same plane, not sticking out of the page at some angle.
- Note that the amount of light reflected cannot be predicted from these laws. That depends on the reflectivity of the surface.

2-4 REFRACTION

We have looked at the laws which govern the paths of specularly reflected light; we now consider what happens to the part of the light which goes into the other material. You already know that it could be partly absorbed, but which direction does it go? Does it go straight ahead or in some other direction or directions. The answer is that if the boundary is smooth enough to be a specular reflector, then the direction of the transmitted light is uniquely determined by the nature of the two materials, the frequency (or the wavelength) of the light and the angle of incidence. Furthermore, the light does not go straight ahead; instead the rays bend at the boundary so that the light goes on in a new direction. The new direction is described by two laws which are almost as simple as the laws of reflection.

- Firstly, the incident ray, the normal, and the refracted ray (as well as the reflected ray) all lie in the same plane. So we can draw all three rays on one plane diagram (figure 2.8).

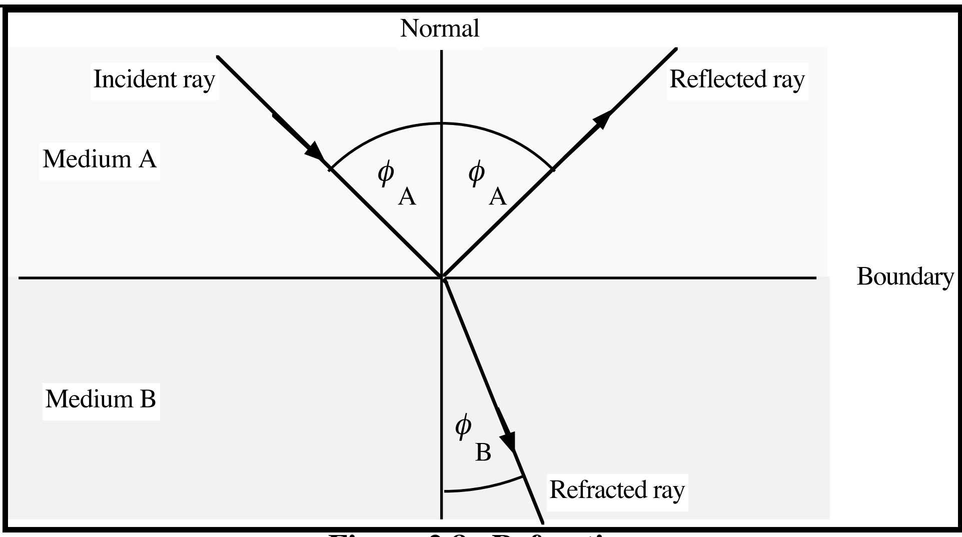


Figure 2.8. Refraction

• Secondly, the direction of the refracted ray is determined by the direction of the incident ray and the ratio of the speeds of light in the two materials:

$$\frac{\sin\phi_A}{\sin\phi_B} = \frac{v_A}{v_B} \, . \qquad \text{... (2.1)}$$

Note that if the light slows down when it goes into the second medium the rays will bend towards the normal, but if it goes faster then the rays will bend away from the normal. This immediately points to a problem with the equation, because it seems to say that we could get a situation where the sine of the angle of refraction, ϕ_B could have a value greater than 1 - which does not make sense. The proper interpretation of this is that in such a case, the refracted ray cannot exist; i.e. that the light will not penetrate the second medium at all. We return to this point shortly.

The law of refraction is a simple consequence of the wave theory of light. You can see in figure 2.9 how plane wave fronts must change their orientation if the light slows down as the wavefronts go from one material into another.

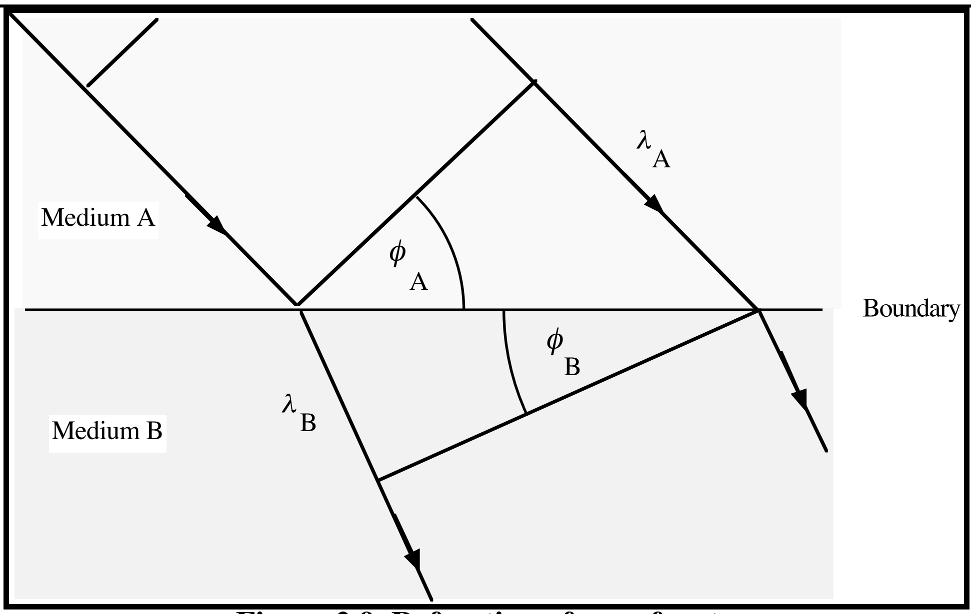


Figure 2.9. Refraction of wavefronts.

The diagram shows two consecutive wavefronts which are one wavelength apart. Since the frequency of the waves remains the same, no matter what medium they travel through, and since the wave speed is equal to the product of frequency and wavelength, the wavelength is proportional to the wave speed. Hence the wavelength in medium B is less than that in medium A. So as each wavefront crosses the boundary, it is pulled around to make a smaller angle with the boundary. Hence the rays of light, which are perpendicular to the wavefronts, must also bend as they enter the new medium.

This law of refraction was known from experiments long before the wave theory of light was invented. In its original form the law was expressed in terms of a property of the two materials called **refractive index** (symbol n) through the equation:

$$n_A\sin\phi_A = n_B\sin\phi_B \, . \qquad \text{... (2.2)}$$

Clearly, there must be some relation between the refractive index of a material and the speed of light in that material. The refractive index of a material can be defined the ratio of the speed of light in empty space (c) to the speed of light in the material (v):

$$n = \frac{c}{v} \qquad \text{... (2.3)}$$

This definition links the two forms of the refraction equation.

Notes

- The law of refraction expressed in terms of refractive index, $n_A \sin\phi_A = n_B \sin\phi_B$, is known as **Snell's law**.
- The symmetrical form of this equation, in which swapping the labels A and B makes no difference, indicates that the incident and refracted light paths are reversible - light can travel either way along the path defined by the incident and refracted rays. See figure 2.10, which (except for the reflected ray) is similar to figure 2.8 with the ray directions reversed.

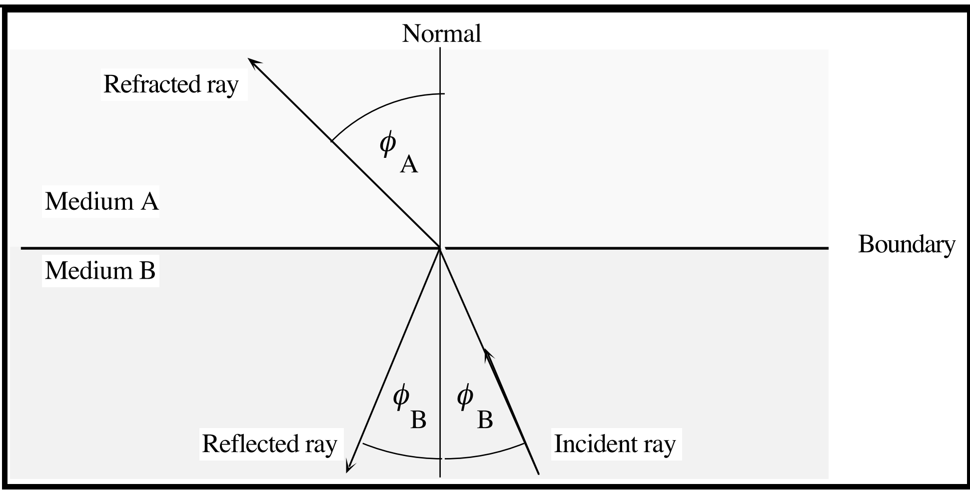


Figure 2.10. Refraction from a medium with high refractive index

- Light always travels slower in a material than it does in a vacuum. Consequently all values of refractive index are greater than one.
- The speed of light in a material depends on the chemical composition of the material, its physical state and also on the frequency of the light. The dependence of speed on frequency has some interesting consequences which we consider in §2-7 under the heading of dispersion. However for many practical applications it is good enough to use a single value of refractive index for each material. The following table shows some measured values of refractive index.

Material	Refractive index
air at STP	1.0003
ice	1.31
liquid water	1.33 to 1.34
olive oil	1.46
optical glasses	1.50 to 1.75
quartz	1.54 to 1.57
diamond	2.42

- You should remember that the speed of light in air differs from its speed in vacuum by less than 0.1%. Therefore in most calculations you can regard air and vacuum as having the same refractive index.
- The frequency of light does *not* depend on the medium.
- It follows that, since the product of wavelength and frequency is equal to the wave speed, the wavelength does depend on the medium. You can see that in figure 2.9.

2-5 REFLECTION AT A BOUNDARY BETWEEN TRANSPARENT MATERIALS

Specular reflection occurs every time light meets a smooth boundary at which the refractive index changes. The reflectivity depends on the refractive indices of the materials on either side of the boundary, the angle of incidence and the polarisation of the incident light. For a given pair of materials it also depends on which way the light goes through the boundary.

Consider first, the case where the incident light comes through the medium with lower refractive index, from air to glass for example. You can easily verify the dependence of reflectivity on angle of incidence by studying the intensity of reflections in a window as you change your angle of view from very small angles of incidence to **grazing incidence** (almost 90°). The reflectivity of glass in air is small for small angles of incidence and increases with increasing angle until it becomes almost 100% at grazing incidence.

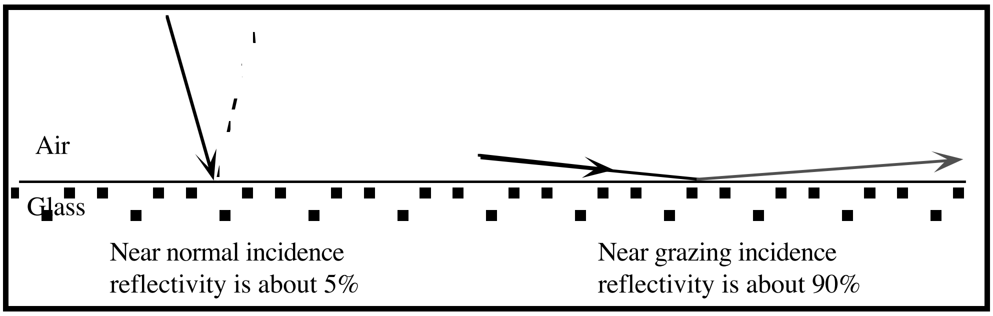


Figure 2.11. Effect of angle of incidence on the reflectivity of plain glass

If the refractive index decreases across the boundary, (e.g. from glass to air), then at small angles of incidence the reflectivity is again low. But this time as the angle of incidence increases the reflectivity reaches 100% well before grazing incidence. Complete reflection happens when the angle of incidence is greater than a value called the **critical angle**, denoted by ϕ_c in figure 2.12.

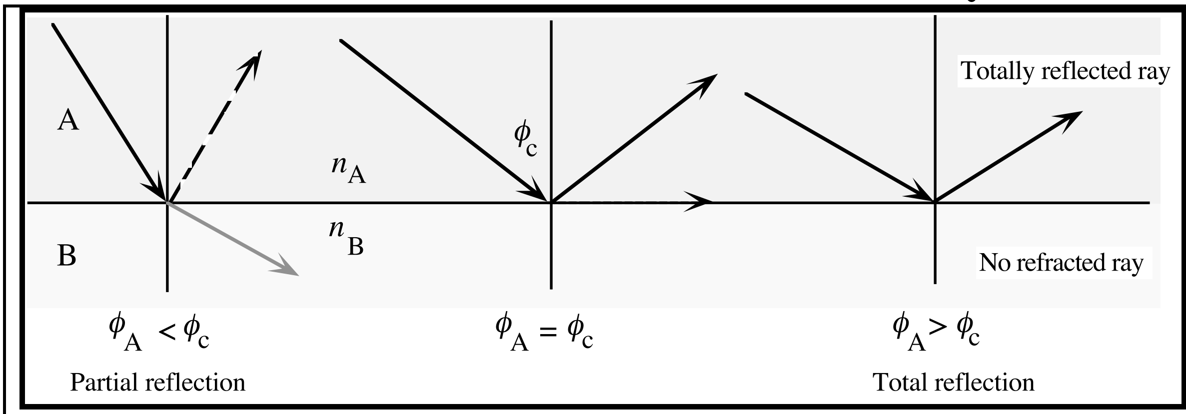


Figure 2.12. Reflection at a boundary where refractive index decreases

Beyond the critical angle all the incident light is reflected and there is no refracted ray, so the phenomenon is called **total internal reflection**. The relation between critical angle and the refractive indices of the two media can be found by inserting the maximum possible value for the angle of refraction, 90°, into Snell's law which gives

$$\sin\phi_c = \frac{n_B}{n_A} \qquad \text{...(2.4).}$$

Remember that total internal reflection can occur only when light strikes a boundary where the refractive index decreases; reflection is back into the medium with the higher refractive index.

2-6 APPLICATIONS OF TOTAL INTERNAL REFLECTION

Prism reflectors

An ordinary glass mirror consists of a reflective metallic coating on the back of a sheet of glass but that is not the only way to make a mirror. Total internal reflection can be exploited to make a perfectly reflecting mirror using only glass, with no metal backing. Figure 2.13 shows how: light enters a prism perpendicular to the first surface so it is not refracted. When the light reaches the next face, the angle of incidence is greater than the critical angle so all the light is reflected. In this example, when the light gets to the third face of the prism it is refracted as it leaves the prism. That final refraction could be a problem because the refractive index is slightly different for different frequencies of light.

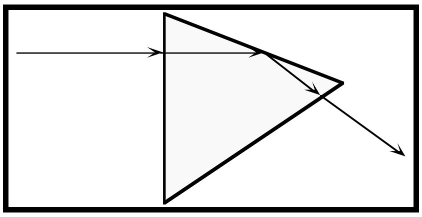


Figure 2.13. Total internal reflection in a prism

However if we use a right-angled prism and a suitable type of glass (figure 2.14) the light can be made to undergo two total reflections with no net refraction before it emerges in a direction which is always exactly opposite to that of the incident light. Such a device is often called a **corner reflector** or retroreflector. Retroreflecting beads are exploited in reflective road signs and "cat's eyes".

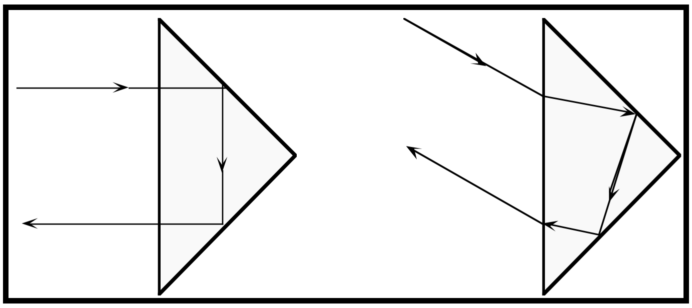


Figure 2.14. A corner reflector

The direction of a reflected ray is always reversed.

A pair of corner reflecting prisms can be used to displace a beam of light sideways without altering its direction of travel or to compress the path of a light beam into a small space. This arrangement, which is often used in binoculars, is an example of a device called an **optical relay** - a device which simply alters a light path without contributing to the formation of an image (see also chapter L7).

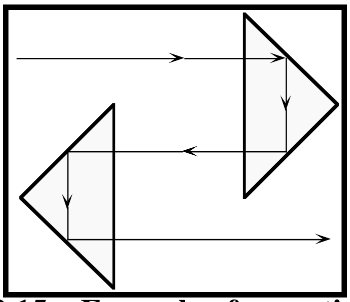


Figure 2.15. Example of an optical relay
In this example the path of the ray is displaced sideways but its final direction is unaltered.

Light pipes

Another important application of total internal reflection is the **optical fibre** or light pipe. Here a light ray enters one end of a transparent rod or fibre and is totally reflected many times, bouncing from side to side until it reaches the other end. This alone is not very useful, but what is important is that when the pipe is bent, the light path can be bent with it, staying within the pipe. The light pipe still works provided that each angle of incidence remains greater than the critical angle, so the light cannot get out until it reaches the flat end of the light pipe. Although there is a high contrast in refractive index between the material of the fibre and air, fibres often need to be coated with a protective medium which reduces the ratio of refractive indices and hence, also, the value of the critical angle. In order to make sure that the angles of incidence remain large enough, the fibre should not be bent too severely.

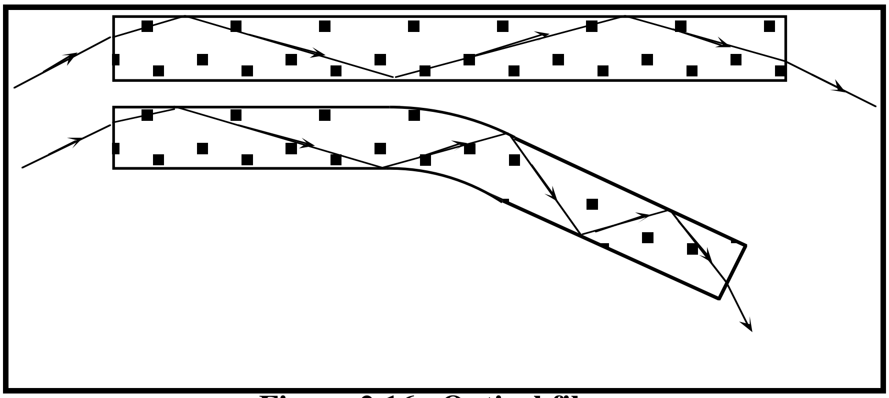


Figure 2.16. Optical fibres
Light is trapped in the fibre by total internal reflection - even when the fibre bends.

Optical fibres have many uses including data transmission, an alternative to sending electrical signals along conducting cables. The advantage of optical fibres here is that the capacity of the medium to carry information is vastly greater. Many different signals can be sent along the same fibre; in more technical terms, optical fibres have large bandwidths.

An important medical application is the fibre-optic **endoscope**, a device for transmitting images of inaccessible internal organs. A typical endoscope contains two bundles of optical fibres - one to carry light to illuminate the object and another bundle to transmit the image. The image is formed by a small lens attached to the end of a collection of thousands of individual fibres. Each fibre carries light from one part of the image, which can be viewed at the other end where the light emerges. In order to get a useful image at the output end, the fibres must be arranged in the same way that they were at the input end. Images seen this way are necessarily grainy, since the final image consists of a collection of light or dark coloured spots, one spot for each fibre.

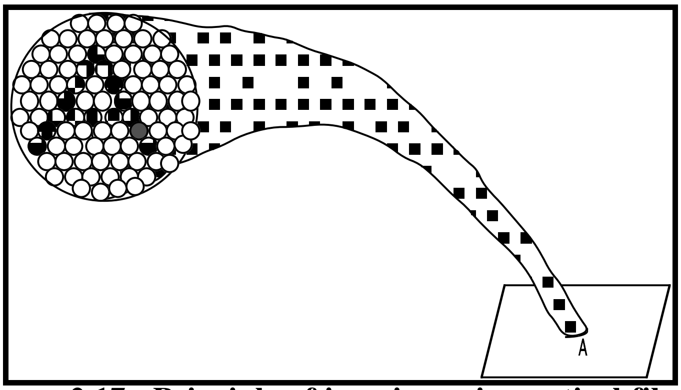


Figure 2.17. Principle of imaging using optical fibres
The resolution is limited by the diameter of each of the fibres.

2-7 DISPERSION

The dependence of refractive index (and wave speed) on the frequency of light produces some important effects which are often very useful and occasionally a nuisance, but nearly always pretty. The beautiful effects can be explained in terms of the notion that the perceived colours of light are related to the mixture of frequency components that the light contains.

The classic example is the production of a **spectrum** of many colours when ordinary white light passes through a prism of clear (colourless) material such as glass. Each ray of light which passes through the prism is refracted twice, once as it enters and again as it leaves. See figure 2.18. The amount of bending or refraction depends on the frequency of the light (as well as the nature of the glass). So white light, which can be described as a continuous distribution of many different frequency components, will bend by many different amounts; one ray of white light becomes a continuous collection of rays with a continuous range of frequencies. Only a few such rays can be shown in the diagram.

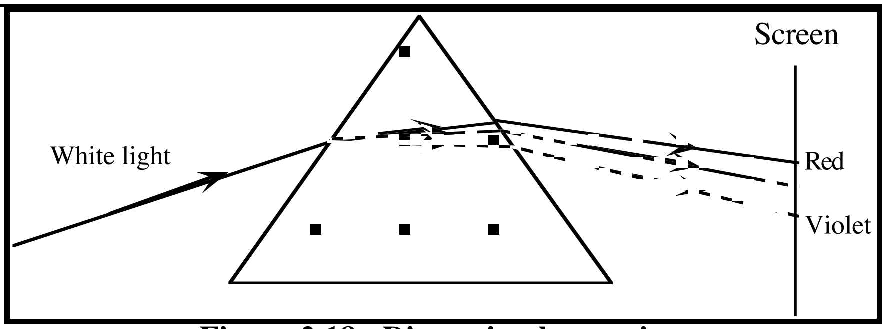


Figure 2.18. Dispersion by a prism
The angular separation between rays is exaggerated.

When a beam of white light is sent into a prism and the refracted light is allowed to strike a diffuse reflector such as a white card, a spectrum of light is formed on the screen. The colours of the spectrum range from red, corresponding to the light which is refracted least, through yellow, green and blue to violet which is refracted the most. Since we know from independent evidence that violet corresponds to high frequency radiation, we can conclude that the refractive index of glass is higher for higher frequency light. The relationship between frequency and refractive index is not, however, a simple linear one, see figure 2.19.

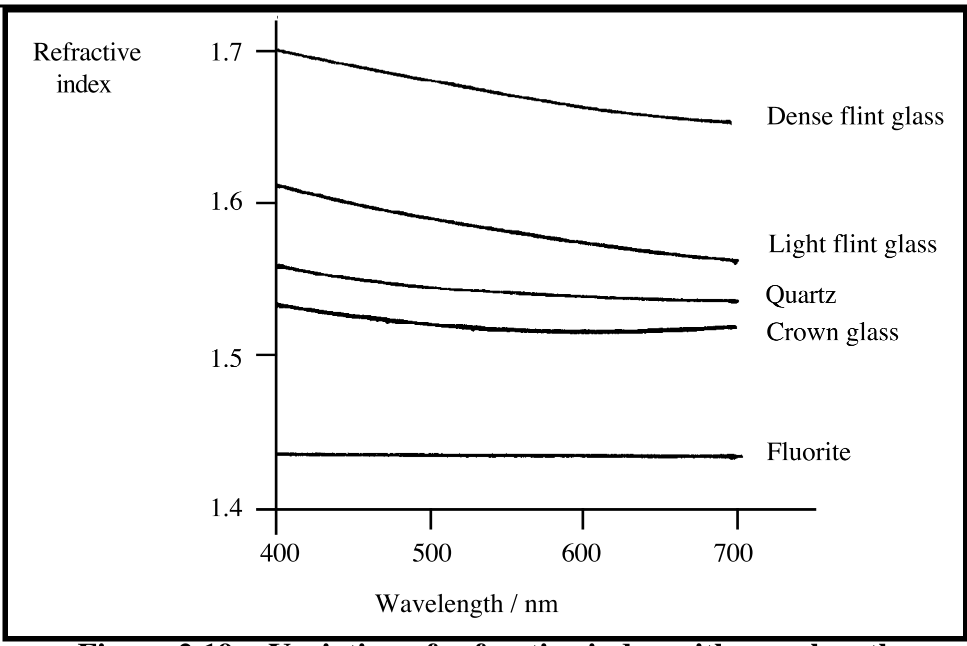


Figure 2.19. Variation of refractive index with wavelength

Since the frequency of light is not easily measured directly, it is traditional to specify properties like refractive index which vary with frequency in terms of the variation with the wavelength instead. (Wavelengths of light can be measured using interference and diffraction techniques described in chapters L4 and L5.) Values of wavelength used in such descriptions are always the wavelength in vacuum corresponding to $\lambda = c/f$. They do *not* refer to the actual wavelengths of the light in the glass.

Glass prisms are used in **spectroscopes** and **spectrographs -** instruments which disperse the spectrum of a light source into components with different frequencies. A simple arrangement is illustrated in figure 2.20.

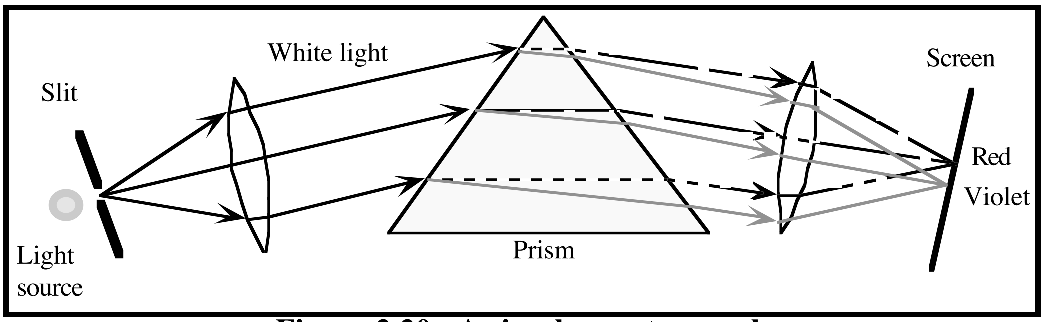


Figure 2.20. A simple spectrograph

Angular separations between rays are exaggerated.

Rainbows

The colours of the rainbow are formed by dispersion in small water droplets. A complete explanation involves some complicated ray tracing, but it is clear that whatever the light paths are, they are different for different frequencies. Figure 2.21 shows how dispersion in a raindrop produces a primary rainbow. (The primary rainbow is the brightest bow, sometimes the only one that you can see.)

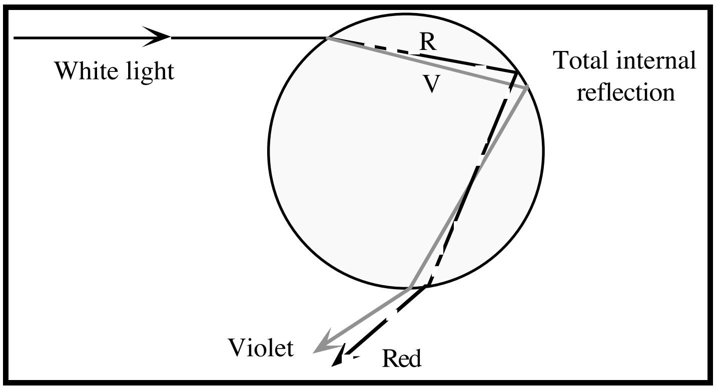


Figure 2.21. Dispersion in a water droplet

A ray of white light from the sun is refracted as it enters a spherical raindrop (figure 2.21) and dispersion occurs. The dispersed light rays are totally internally reflected and are then refracted again as they leave the drop. The dispersed rays which come out are now travelling in different directions, depending on their frequencies, so they appear to come from different parts of the sky. The angles between the incident rays from the sun and the rays from the rainbow are essentially fixed by the refracting properties of water and are on average about 138° for the primary rainbow. This fixed value for the scattering angle accounts for the shape of the rainbow.

2-8 MIRAGES

There are several kinds of mirage. Probably the commonest type is the illusion that light from distant objects is being reflected by a pool of water which is not really there. This kind of mirage is caused by refraction in a hot layer of air close to the ground. Although the refractive index of air is very close to 1.000, it is not exactly 1. Furthermore the refractive index of the air depends on its temperature. Light coming from the sky at an angle not much above the horizon travels into air whose refractive index gets less as the light gets closer to the ground. See figure 2.22. The variation in refractive index makes the light rays bend away from the ground so that eventually they will be totally internally reflected within the air and will travel upwards. You can see this effect most noticeably on a long horizontal bitumen road on a hot day. The black bitumen absorbs a good deal of the sunlight which hits it and it gets very hot. The road surface then heats the air immediately above it, the hottest air being closest to the road, so the refractive index is least near the hot road surface.

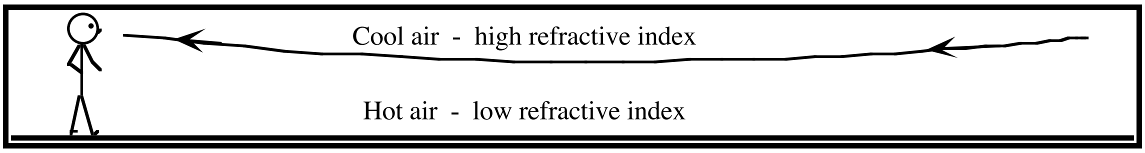


Figure 2.22. Path of a light ray in a common mirage

Although the variation in refractive index is continuous, the process can be understood in terms of many different layers with different refractive indices. Imagine a ray coming to the boundary between two such layers, as in figure 2.23. If the ray is close to horizontal it has a large angle of incidence, so when it goes into the hot air of lower refractive index the angle of refraction is even larger. In the lower part of figure 2.23 an incident ray meets a boundary at an angle greater than the critical angle so it is totally reflected. Then as the ray continues back up through the air the refraction process is reversed and the angle to the horizontal gets larger. A person seeing the refracted light perceives that it is coming up from the ground, but it looks like light from the sky, or some object near the horizon, creating the illusion that the light has been reflected by a pool of water.

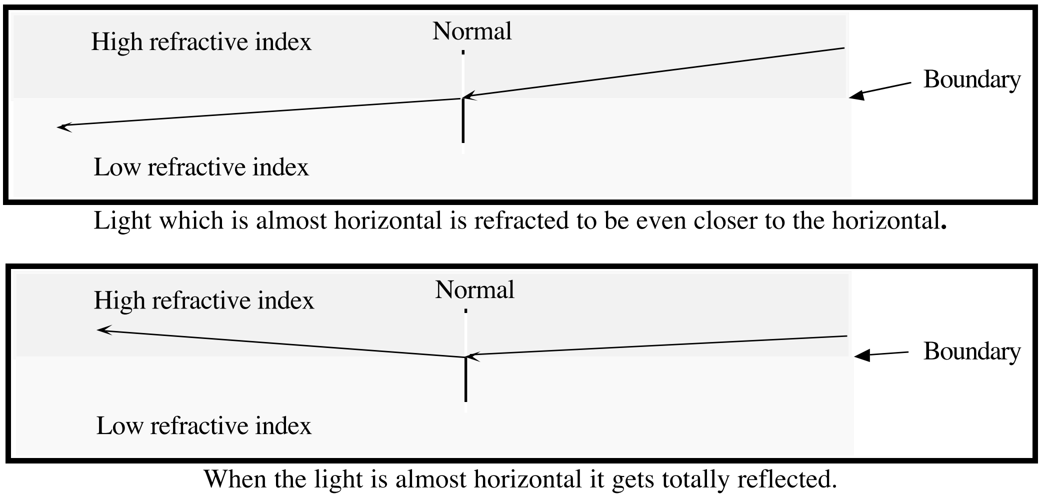


Figure 2.23. Refraction and total internal reflection in a mirage

Other kinds of mirage are more complex than this but all can be explained in terms of variations in the refractive index of the atmosphere.

2-9 IMAGES

We see things by the light that comes from them into our eyes. Although the process of seeing is a complex one involving both eye and brain, some aspects of seeing can be discussed in terms of ray optics. When you see an object your eye collects light from all over the object. Light rays go out in all directions from each point on the surface of the object, but only some of those rays enter the eye and those that do are contained within a cone. The angle of that diverging cone of rays depends on the distance from the object point to the eye - the further away the object, the smaller is the angle. Although the eye-brain system does not respond directly to that angle, or the degree of divergence of the rays, it does produce perceptions of depth by much more complex mechanisms. We can, however, model or calculate the apparent distances of object points from an eye by considering the diverging cone of rays from each object point to the eye.

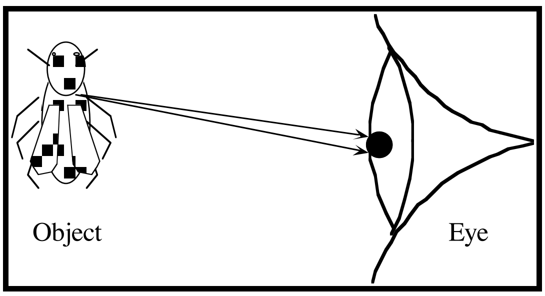


Figure 2.24. Seeing an object
The direction and the divergence of the rays indicate the perceived position of the object point.

The apparent location of an object point can be found by considering rays from the same object point arriving at the eye from different directions. Those rays can be extended back until they meet, in order to find out where they appear to come from. The point where they meet is called an **image point**. When there is no refraction or reflection of the light rays as they travel from the object to the eye, through still air for example, the positions of the object and its image coincide. However if the light is reflected or refracted on its way to the eye, then object and image are in different places.

Specular reflection by a plane mirror

Although many rays of light are involved, the image point corresponding to each object point can be found using any two rays. All other rays from the same object point will, after reflection, appear to come from the same image point. The diagram shows how two rays coming from an object point are reflected in a plane mirror.

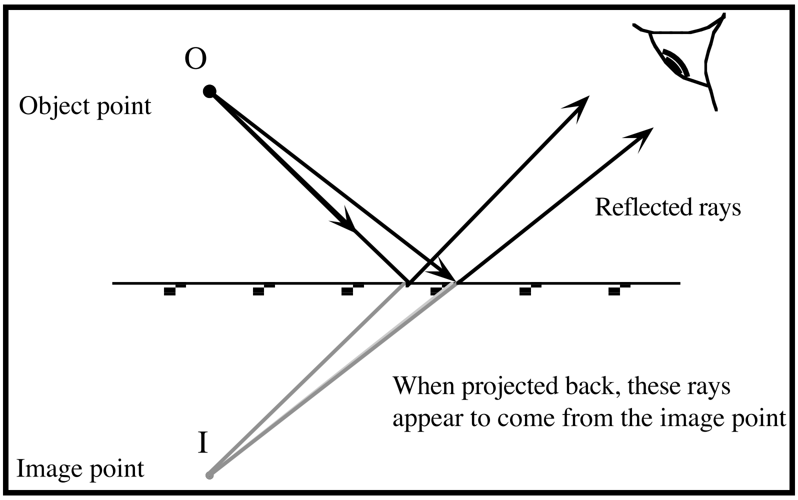


Figure 2.25. Reflection in a plane mirror

When the reflected rays are projected back behind the mirror, they appear to diverge from an image point I which is as far behind the mirror as the object point O is in front. Note that in this and other diagrams the actual rays of light are drawn in black while their projections back into places where the light does not really go (or come from) are shown in grey. Since the light does not actually come from the image in this case, it is called a **virtual image**. This method of locating the image by following the paths of different rays is called **ray tracing**.

Images affected by refraction

Objects located inside a refracting medium, such as water, seem to be in the wrong place and they also look distorted. You can easily observe that by putting an object in a dish of water. The diagram shows how light rays coming from an object point under water are bent as they leave the water so that they seem to be coming from an image point which is not at the position of the object. In this example the image of one object point is actually somewhat spread out - the cone of rays no longer diverges from a unique point after refraction. Since the eye collects only a very narrow cone of rays, the spreading out effect is not noticeable if you keep your eye in one place. But if you move your head, you will see the image move! Contrast that with normal viewing in which the brain perceives that fixed objects stay put when you move your head.

Other examples of virtual images formed by refraction at plane boundaries include the apparent bending of straight objects placed partly underwater and the pair of images of one object seen through adjacent sides of a fish tank.

For more about images see chapter L3.

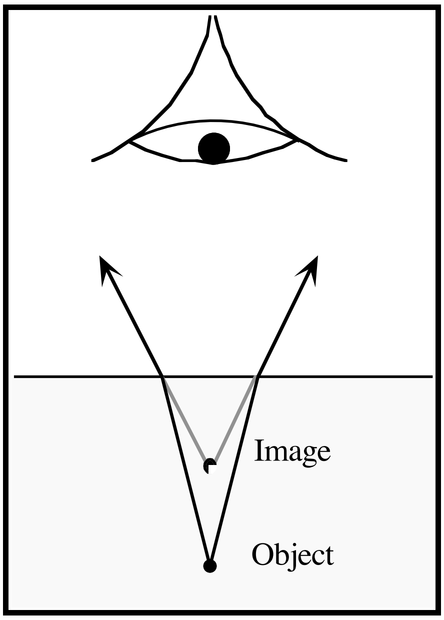


Figure 2.26. Viewing an object under water

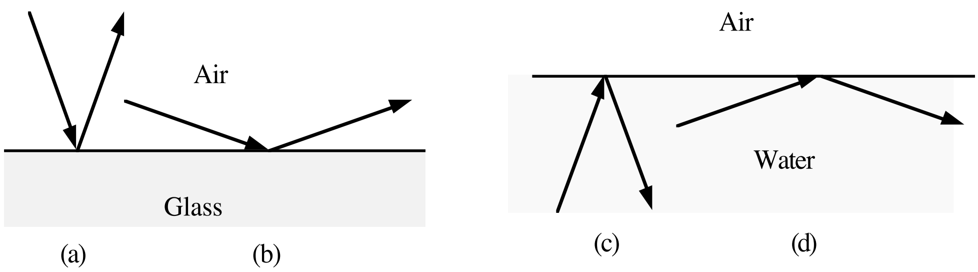
The angular width of the cone of rays is exaggerated. Only a small cone of light enters the eye.

QUESTIONS

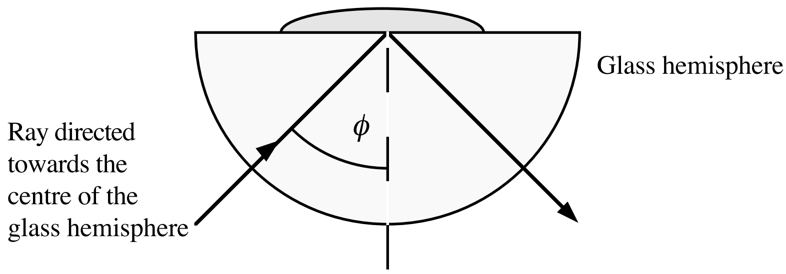
The following questions do not have answers that have to be learned. They are designed to help you to think about the relevance and applications of principles covered in this chapter.

Q2.1 In the corner reflector of figure 2.14, the angles of the prism are 90°, 45° and 45°. What can you say about the refractive index?

Q2.2 Look at the diagrams below and in each case, determine whether little, almost all, or all of the incident light is reflected.



Q2.3 The refractive index of small quantities of liquid can be measured by finding the critical angle of reflection.

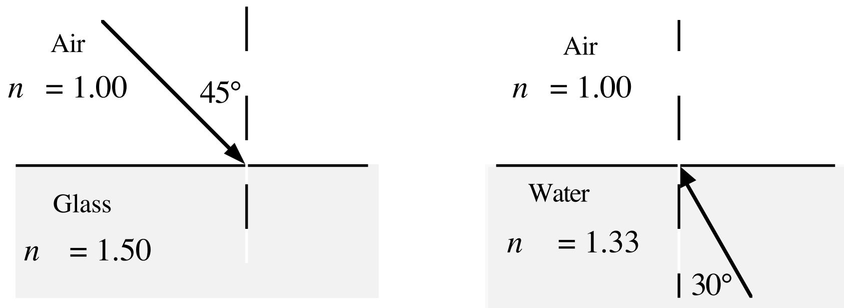


Total internal reflection takes place for all angles of incidence ϕ greater than the critical angle. The critical angle with a drop of liquid present is 59°. The refractive index of the glass is 1.56. Calculate the refractive index of the liquid.

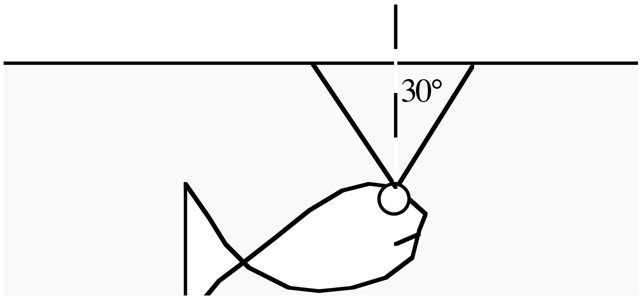
Over what range of values of the refractive index of the liquid can this method be used?

Q2.4 Recently, in one year, eight people in N.S.W. suffered severe spinal injuries caused by diving into shallow water and landing on their heads. In some cases the water was clear and the bottom of the pool was plainly visible. Why is it surprising that people should make that mistake?

Q2.5 Calculate the angle between the refracted ray and the normal and sketch the path of the refracted ray in the two examples below.



Q2.6 A fish views the outside world through a water-air boundary so its view is distorted. Suppose the fish's eye has a field of view in water which is a cone of half angle 30°. When the fish looks straight out of the water how much of the outside does it see? What would the fish see if its field of view in water were a cone of half angle 50°?



Q2.7 Refer to the graph of refractive index as a function of wavelength for various materials (figure 2.19). Which would give a more spread-out spectrum, a prism of dense flint glass or a prism of crown glass?

Q2.8 A book quotes the refractive index of an optical glass as 1.48626 at a wavelength of 587.6 nm. What is the frequency of the light used? What is its actual wavelength in the glass?

Discussion questions

Q2.9 When you look at reflections in a sheet of glass, you can often see a double image. Why?

Q2.10 When you look over the top of hot object, such as a bitumen road on a summer's day, the view of things beyond seems to wobble or shimmer. Explain.

Q2.11 Do you think that sound waves should obey the same laws of reflection and refraction as light waves?

Q2.12 Can you invent an experiment to measure the wavelength of light using specular reflection? Could you do it with refraction?

Q2.13 Can total internal reflection occur when light travels through water to a boundary with glass? How would you specify the kind of material where total reflection of light travelling through water can occur?

Q2.14 Does the value of critical angle for a given pair of materials depend on the frequency of light? Does total internal reflection cause dispersion? Can there be any dispersion in light which has been totally internally reflected?

Q2.15 The usual way to make a spectrum using a slab of glass is to make a prism in which there is an angle (not zero) between the faces where the light goes in and out. Does that mean that you can't get a spectrum from a piece of glass with parallel faces (zero angle) like a window pane? What is the advantage of having the two faces at an angle?

Q2.16 Making a spectrum by just putting a prism into the path of some white light doesn't give the best results. What else should you do to make a really nice spectrum?

FURTHER READING

Katzir, Abraham; **Optical Fibers in Medicine**, *Scientific American,* May 1989, 86 - 91.

Fraser, A.B. & W.H. Mach, **Mirages**, *Scientific American,* January 1976, 102 - 111.

Walker, Jearl, **The Amateur Scientist, How to create and observe a dozen rainbows ...**, *Scientific American,* July 1977, 138 - 144.

Walker, Jearl, **The Amateur Scientist, Mysteries of rainbows ...**, *Scientific American,* June 1980, 147 - 152.

L3 IMAGES

OBJECTIVES

Aims

From this chapter you should develop an understanding and appreciation of what images are and how they are formed. In particular you will learn how lenses work. You will also learn how to solve simple quantitative problems involving image formation by lenses, using both geometrical and algebraic methods.

Minimum learning goals

1 Explain, interpret and use the terms:

(a) *image*, *object, real image, virtual image,*

(b) *plane mirror, lens, convex lens, converging lens, concave lens, diverging lens, positive lens, negative lens, thin lens,*

(c) *paraxial approximation, paraxial rays, converging beam, parallel beam, diverging beam,*

(d) *principal point, principal axis, principal plane, focal point, focal plane, focal length, power, dioptre,*

(e) *lens equation, sign convention,*

(f) *object distance, image distance, object position, image position, linear magnification, lateral magnification, longitudinal magnification, virtual object.*

2 Explain the nature of images and how they are formed.

3 Describe and explain how lenses and plane mirrors produce images.

4 Describe and apply standard ray-tracing procedures for both thin and thick lenses.

5 Describe and explain those properties of a lens which determine its power. State and apply the lensmaker's formula.

6 Solve simple problems involving image formation and magnification, by single thin lenses or a set of thin lenses with the same principal axis, using both ray-tracing and algebraic techniques. Describe and explain the rules and techniques used in such problems.

TEXT

3-1 IMAGES

The reflection of your face in a mirror, the view of a small insect under a microscope and the picture on the big screen at the movies are all optical **images**. They are formed by light rays whose paths have been altered by the action of a mirror or a lens on the light coming from an **object**.

It is convenient to classify images into two types, called real and virtual. In a **real image** the light actually goes to (or through) the image. Examples include the image on the cinema screen and the light image formed on the film in a camera. Real images are often formed on screens or other solid surfaces, but a screen is not essential. Even if the cinema screen were to be suddenly whisked away in the middle of a movie, a real image would still exist in the same place, but to see it you would have to leave your seat and go to a place, well behind where the screen used to be, where you could look back towards the projector. Images formed in the eye are also real images.

The light rays which form a **virtual image** do not actually pass through the image - they only appear to be coming from it. The most familiar example of a virtual image is a "reflection" in a mirror. Although the image is located some distance behind the mirror, the light does not actually go there or come from there. It is an *illusion* that the light comes from behind the mirror.

The most important image-forming system of all is the human eye. Whenever you see something, real images are being formed on the retinas of your eyes. Sometimes this means that the eye forms an image of an image. This happens, for example, when you see the virtual image of your face in the mirror or an image formed by a lens.

3-2 RAY TRACING

The most basic technique for calculating how images are formed is **ray tracing**, which is just the construction of diagrams showing where light rays from an object go. Starting from some point on the object, a ray is drawn towards the optical device. Where the ray meets a reflecting or refracting boundary the laws of reflection and refraction give the new direction of the ray. The process continues until the ray does not meet any more surfaces. The whole procedure is then repeated for another ray, starting at the same object point, but heading off in a different direction. In principle, this procedure should be repeated for a large number of rays. Once all the rays have been plotted you can look for a point where they all meet. If such a point can be found then a real image of the object point exists there. If the rays do not actually meet at a point, there is no real image but there may be a virtual image. To look for a virtual image, the rays which finally come out of the optical system are projected backwards to find out if the those projections meet. If they do, then their intersection gives a virtual image point. The tracing of many rays leads to the location of one image point for each object point. To locate the whole image the complete procedure is repeated for other object points, gradually building up a map of the complete image. The complete process is tedious, time-consuming and very accurate. It is one way that professionals can check the design of complex and expensive optical systems. In practice, ray-tracing calculations and plotting are mostly done by computers.

Ray tracing is also the basis of much simpler procedures for calculating the approximate locations of images in simplified optical systems. In some simple cases, exact answers can be found very quickly with little effort. A good example of ray tracing is to locate the virtual image formed by a plane mirror.

3-3 PLANE MIRROR

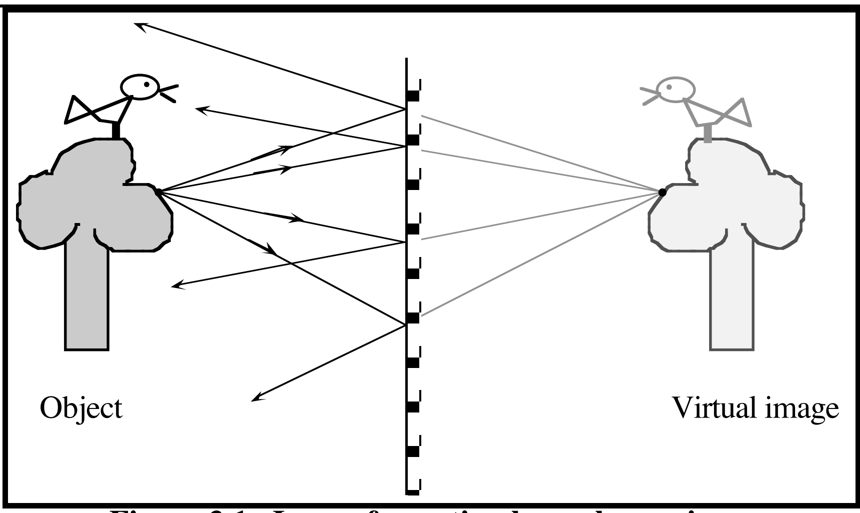


Figure 3.1. Image formation by a plane mirror

Figure 3.1 shows the image formed by a flat reflecting surface. The diagram shows just a few of the many rays diverging from the same object point. Each ray is reflected so that its angle of reflection is equal to the angle of incidence. The rays themselves do not meet so there is no real image, but if the reflected rays are projected backwards (shown by gray lines) they do meet. Since light does not really come from that point behind the mirror it is a virtual image point. If more rays are added to the diagram it is found that all the reflected rays diverge from the same virtual image point; so each image point is unique. That result indicates that a really flat mirror produces a sharp image - does that match your experience?

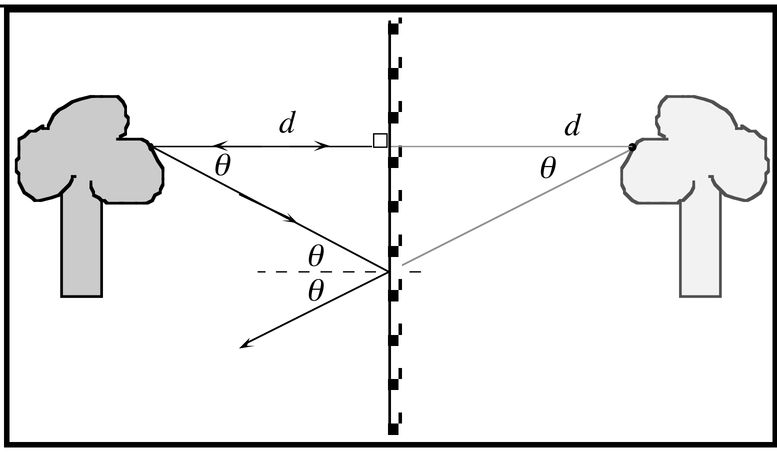


Figure 3.2. Geometry of mirror reflection

It is actually quite easy to show that with a plane mirror all the rays from one image point are reflected so that they diverge from one virtual image point. Figure 3.2 shows just two reflected rays, one normal to the mirror and another one. Marking up equal angles θ and spotting the similar triangles in the diagram shows that the distance of the image point from the mirror is equal to the distance of the object point from the mirror and that the line joining them is perpendicular to the reflecting surface. Since that result is true for any value of the angle θ it is true for all reflected rays. These diagrams also illustrate why the study of image formation is called geometrical optics.

3-4 REFRACTION AT A CURVED BOUNDARY

Whenever light crosses the boundary between two optical media which have different refractive indices the light bends or refracts. As a light ray goes from low to high refractive index it bends towards a line normal to the surface; when it goes from high to low refractive index it bends away from the normal. See figure 3.3.

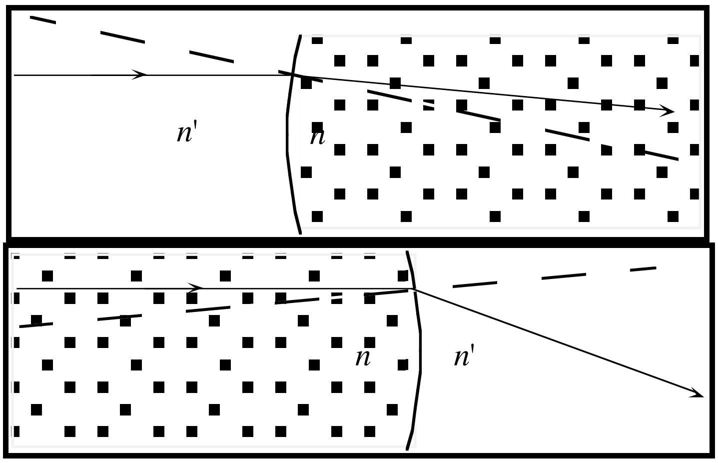


Figure 3.3. Refraction at a curved boundary
The medium with the higher refractive index, *n*, is shaded. The broken line is normal to the surface.

Parallel beam

When we consider light from a very distant object all the rays from one point of the object are practically parallel to each other and are said to form a **parallel beam** of rays. Since it is often said that parallel lines "meet at infinity" an object point which produces a parallel beam can be described as being at infinity. In practice that means that the distance to the object is very large compared with the other relevant distances. Many aspects of the behaviour of optical systems can be explained in terms of what they do to a parallel beam, or rays from objects at infinity.

Figure 3.4 shows what can happen to a parallel beam of light (coming from an object point at infinity) when it crosses a spherical boundary between two materials with refractive indices n' and n. In this case the boundary is convex towards the incoming beam and n' is less than n.

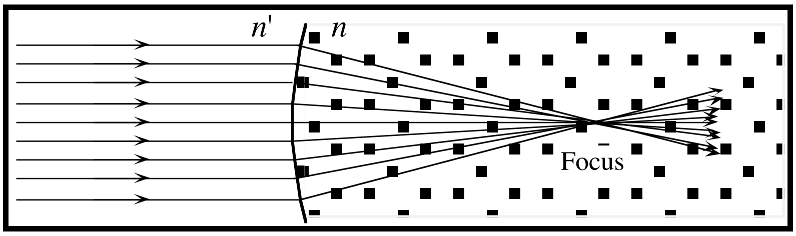


Figure 3.4. Focussing by a spherical refracting boundary
A narrow beam can be focussed to a point.

Each ray obeys the law of refraction and the refracted rays form a converging cone of light. If the beam is narrow compared with the radius of curvature of the surface all the rays will pass close to the same point. The rays are said to come to a **focus** - the beam has been **focussed** to form a real image of the object point. The focussing effect or degree of convergence produced by the surface depends on its **curvature**. A flat surface has no curvature so it can produce no convergence - an incident parallel beam will still be parallel after refraction, although it may be travelling in a different direction. The greater the curvature the greater is the converging effect. Strong focussing is also produced by a large contrast in refractive indices. If the refractive indices of the two materials were equal there would be no focussing effect, but a large ratio n/n' can produce strong focussing provided that there is also some curvature.

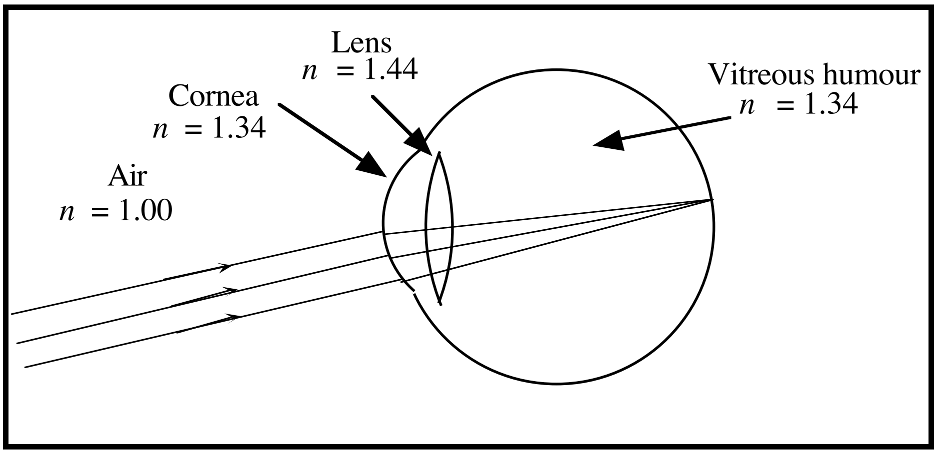


Figure 3.5. Simplified optical structure of the eye
Most of the refraction occurs at the front surface of the cornea.

An example of focussing by a curved surface is the action of the **cornea** of the eye (figure 3.5). The cornea has a curved boundary with the air and most of the eye is filled with transparent materials which have fairly uniform refractive index. Therefore most of the refraction of incoming light occurs at the front surface of the cornea. Some refraction also occurs at the the two surfaces of the lens of the eye, but the contrast of refractive index there is quite small so the focussing by the lens is weak. The lens is used essentially for fine adjustment in the focussing of visual images. For more about the eye see chapter L7.

3-5 LENSES

A simple **lens** is an optical device, usually made of glass or clear plastic, which can form images. Most lenses have a circular outline and two curved, often spherical, faces. They form images by refracting rays of light at both of their surfaces. If you knew the precise shape of a lens and the refractive index of the material in it you could use ray tracing to calculate (or make a computer calculate) the locations of images. In general you would find that lenses do not produce perfectly sharp images and there are two kinds of reason for that. Firstly, diffraction effects (which will, be considered in chapter L5) produce a blurred image for every object point. The effects of diffraction can be minimised, but not eliminated, by using large lenses. Secondly, even if diffraction did not exist there would still be the geometrical restriction that it is not possible to design a lens, or a system of lenses, which produces a unique image point for every possible object point. On the other hand it is possible to design lenses and systems of lenses which produce satisfactory images. The quality of the image can be improved using a more complex design - at greater cost.

Types of lenses

There are two basically different kinds of lens. When used in air, **converging lenses**, which are thicker in the middle than at the edges, make a parallel beam of light converge. A **diverging lens**, which can turn a parallel beam into a diverging beam, is thinner in the middle.

Action of a converging lens

Figure 3.6 shows how a converging lens affects a light ray passing through it. The total effect is just that of refraction at two successive curved surfaces.

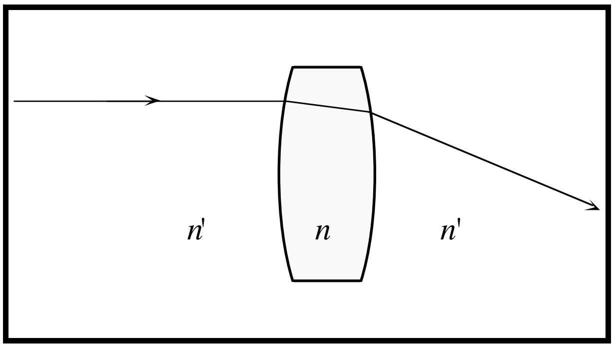


Figure 3.6.

Refraction by a converging lens

As the ray passes through the first surface it is refracted towards the normal and it then continues in a new direction through the glass until it arrives at the other side. There the ray goes from glass to air so it is refracted away from the normal and emerges in a new direction. These changes in direction of the ray can be calculated using the law of refraction (Snell's law) so the path of any ray can be traced out in this way. In principle all you need to do to find out if an image exists, and where it is located, is to trace out rays. Ideally we would like to have all rays from one object point arriving at a unique image point. In reality that is impossible to achieve for all object points, but with good design, it can be achieved approximately.

Principal axis

Most lenses are symmetrical about an axis or line through the middle of the lens. If you rotate the lens around this axis it looks just the same. In optics that axis of symmetry is called the **principal axis** of the lens. From now on we think of every lens as having rotational symmetry about its principal axis. That is a fortunate simplification because it allows us to describe and work out the optics of lenses using two-dimensional drawings and constructions. You need to remember, however, that in reality lenses, objects and images are three-dimensional structures. The usual way of drawing lens diagrams is to draw a line across the paper to represent the lens's principal axis (figure 3.7). Rays are drawn on a two-dimensional diagram which represents a slice through the lens and the three-dimensional bundles of rays.

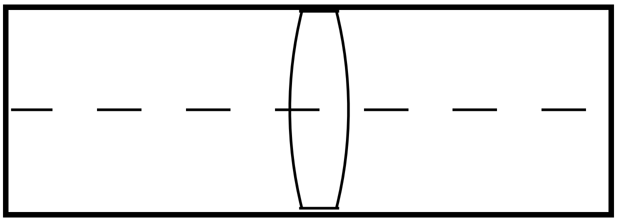


Figure 3.7.
Lens and principal axis

Focal point and focal length

The meaning of the term *converging lens* is illustrated by the lens's action on a parallel beam of rays coming from a distant object point that is located on the principal axis. When the rays come out the other side of the lens they form a converging cone of light (figure 3.8).

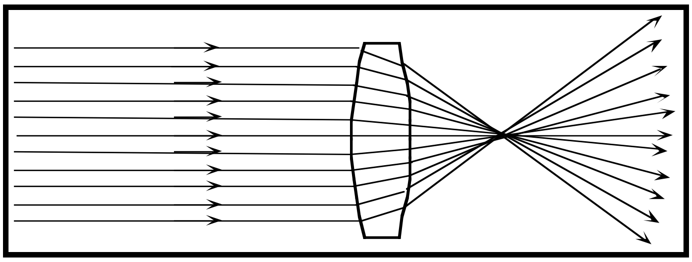


Figure 3.8. Function of a converging lens
If the whole lens is used the rays will converge but they will not all go through the same point.

An ideal converging lens would refract all the rays in a beam parallel to the principal axis so that they pass through one real image point, which is called the **focal point** of the lens. (In reality the image "point" is always a bit blurred.) The light is said to have been **focussed** by the lens. The distance from the focal point to the middle of the lens is called the **focal length** of the lens.

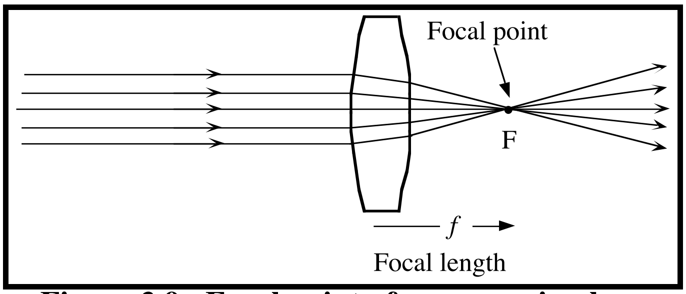


Figure 3.9. Focal point of a converging lens
A narrow (paraxial) beam parallel to the principal axis is focussed on to the focal point.

Paraxial rays

The somewhat blurred image "point" corresponding to a point object can be made sharper by restricting the rays which form the image so that they are close to the principal axis and also by making sure that the angle between any ray and the principal axis is small. Such rays are said to be **paraxial**. Paraxial rays don't have to be parallel to the principal axis but they do have to be close to it. In most of the lens diagrams in this and other texts the angles between the rays and the principal axis are often quite large and the rays may be a fair distance from the axis; so the drawings are not good representations of the paraxial condition. The angles and off-axis distances are generally exaggerated so that you can see the features of the ray diagrams more clearly.

If the distant object point is not located right on the principal axis, but is off to one side, the incoming bundle of parallel rays will be at an angle to the principal axis. If that angle is small the rays satisfy the **paraxial approximation** and the converging lens will produce a reasonably sharp image point, as shown in figure 3.10. In this case the focus or image point is not on the principal axis but it does lie in a plane, called the **focal plane**, which is perpendicular to the principal axis. The focal point of the lens is at the intersection of the focal plane and the principal axis.

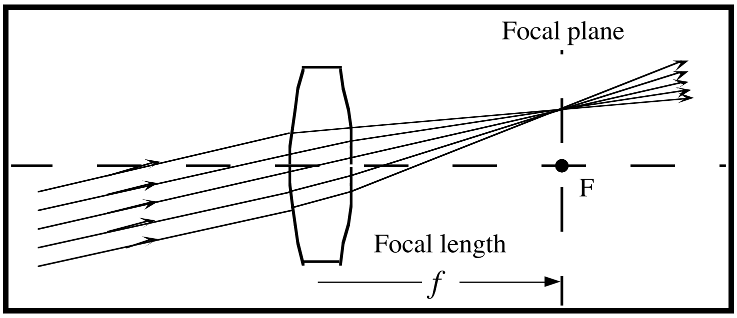


Figure 3.10. Focussing by a converging lens
Incoming parallel rays from any small angle come to a focus in the focal plane.

An important feature of all ray diagrams is that if you reverse the directions of all the rays then you get another valid diagram. The light paths are said to be reversible. So, for example, to find out how a lens affects the rays coming from an object point at a lens's focal point, you could just reverse all the rays in figure 3.9. That would give light going from right to left instead of the usual left to right, so for consistency the diagram is reversed left-to-right, which gives figure 3.11.

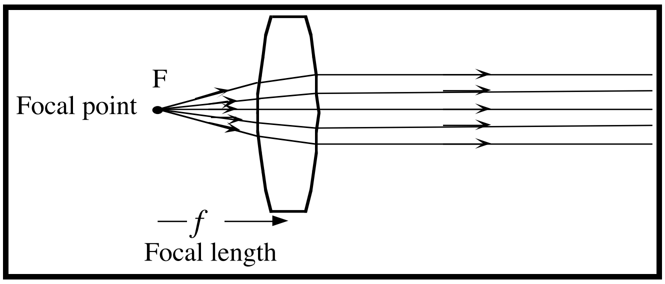


Figure 3.11. Object point at the focal point of a converging lens
This diagram is like figure 3.9 with the rays reversed. Light paths are reversible.

Thus every lens has two focal points, one one each side. Provided that the lens is immersed in the same medium on both sides, the two focal lengths are equal.

Converging and diverging lenses

Figures 3.9 and 3.10 show how a converging lens makes a parallel beam of light into a converging beam. A converging lens can also make a converging beam into an even more converging beam (figure 3.12). It can also refract a diverging beam into less diverging beam, a parallel beam or even into a converging beam (figure 3.14, below). In summary, a converging lens increases the convergence of any light beam which passes through it.

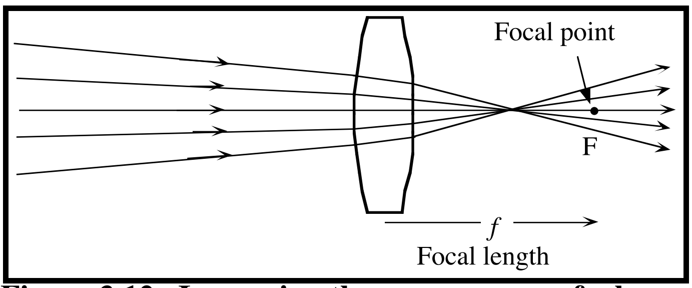


Figure 3.12. Increasing the convergence of a beam
The beam is focussed before the focal point.

A **diverging lens** is thinner in the middle than it is at its edge. It bends the parallel beam from a distant point into a diverging cone of rays which (ideally) appear to come from one virtual image point. Here the focal point and the focal plane are on the same side of the lens as the incident light.

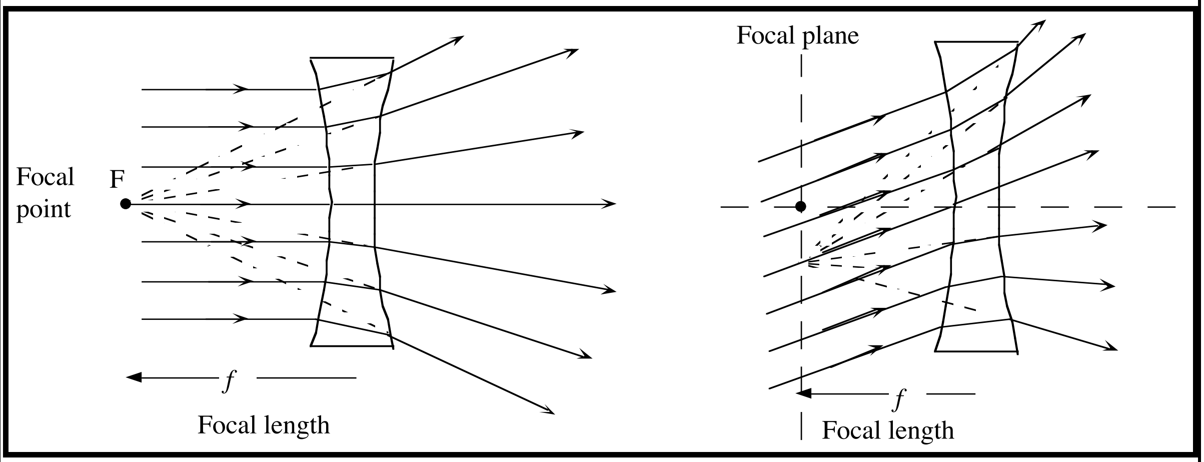


Figure 3.13. Function of a diverging lens

Just as a converging lens increases the convergence of a bundle of rays, a diverging lens decreases the convergence - or you could say that it increases the divergence of the rays.

Power of a lens

The shorter the focal length of a converging lens the better it is at converging light. This characteristic of a lens, its converging ability, is called **power**, which can be defined formally as the reciprocal of focal length:

$$P = \frac{1}{f} \quad . \qquad \text{... (3.1)}$$

The SI unit of optical power is the **reciprocal meter** (m^{-1}). A commonly used alternative name for the unit, is the **dioptre**. For example, if the focal length is +0.500 m, the power is $+2.00\ m^{-1}$ or 2.00 dioptres and if the focal length is -0.040 m, the power is $-25\ m^{-1}$. Converging lenses always have positive values of power and are often known as **positive lenses**. The negative value of power for a diverging lens expresses the fact that it does the opposite of converging; a diverging lens can be called a **negative lens**.

3-6 FORMATION OF IMAGES BY THIN LENSES

Figure 3.14 shows how a converging lens refracts the divergent bundle of rays from one point on an object into a bundle of rays which converge onto a real image point. The diagram shows only five of the many rays which could be drawn from the object point.

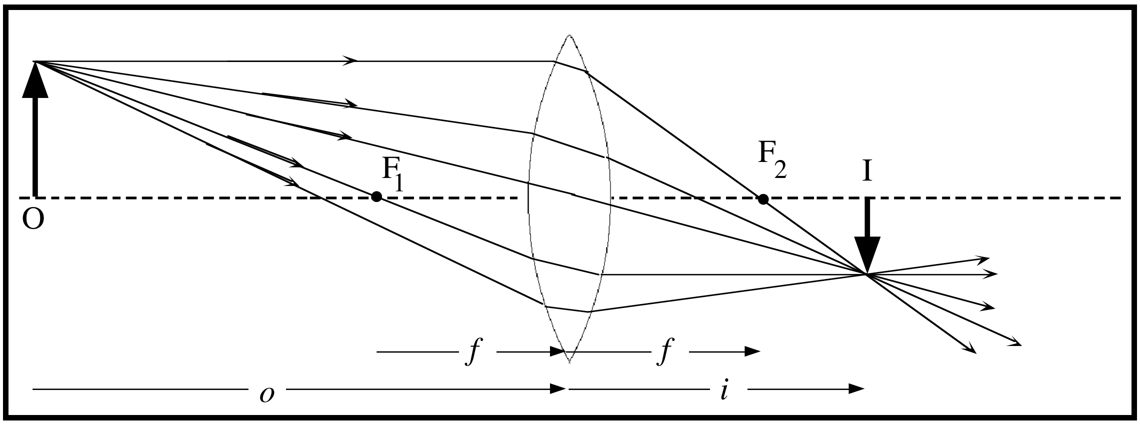


Figure 3.14. Image formation by a converging lens
Ideally, all the rays from one object point converge to one image point.

If all the rays are paraxial they will all come to a focus at the same image point. The actual paths of all these rays could be worked out by ray-tracing. Image points for other object points could be located in the same way. There are some features of this example which are worth noting; for an ideal **thin lens** and paraxial rays we get the following results.

- All object points which are at the same distance from the lens produce image points at equal distances from the lens. So the image position can be located using one object point and its image point. (Different object distances give different image distances.)
- If the object is placed further from the converging lens than the focal length its image will be real.
- The orientation of the real image is opposite to that of the object - it is said to be **inverted**.

Simplified model of a thin converging lens

If all the rays are paraxial, detailed ray tracing is not necessary. The function of a lens can be described completely in terms of a geometrical construction in which the lens is represented by a set of points, lines and planes and the location and size of an image can be found by tracing only two rays. Figure 3.15 shows the essential features of the model for a thin lens. The **principal plane** is a plane perpendicular to the principal axis located centrally within the lens. The **first focal plane** and the **first focal point** are located on the side of the lens where the light comes from. Light diverging from a point source in the first focal plane will emerge as a parallel beam. The **second focal plane** contains all the points where an incoming beam of parallel rays can come to a focus. The focussing properties of the lens are determined by the locations of the principal plane and the focal planes relative to the principal axis.

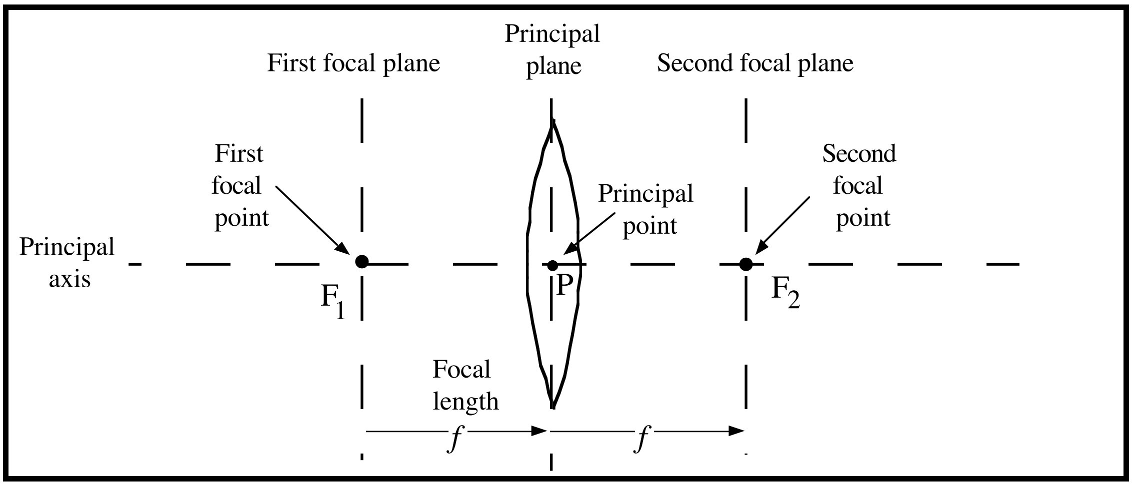


Figure 3.15. Geometrical model for the function of a lens

Standard ray tracing

The following procedure produces accurate answers provided that the actual situation is restricted to paraxial rays. To get accurate results you need to make a scale drawing - graph paper helps - but the procedure can also be used to make rough sketches to work out, for example, whether images are real or virtual and whether they are upright or inverted.

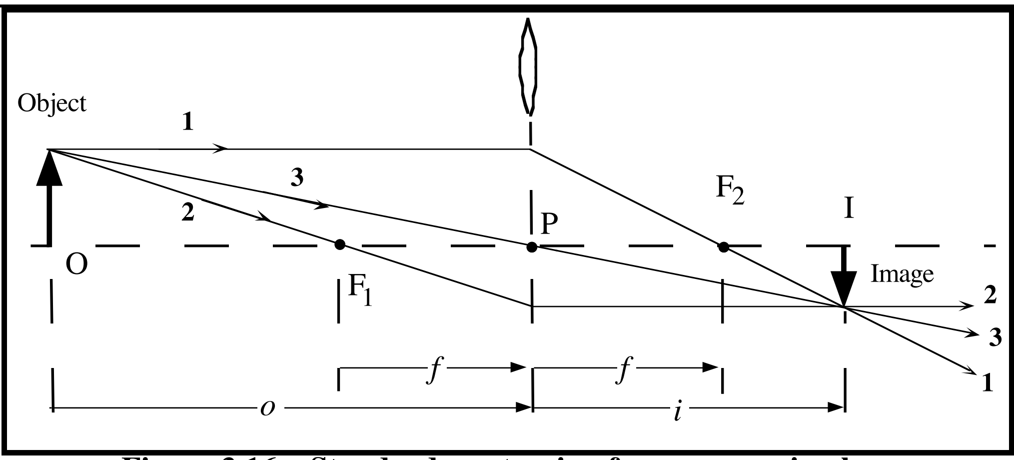


Figure 3.16. Standard ray tracing for a converging lens

- First draw two perpendicular lines to represent the **principal axis** and the **principal plane** of the lens. The **principal point** is at the intersection of those lines. If you like you can include a small sketch above the principal plane to indicate the type of lens. Measure out and mark the positions of the two focal points. (For a converging lens the first focal point is the one on the side where the light comes from; for a diverging lens it is on the other side.)
- Mark the object position, O, and draw a line representing the object, perpendicular to the principal axis with one end on the axis.
- Choose an object point somewhere off the principal axis.
- From that object point construct *any two* of the following three rays (figure 3.16).

  1. An incident ray from the object point parallel to the principal axis is refracted so that it intersects the second focal point.

  2. An incident ray which intersects the first focal point is refracted parallel to the principal axis.

  3. An incident ray which intersects the principal point continues undeviated.
- The image point will be at the intersection of the two refracted rays. If those rays actually cross you have a real image. If the rays don't meet you will need to extend them to find where their extensions cross. In that case you have a virtual image. (It is a good idea to show these extended rays in a different style; in diagrams in this book they are printed grey.)
- Draw a line perpendicular to the axis to represent the image.

Diverging lens

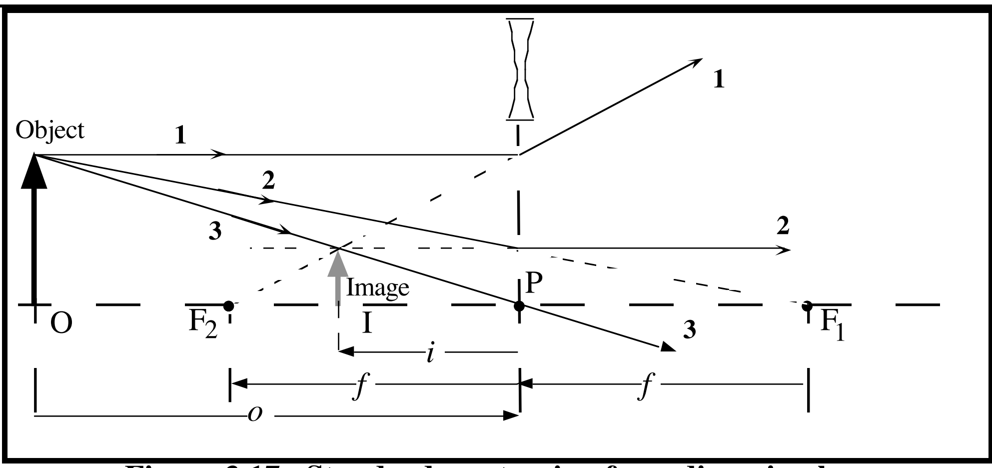


Figure 3.17. Standard ray tracing for a diverging lens

Note that the first focal point is on the far side of the lens.

Figure 3.17 illustrates the construction for a diverging lens. Note that the first focal point is now on the far side of the lens. The rules for constructing rays are exactly the same. Again, any two of the three rays will do. In this case real rays do not intersect, they have to be extended backwards (shown here as grey lines) to find the image. If any ray has to be extended back to the image point, that always indicates a virtual image. You will find that a diverging lens always produces a virtual image of a real object. The construction in figure 3.17 also shows that the image is **upright**.

This ray tracing procedure works perfectly only if you bend the construction rays at the principal plane rather than at the curved surfaces of the lens. When you do that you no longer need to worry whether the rays in your construction are truly paraxial or not; the rays can be drawn at any angle and at any distance from the principal axis and they will still give the same answer as a realistic construction using paraxial rays refracted twice at the surfaces of the lens.

The lens equation

Image formation by a thin lens using paraxial rays can also be described by the **lens equation**:

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} \; . \qquad \text{... (3.2)}$$

Here o and i are the distances of the object and the image from the lens (see figure 3.17) and f is the focal length. This equation is simply the algebraic equivalent of the ray tracing procedures described above and will give exactly the same answers for paraxial rays. The equation works for both converging and diverging lenses provided that a suitable sign convention is used.

Sign convention

The lens equation requires the following **sign convention** (or an equivalent one).

- The object distance o is measured from the object to the lens.
- The image distance i is measured from the lens to the image.
- Object and image distances are positive if they are measured in the same general direction as that in which the light goes, negative otherwise.
- A negative value of image distance indicates a virtual image and a positive value means that the image is real.
- The focal length, f, is always positive for a converging lens, and negative for a diverging lens.

The convention should be easy to remember because you just follow the light and use the natural sequences: (a) object - lens - image and (b) first focal point - lens - second focal point. Draw a single headed arrow to represent each distance. Then if any arrow points backwards (against the light) the corresponding distance value is negative. The convention means that the object distance is normally positive - for one real object and a single lens it is *always* positive. (It can have a negative value only in a compound optical system where there is an intermediate image which acts as a virtual object for the next component of the system. See chapter L7.)

The sign convention is illustrated by the directions of the arrows in figures 3.16 and 3.17. For the formation of a real image by a converging lens (figure 3.16) f, o and i are all positive. In the case of the diverging lens (figure 3.17), f is negative, o is positive and i is negative.

Example

Find the image formed by a converging lens of focal length 35 mm when the object is placed 85 mm from the lens.

Answer

Rearrange the lens equation to get: $\frac{1}{i} = \frac{1}{f} - \frac{1}{o}$.

You can either start substituting in this equation, or continue the algebraic manipulation, making the image distance the subject:

$$\frac{1}{i} = \frac{o - f}{o.f} \; ;$$

so

$$i = \frac{of}{o - f}$$

$$= \frac{(85\text{ mm}) \times (35\text{ mm})}{(85 - 35)\text{ mm}} = 60\text{ mm}.$$

Since value of i is positive the image is real and it is located 60 mm from the lens, or 145 mm from the object.

3-7 MAGNIFICATION

The **magnification** produced by a lens is defined as the ratio of the image size to the object size. If the sizes are specified in terms of lengths, the corresponding magnification is called a **linear magnification** whereas sizes described as angles subtended at the lens give **angular magnification**. Angular magnification will be considered in chapter L7 - here we concentrate on linear magnification. Linear sizes can be measured in different ways. If the dimensions of object and image are specified using measurements perpendicular to the principal axis, the linear magnification is called **lateral magnification**. It can easily be worked out from a standard ray diagram such as figure 3.18.

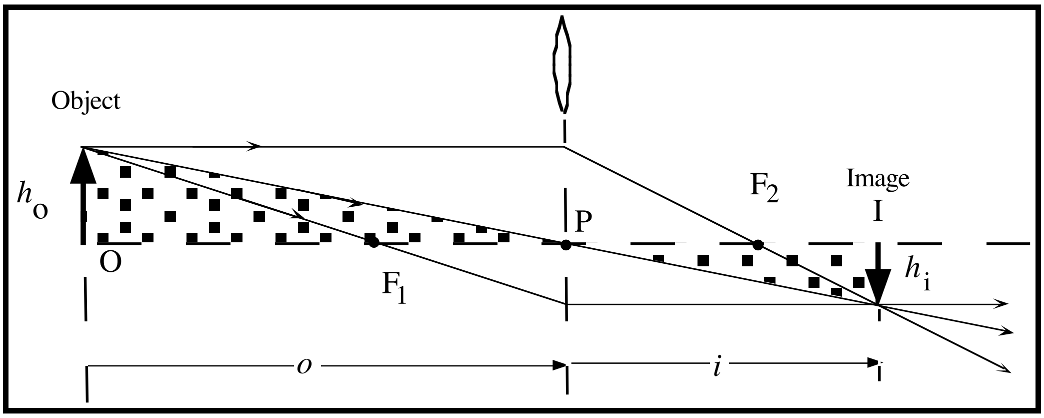


Figure 3.18. Lateral magnification by a converging lens
The similar triangles (shaded) show how magnification is related to object and image distances.

The magnitude of the lateral magnification is defined as the ratio of the image size to the object size:

$$|m| = \left|\frac{h_i}{h_o}\right| . \qquad \text{... (3.3)}$$

The pair of similar triangles, shaded in figure 3.18, also shows that the magnification is given by the formula:

$$m = -\frac{i}{o} \qquad \text{... (3.4)}$$

where the minus sign is inserted so that a negative value for magnification indicates an inverted image. In figure 3.18 the real image is inverted and magnification is negative.

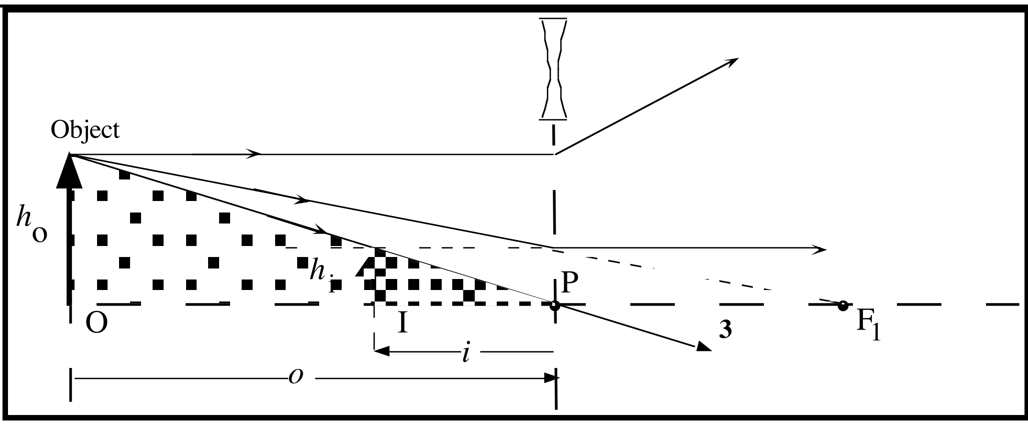


Figure 3.19. Lateral magnification by a diverging lens

Figure 3.19 shows the same situation as figure 3.17; here the virtual image is upright and the magnification is positive. The two shaded triangles can be used to show that the magnification formulas (equations 3.3 and 3.4) are the same as those for the converging lens.

3-8 COMBINATIONS OF LENSES

The general case

To find an image formed by two or more lenses which are not in contact, you can't just add the powers. But the image can be located by standard ray tracing or by repeated application of the lens equation. Both methods proceed in the same way. The first step is to calculate the image formed by the first lens alone, ignoring the second lens. Then use that image as the object for the second lens. The new object distance is found by subtracting the previous image distance from the lens separation. Then calculate the second image. Repeat the process for each lens in the system.

Example

Two converging lenses with focal lengths of 20.0 mm and 30.0 mm are placed 10.0 mm apart with their principal axes coinciding. An object is located 60 mm along the principal axis from the stronger lens. Find the position and lateral magnification of the image.

Answer

The first step is to calculate the image formed by the first lens alone - call it lens A. Proceed as though the second lens is not there. Rearranging the lens equation gives

$$\frac{1}{i_A} = \frac{1}{f_A} - \frac{1}{o_A} .$$

Substituting $f_A = 20.0$ mm and $o_A = 60$ mm gives $i_A = 30$ mm. The result means that there would be a real image 30 mm from lens A. Now imagine that lens B is put in place 10.0 mm beyond lens A and use the image formed by lens A as the object for lens B (figure 3.20).

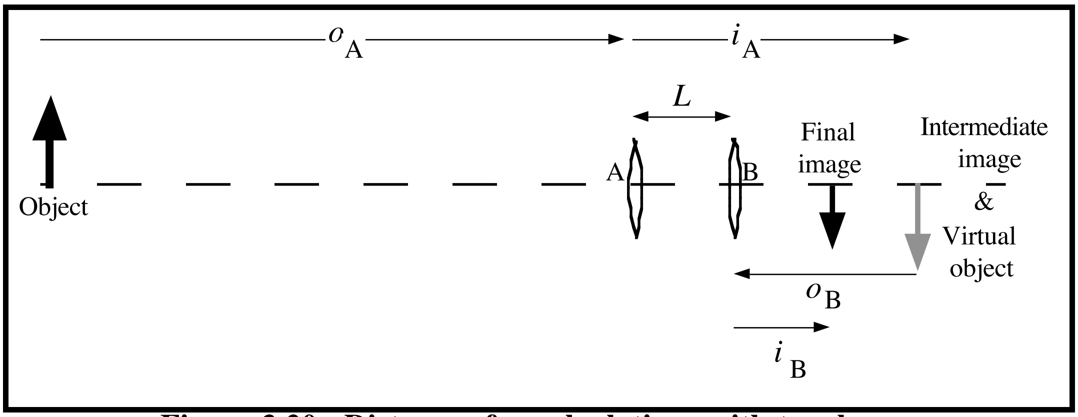


Figure 3.20. Distances for calculations with two lenses

The image formed by lens A acts as the object for lens B.

In this example there is a problem: lens B will intercept the light before it gets to the location of image A. The problem can be handled by saying that the object distance is negative and that the object for B is a **virtual object**.

Now calculate the new object distance from the previous image distance and the separation L between the lenses: $o_B = L - i_A = 10.0 \text{ mm} - 30 \text{ mm} = -20$ mm. Use the lens equation again:

$$\frac{1}{i_B} = \frac{1}{f_B} - \frac{1}{o_B} .$$

The final image distance is $i_B = 12.5$ mm.

The answer means that the image is real and is located 12.5 mm past lens B, or 22.5 mm from lens A.

The final magnification is the product of the two individual magnifications:

$$\begin{aligned} m &= m_A m_B \\ &= \left(-\frac{i_A}{o_A}\right)\left(-\frac{i_B}{o_B}\right) \\ &= \left(\frac{i_A}{o_A}\right)\left(\frac{i_B}{o_B}\right) \\ &= \left(\frac{30 \text{ mm}}{60 \text{ mm}}\right)\left(\frac{12.5 \text{ mm}}{-20 \text{ mm}}\right) = -0.31 . \end{aligned}$$

The negative value here indicates that the image is inverted.

The same answers can be obtained using standard ray tracing diagrams drawn to scale. Figure 3.21 shows how that can be done, using separate diagrams for each image, in order to avoid cluttering the construction too much.

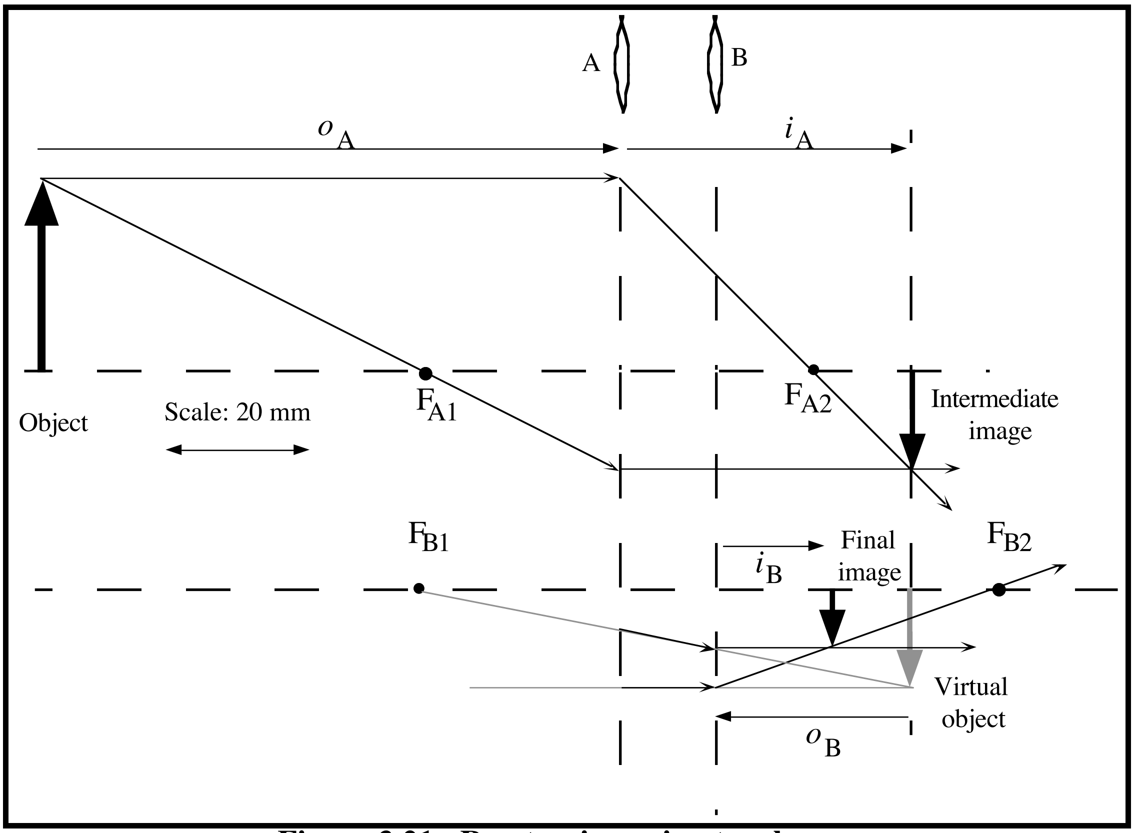


Figure 3.21. Ray tracing using two lenses

The top part of the diagram shows how the intermediate real image is located, by ignoring the presence of lens B. When lens B is in place the intermediate image becomes a virtual object for lens B, so the lower part of the diagram shows virtual rays from the object extended back to lens B. On the front (left) side of the lens those rays can be drawn as real rays. The construction rules are the still the same. A ray from the object parallel to the axis is bent to go through the second focal point. Another ray extended back from the object to intersect the first focal point emerges parallel to the axis. Since these two refracted rays are real their intersection gives the location of the real final image point.

Thin lenses in contact

When the separation between two thin lenses is small compared with each of their focal lengths calculations like the one above become much simpler - the two lenses can be treated as one! When two thin lenses of powers P_1 and P_2 are placed *in contact* the resulting power of the **compound lens** is the sum of the individual powers:

$$P = P_1 + P_2 . \qquad \text{... (3.5)}$$

Example. The combined focal length of two converging lenses in contact, with individual focal lengths of 50 mm and 35 mm is given by

$$\frac{1}{f} = \frac{1}{50\ \text{mm}} + \frac{1}{35\ \text{mm}} .$$

So $$f = \frac{50\ \text{mm} \times 35\ \text{mm}}{50\ \text{mm} + 35\ \text{mm}} = 21\ \text{mm}.$$

Note that the power is greater than that of either lens and the focal length is shorter. Now treat the combination as a single lens with focal length 21 mm.

Example. When a converging lens with power +5.0 m^{-1} and a diverging lens with power -5.0 m^{-1} are placed in contact, the light emerges unchanged: $P = +5.0\ \text{m}^{-1} + (-5.0\ \text{m}^{-1}) = 0.0\ \text{m}^{-1}$.

3-9 LENS DESIGN

Types of lenses

Lenses are sometimes described in terms of the shapes, **convex**, **concave** or **plane**, of their two refracting surfaces, as illustrated in figure 3.22.

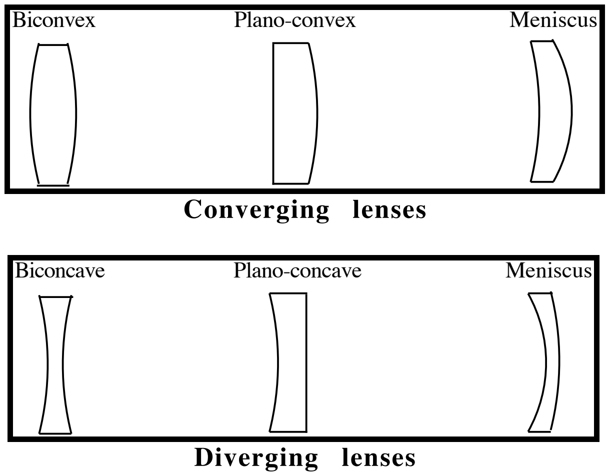


Figure 3.22. Naming some lens shapes
In all cases the refractive index of the lens is greater than that of the surrounding medium

Remember that all converging lenses are thicker in the middle than at the edges whereas diverging lenses are thin in the middle.

Lensmaker's formula

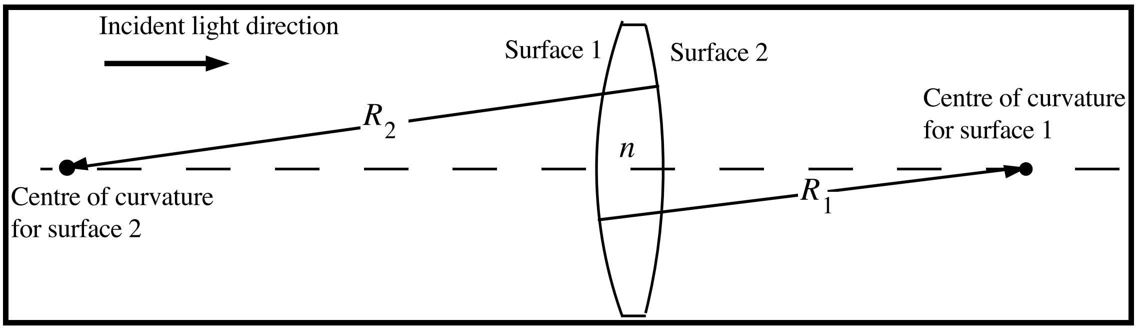


Figure 3.23. Specifying the radii of curvature of a lens

The focal length of a thin lens with spherical surfaces can be calculated from the curvature of the two faces and the refractive index of the lens material. Figure 3.23 shows a double convex lens made of material with a refractive index n surrounded by a medium of refractive index n'. The radii of curvature of the two surfaces are R_1 and R_2. The power is given by the **lensmaker's formula**:

$$P = \frac{1}{f} = \left(\frac{n}{n'} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) .$$

If the lens is surrounded by vacuum or air, n' will be 1.000 and the formula can be written as

$$P = \frac{1}{f} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) . \qquad \text{... (3.6)}$$

The sign convention needs to be extended, as follows, to give signs to the values of the radii. Imagine the light coming from a particular direction and label the first surface reached by the light number 1; the other is number 2. Measure each radius from the surface to the centre of curvature. An equivalent statement is that if a surface is convex to the incident light, its radius is positive and if it is concave, the radius is negative.

In the example shown in figure 3.23, imagine the light coming from the left so the radius, R_1, of the first refracting surface is positive and R_2 is negative. The negative value of $1/R_2$ will be subtracted from the positive value of $1/R_1$, which must give a positive answer. You will get the same result if you take the light coming in from the right and swap the labels 1 and 2.

Example

Find the power and focal length of a biconvex lens made of glass with refractive index 1.53 whose surfaces have radii of curvature 0.250 m and 0.400 m.

Answer

Use the lensmaker's formula with R_1 = 0.250 m; R_2 = - 0.400 m; n = 1.53 .

$$P = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = 0.53\left(\frac{1}{0.250 \text{ m}} - \frac{1}{-0.400\text{m}}\right)$$

$$= +3.4 \text{ m}^{-1}. \qquad \text{[Keep 3.445 in the calculator.]}$$

$$f = \frac{1}{P} = 0.29 \text{ m}.$$

You can check that you will get the same answer if you swap the sequence and put R_1 = 0.400 m; R_2 = -0.250 m.

3-10 ABERRATIONS

Aberrations are departures from the desired or ideal shape of an image produced by an optical system. They are not necessarily faults in the design of a lens but are due simply to the inability of any lens to produce perfect images from a wide range of objects. In fact the only perfect image-forming optical device is the plane mirror, which is unable to magnify.

Spherical aberration

When a wide parallel beam is focussed by a converging lens with spherical surfaces the focal length for rays which go through the outside margins of the lens is shorter than that for paraxial rays. See figure 3.24. Only rays which are close to the principal axis and nearly parallel to it are brought to a focus at a well-defined focal point F_p. So when a wide beam is used parts of the image will be out of focus.

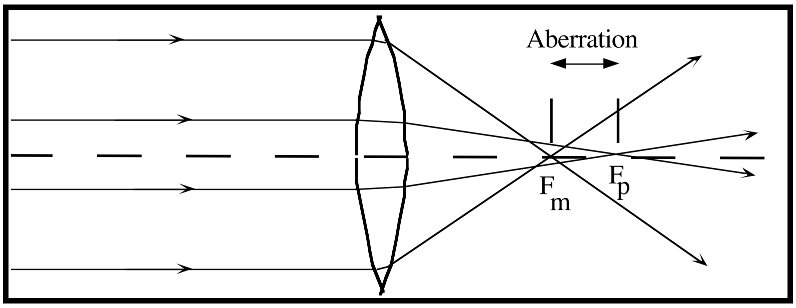


Figure 3.24. Longitudinal spherical aberration

Marginal rays come to a focus (F_m) before the focus of the paraxial rays (F_p) .

There are three ways to minimise spherical aberration.

1. Use a lens whose curved surfaces have specially computed non-spherical shapes. Such a lens is expensive to make and works only for a limited range of object positions. However, the aberrations can be minimised if the rays are bent approximately the same amount by both surfaces of the lens. So if you have a plano-convex lens spherical aberration will be less if you put the curved face rather than the plane face towards the light.
2. Correct the aberrations due to a positive lens by placing next to it a negative lens whose aberrations are in the opposite sense. This technique is not very efficient as you end up with a lens of very little power.

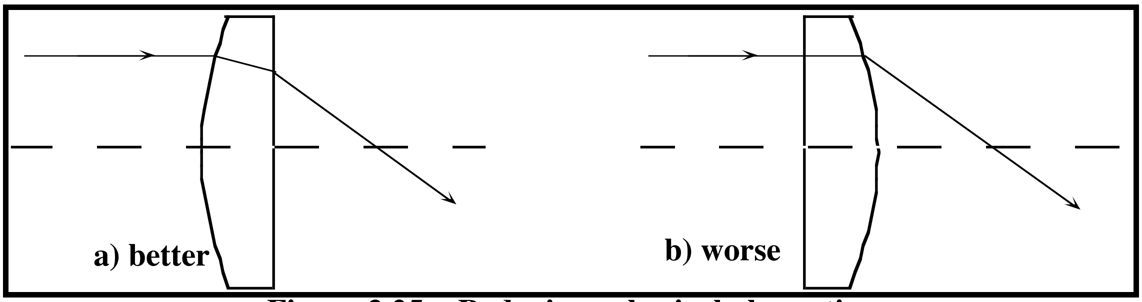


Figure 3.25. Reducing spherical aberration
It is better to have some refraction at both surfaces (a) rather than all at one surface (b).

3. Restrict the angles of incidence by using only the central section of the lens and placing the object close to the principal axis. All rays will then be paraxial rays, so the lens equation works for all object and image points. The disadvantage is that you don't get much light through the lens, so the image is not very bright.

Chromatic aberration

The focal length of a simple lens depends on the refractive index of the material but that varies with the frequency of light. So there are different focal points for different frequencies. Consequently the magnification will also change for different frequencies. The image of an object illuminated with white light will be surrounded by a coloured halo.

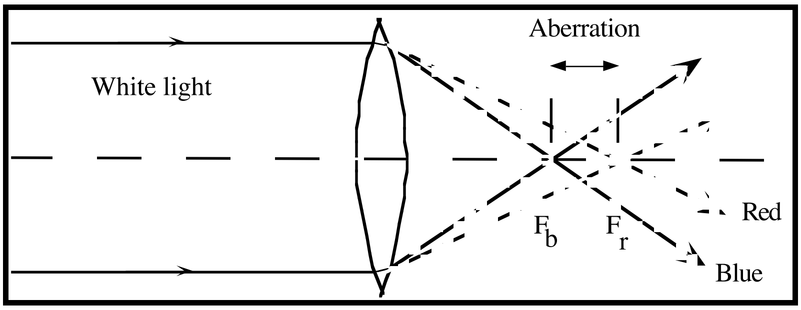


Figure 3.26. Longitudinal chromatic aberration
The focal length for blue light is less than that for red light.

There are two ways to minimise chromatic aberration.

1. Make the lens from a material whose dispersion, the variation in refractive index with wavelength, is small.

2. Make an **achromatic doublet** which consists of a converging lens and a diverging lens in contact. The dispersion in the two kinds of glass is different. A typical combination is a positive lens made of crown glass and a negative lens made of flint glass. An achromatic doublet can completely eliminate chromatic aberration for only two frequencies of light, but if those two frequencies are suitably chosen, then the aberration for other frequencies is reduced. Although an achromatic doublet consists of a positive and a negative lens, the combination still has some power.

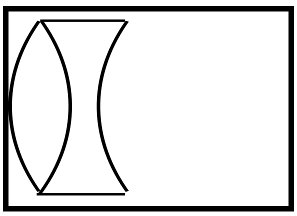


Figure 3.27.
An achromatic doublet

THINGS TO DO

Magnifying glass

Beg, borrow or buy a cheap magnifying glass. You can use it to observe many of the things described in this chapter and you can measure its focal length.

Measuring focal length

You can easily measure the focal length of a converging lens by forming a real image of a distant object on a small screen such as a piece of paper. (You may need a support to prop up the screen.) On a bright day in a room without too much light, hold the lens so that it faces a bright scene outside the room. Place your screen parallel to the principal plane of the lens and move the the lens until you get a sharp image of the outside scene on the screen. Measure the distance from the lens to the screen to get an estimate of the focal length. Strictly, you have measured the image distance, but if the object distance is large by comparison, then the image distance is practically equal to the focal length. A variation on the method is to put a piece of paper on the floor under an unshaded light globe and turn the light on. Put the lens between globe and paper and move it until you see a sharp image of the globe's filament on the paper.

On a sunny day you can use the lens to focus an image of the sun on to a piece of paper and measure the focal length. If you leave the image there long enough you can char the paper or even set it alight! That will show how the energy flux from the sun is being concentrated in the image. NEVER look through the lens at the sun - in fact you should never look directly at the sun at all.

Virtual and real images

Look through the magnifying glass at an object such as a pencil. Start with the pencil closer to the lens than the focal length. You should see a virtual, upright image. Gradually move the object further from the lens and note what happens to the image. There will be a position, or a region, where you lose the image. How far from the lens is the pencil when that happens? You would expect it to happen when the object is in the focal plane. Where do you expect the image to be?

If you keep moving the pencil further away you will find another, inverted, image. Although it is not formed on a screen it is a real image. To demonstrate the real image in space more forcefully make an image of a bright object such as a light globe on a piece of translucent paper. Look at the image from behind the paper. Now slowly slide the paper sideways out of the light path, letting part of the image go off the edge of the paper. You can still see the image there in space. The fact that the image could be picked up by the paper shows that it is real.

Combining lenses

If you can scrounge another magnifying glass, measure its focal length. Predict what the combined focal length will be if you put the two lenses in contact. Do it and check your prediction experimentally.

Making a lens

Find a clear bottle and look at things through it. Then fill it with water and look at the same things again. The images you see will be distorted but they are nevertheless images. Use the bottle of water to produce real images of bright objects on a piece of paper. Hence estimate the focal length of the cylindrical bottle-lens.

QUESTIONS

Q3.1 What is the size of the smallest plane mirror you can use to see yourself from head to toe without moving your head or the mirror?

Q3.2 When will a lens give a lateral magnification of +1.00? When will it give a magnification of -1.00? In each case what kind of lens can you use and where do you put the object?

Q3.3 Use the ray-tracing method to locate the image of an object placed in front of a converging lens between the focal point and the lens.

Q3.4 Consider any converging lens. Use the lens equation to demonstrate each of the following results.

(i) An object at an infinite distance from the lens gives a real image in the focal plane on the other side.

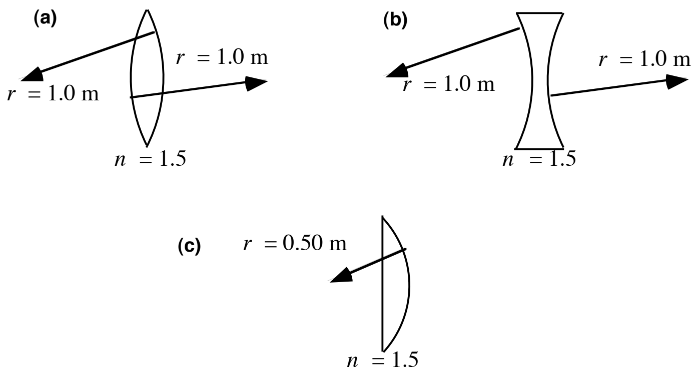
(ii) An object at distance $2f$ gives a real image at a distance $2f$ on the other side.

(iii) An object in the focal plane gives an image at an infinite distance.

(iv) An object between the focal plane and the lens gives a virtual image on the same side of the lens.

Q3.5 Consider any diverging lens. Use the lens equation to show that for all positions of the object, the image is virtual and is on the same side of the lens.

Q3.6 You can always tell whether a lens will be converging or diverging just by comparing its thickness in the middle with that at the edge. The sign convention used with the lensmaker's formula should give the same result. Check that this is so in the following three examples.



Q3.7 a) Calculate the focal length of the lens in Q3.6(a) when it is immersed in water. The refractive index of water is 1.33.

b) Will an air filled plastic bag used underwater by a skindiver serve as a converging or a diverging lens?

Q3.8 The definition of power of a lens matches intuitive ideas of powerful or less powerful lenses. For example suppose that we have two converging lenses. One needs to be held 0.10 m from a wall to focus a distant scene on the wall. The other needs to be held 0.20 m from the wall. The first is bending the parallel rays more strongly - it is the more powerful lens. Sketch the path of a number of rays to illustrate image formation in these two cases. Calculate the powers of the two lenses.

Q3.9 Suppose that a crown glass lens has a focal length of 0.100 m for a typical frequency of red light. What is its focal length for violet light? Refer to the graph of refractive index as a function of wavelength, figure 2.19.

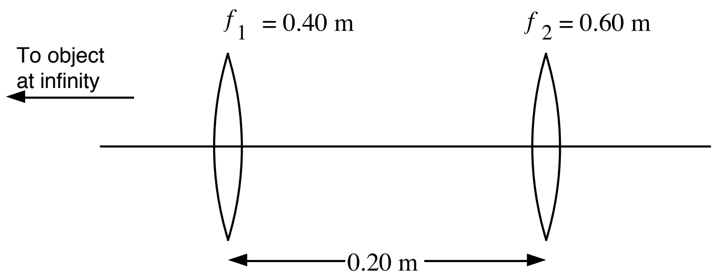
Q3.10 The angular size of the sun seen from earth is about 0.01 rad. A magnifying glass with a focal length of 150 mm and a diameter of 65 mm is used to produce an image of the sun on a card. What is diameter of the image?

Q3.11 Two lenses with powers of 5.0 m^{-1} and 4.0 m^{-1} are arranged so that their principal axes coincide. Light from an object 35 mm high goes first to the 5 dioptre lens. Find the position, nature (real or virtual, upright or inverted) and the size of the image when the object is placed 0.80 m from the first lens,

(a) when the lenses are 0.60 m apart and

(b) when they are 0.10 m apart.

Q3.12 Calculate, by successive applications of the lens formula $\frac{1}{o} + \frac{1}{i} = \frac{1}{f}$, the position of the image formed by the two-lens system below. The object is at infinity.



Discussion Questions

Q3.13 Explain why a lens has chromatic aberration, whereas a mirror does not.

Q3.14 A lens can produce really sharp images only if all the object points are in one plane - i.e. if they are all at about the same distance from the principal plane of the lens. A plane mirror produces sharp images for all objects at once - no matter how far away they are. Explain. How does the lateral magnification produced by a plane mirror depend on the object distance?

Q3.15 Can you devise an arrangement of mirrors which allows you to see the back of your head? Make a sketch and show some rays.

Q3.16 A converging beam of light strikes a plane mirror. Is the image real or virtual? Explain.

Q3.17 Under what conditions does a converging lens produce a real image? Can such a real image ever be upright? Explain.

Q3.18 Which way up is the image on the retina of your eye? Is that a problem?

Q3.19 Under what conditions is the lateral magnification of a lens infinite? Does an image exist if the magnification is infinite?

Q3.20 Can a diverging lens be used to produce a real image? Explain and discuss.

Q3.21 Can a virtual image be viewed on a screen? Can you photograph a virtual image? Discuss.

APPENDIX

Paraxial Rays

Paraxial rays must be close to the principal axis and also they must make small angles with the principal axis. The criterion for a small angle is that $\tan\alpha \approx \sin\alpha \approx \alpha$. The following table illustrates how good the approximation is.

α/degree	α/rad	$\sin\alpha$	$\tan\alpha$
5.00	0.0873	0.0872	0.0875
10.00	0.1745	0.1736	0.1763
15.00	0.2618	0.2588	0.2679
20.00	0.3491	0.3420	0.3640
30.00	0.5236	0.5000	0.5774

L4 INTERFERENCE

OBJECTIVES

Aims

When you have finished this chapter you should understand how the wave model of light can be used to explain the phenomenon of interference. You should be able to describe and explain, with words and and a minimal use of mathematical formulas, some idealised examples of interference, such as that produced by two coherent monochromatic point or line sources or monochromatic fringes in a thin film or wedge.

Minimum learning goals

1. Explain, interpret and use the terms:

   phase, phase difference, in phase, superposition, interference, interference pattern, path difference, optical path difference, coherence, coherent sources, incoherent sources, coherent waves, incoherent waves, fringe, order (of a fringe), amplitude, angular position, fringe separation, thin film, thin film interference.
2. State and explain the principle of superposition.
3. Describe and explain how interference between light waves can produce spatial patterns of varying intensities of light. Describe the conditions which are necessary for the formation of such interference patterns.

4(a) Use words, diagrams and graphs to describe interference patterns produced on a plane screen by two monochromatic, coherent, point or parallel-line sources.

(b) State and apply the relations among (i) wavelength, (ii) slit separation, (iii) angular and linear positions of light and dark fringes, and (iv) the distance from the slits to the screen.

5(a) Describe and explain interference of monochromatic light produced by reflection from thin films of uniform and non-uniform thickness.

(b) State and apply the conditions for maxima and minima in reflected monochromatic light for thin films and wedges.

6. Describe examples and applications of thin film interference
7. Describe and explain the appearance of interference patterns produced by double slit and thin-film arrangements with white light.

PRE-LECTURE

4-1 SUPERPOSITION OF WAVES

So far we have described the behaviour of light in terms of the behaviour of light rays which were usually straight lines although they could change direction at a boundary between two media. In this chapter and the next we look more closely at the wave nature of light and in doing so we will see some of the limitations of the ray model.

Revision

You should make sure that you still understand the idea of an oscillation and the terms *amplitude*, *phase* and *frequency* - see chapter FE7. Also re-read §1-2 in chapter L1.

Addition of waves

Interference, the topic of this chapter, is just the combination of waves. Interference of light waves can be described in terms of electric field (see chapter E1). To see how to calculate the combined effect of two waves think of two simple harmonic waves with the same angular frequency ω and equal amplitudes A as they both pass through the same point in space. Suppose that they have the same polarisation, which means that their electric fields are parallel (or antiparallel) and their electric fields at the point of interest can be described by the components E_1 and E_2 referred to the same direction. The two waves may, however, have different phases. The equations describing how these field components change with time at *one* fixed point in space can be written as

$$E_1 = A \sin(\omega t)$$

and

$$E_2 = A \sin(\omega t + \phi)$$

where ϕ is the phase difference between the waves. (These equations match the wave equation 1.1 with $x = 0$) The total field is found using the **principle of superposition** which says that the total field is equal to the vector sum of the individual wave fields. 'Vector sum' means that we have to take proper account of directions, by using components for example. In this simple example the directions are chosen so that each field can be described using only one component; hence simple algebraic addition gives the answer:*

$$E = E_1 + E_2 = 2A \cos\left(\frac{\phi}{2}\right) \sin\left(\omega t + \frac{\phi}{2}\right).$$

The equation can be rewritten as $E = A_t \sin\left(\omega t + \frac{\phi}{2}\right)$

where the new total amplitude is $A_t = 2A \cos\left(\frac{\phi}{2}\right)$.

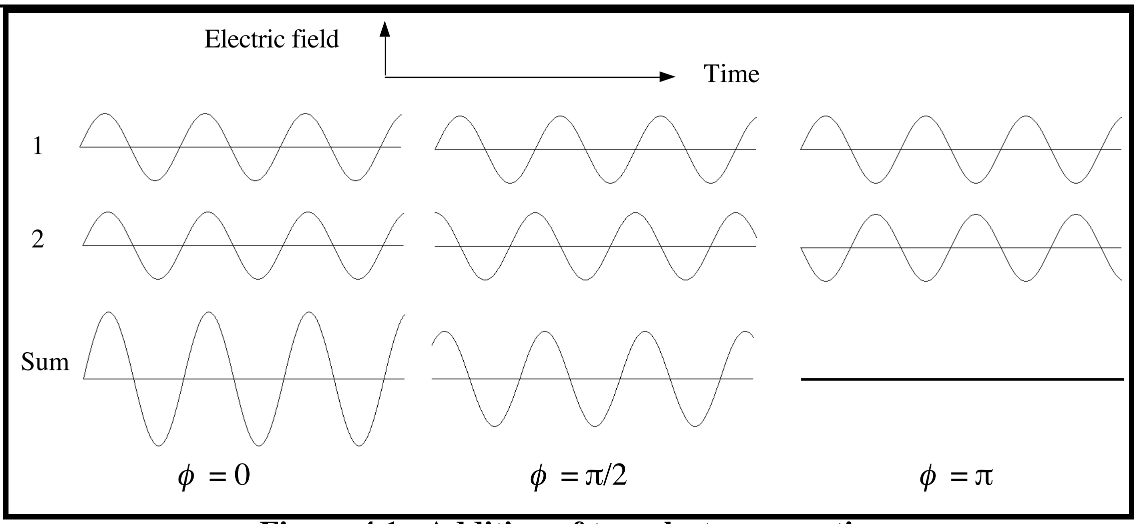


Figure 4.1. Addition of two electromagnetic waves
The amplitudes of the two elementary waves are equal. The sum of waves 1 and 2 is shown below.

Important things to note about this example are that the amplitude of the total electric field depends on ϕ, and the angular frequency of the resultant wave is still ω. Two special cases are of interest.

- If the waves are in phase, then $\phi = 0$, so the resultant amplitude is twice the amplitude of one of the waves:

$$2A \cos\left(\frac{\phi}{2}\right) = 2A .$$

Since the 'intensity' (irradiance in the case of light) of a wave is proportional to the square of its amplitude, the intensity of the resultant wave is four times the intensity of one of the original waves.

---

* The addition is carried out using the identity:

$$\sin\alpha + \sin\beta \equiv 2 \sin\left(\frac{\alpha + \beta}{2}\right) \cos\left(\frac{\alpha - \beta}{2}\right) .$$

- If the waves are out of phase by half a cycle, then $\phi = \pi$ so the total amplitude is

$$2A \cos\left(\frac{\phi}{2}\right) = 0$$

and the intensity of the resultant wave is zero.

Note

The terms 'constructive' and 'destructive' are sometimes used to describe these interference maxima and minima. Those names are avoided here because they can be misleading. Nothing is actually destroyed in wave interference; rather the effect of two waves is always additive, as expressed by the principle of superposition. Certainly energy is never destroyed - if energy seems to be missing from some place it always shows up somewhere else.

TEXT & LECTURE

4-2 YOUNG'S DOUBLE SLIT EXPERIMENT

The most famous of all demonstrations of the wave nature of light is Thomas Young's double slit experiment. Here we describe a modern version of Young's experiment as demonstrated in the video lecture. The arrangement (figure 4.2) consists of a source of light, a coloured filter, an opaque plate with a narrow slit cut into it, another plate with two narrow parallel slits in it and a white screen for viewing the light. Each of the slits is quite narrow, typically less than a tenth of a millimetre, and the two slits in the second plate are usually separated by only a fraction of a millimetre.

Light travels from the source through the filter to the first slit. From there it travels to the plate with two slits where much of it is blocked off but some can get through both the slits. Some of the light which finally gets through is intercepted by the screen.

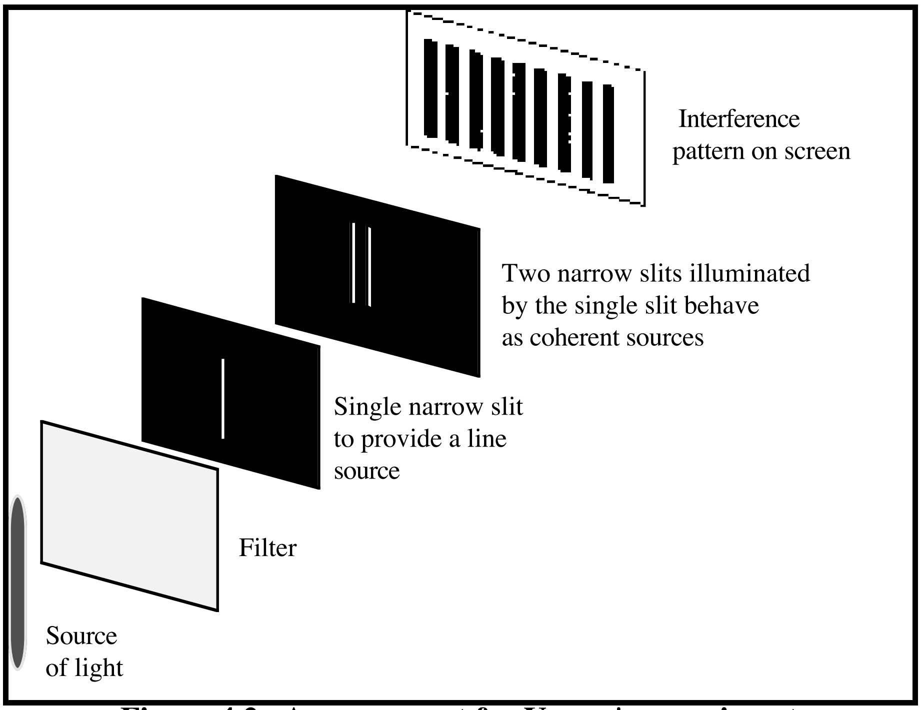


Figure 4.2. Arrangement for Young's experiment

When only one of the slits in the second plate is open there is a diffuse pool of light on the screen. This patch of light is wider than the slit that the light came through. This spreading out is called **diffraction**, a topic which will be taken up in the next chapter.

When light passes through each of the pair of slits in turn (keeping the other slit covered) you can see a pool of light on the screen. The areas covered by these two pools of light overlap considerably, so one would expect naively that with two slits open the resulting pool of light would just be a merging of the two pools already seen, bright in the middle and falling off at the edges. This expectation turns out to be quite wrong - instead of a continuous patch of light there is a pattern of light and dark stripes, called interference **fringes**. The ray model of light has no hope of accounting for that!

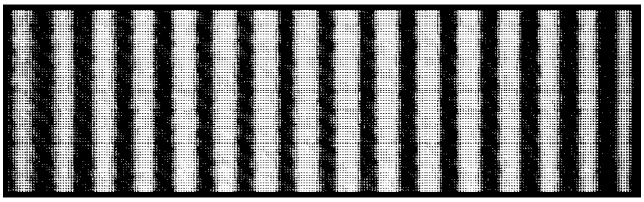


Figure 4.3. A twin slit interference pattern

At some places where there used to be light from each slit separately there is now darkness, but the energy in the light has not been destroyed. The brightest parts of the fringe pattern are now more than twice as bright as the brightest part of the light pool from one slit. The energy of the light has just been redistributed.

Although Young's original experiment used white light, the contrast between light and dark fringes is enhanced if a suitable coloured filter is used to restrict the range of frequencies in the light. The fringes are sharpest when a very narrow range of frequencies - monochromatic light - is used. (On the other hand using a broad range of frequencies produces some pretty multi-coloured effects - more about that later.)

The explanation of Young's experiment needs the wave model of light. To see how the wave model works it is useful to study a similar experiment using water waves instead of light, in which the superposition of waves can be seen directly.

Interference in water waves

Many aspects of wave behaviour can be observed in water waves. In a simple direct analogue of Young's experiment straight-line water waves (analogous to plane waves in three dimensions) are generated by a long paddle. The waves travel to a barrier with two narrow slots in it. When only one of the slots is open (figure 4.4) diffraction can be observed; the waves spread out instead of forming a sharp shadow. Each of the wave fronts coming out the other side has a roughly circular shape, but the amplitude of the waves is weaker at the sides than it is in the straight-through direction.

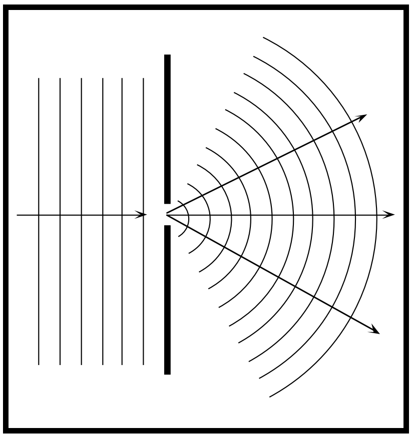


Figure 4.4. Diffraction of water waves at a hole

When both slots are open a new feature, called interference, is seen: there are some places where there is practically no wave disturbance and others where it is quite strong. See figure 4.5.

In the case of water waves the wave disturbance can be taken to be the displacement of the surface of the water from its equilibrium level. The amplitude of the resultant wave pattern varies from a minimum of zero at some places to a local maximum at other places. Furthermore, although the wave disturbance at any place varies in time, the amplitude at any single place is fixed. At a place where the amplitude is a maximum, say *A*, the wave disturbance varies from a minimum of -*A* to a maximum of +*A*. At a point where the amplitude is equal to zero, there is no net wave disturbance at any time. The amplitude of the resultant wave changes continuously from point to point between the locations of the maxima and minima.

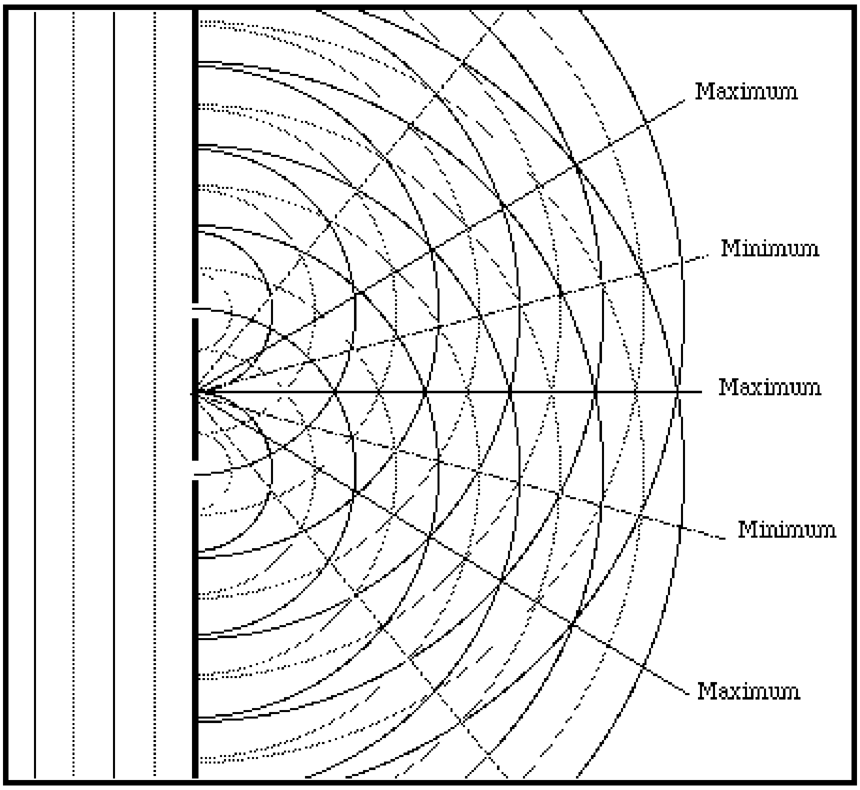


Figure 4.5. Interference of water waves from two slots
Maxima and minima in the amplitude occur at fixed locations, along lines radiating from the mid-point of the two slots. Wave crests are shown with full lines, troughs with broken lines.

These maxima and minima correspond to the fringes in Young's experiment.

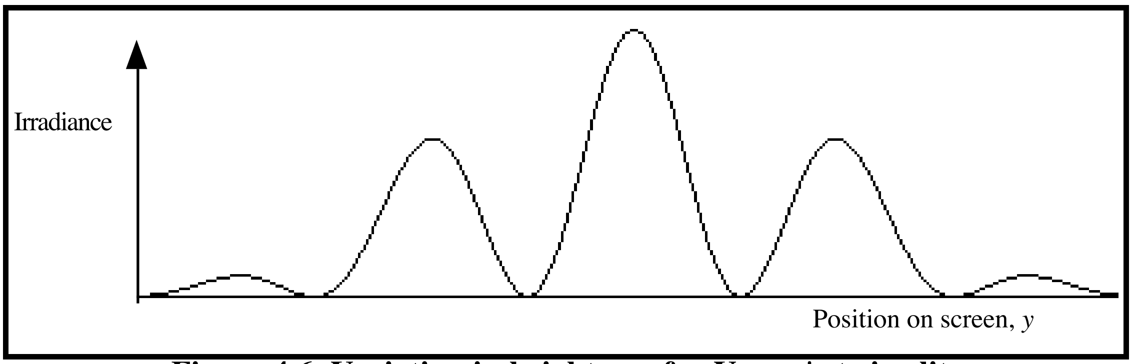


Figure 4.6. Variation in brightness for Young's twin slits

In this example the locations of the points of maximum and minimum amplitude all lie on approximately straight lines radiating from a point mid-way between the slots. If you look at the resultant wave at places in between the positions of the maxima and minima you will see that the amplitude varies smoothly with position. Figure 4.6 shows how the intensity varies with position on the screen for a typical Young's slits experiment using light.

4-3 SUPERPOSITION

The key to understanding interference is the **principle of superposition** which says simply that the combined effect of several waves at any place at a particular instant of time is given by the sum (vector sum if the wave property is a vector) of the wave property for the individual waves. The contribution from one wave is just that which would occur if the other waves were not there. In the case of the water waves the appropriate wave property is the linear displacement (change in position) of the surface of the water from its equilibrium position; for example if one wave produces an upward displacement of 2.0 mm and the other gives a downward displacement of 1.5 mm, the net effect is 0.5 mm up. In the case of light waves the appropriate wave property is the electric field (which can point in any direction perpendicular to the direction of propagation).

The interference pattern can be understood in terms of superposition. There are some places where a crest of one water wave arrives at exactly the same time as the trough of the wave from the other slot. At any one of those points the net displacement for these two waves is a minimum - zero if the amplitudes are equal. Although the crest of wave 1 and the trough of wave 2 move on, the sum of the two wave disturbances at the same fixed point in space remains zero. There are other points where a crest always arrives in step with a crest and a trough with a trough. At those points the amplitude of the resultant wave turns out to be the sum of the individual amplitudes. There are many other places where the individual waves add to give other values of the resultant amplitude. The resultant wave at any point depends on the phase difference between the individual waves (as well as their amplitudes).

Note that interference occurs only at places where both waves are present. Outside the region where the waves overlap there is no interference.

It is very important to note that although the two waves add up at any point in space that does not stop the progress of the waves. Each wave is quite unaffected by the other!

4-4 ANALYSIS OF TWIN SLIT INTERFERENCE

In Young's experiment with light, the function of the single slit is to ensure that the light falling on each of the pair of slits is **coherent**. Although the light consists of a continuous distribution of component waves with different frequencies and wavelengths, light reaching each of the twin slits from the narrow single slit has the same composition. If the pair of slits is placed symmetrically then any change in any component of the light, including any change in phase, occurs simultaneously at both slits. So the slits behave as **coherent sources**.

In the water wave experiment, the waves are much less complex, being essentially composed of only one frequency component. Since the original wave had straight wavefronts the waves emanating from the two slots are exactly in phase at all times. (Their amplitudes at the slots are also equal provided that the slots are equally wide.)

Conditions for interference maxima and minima

It is easy to calculate the points in space where maxima and minima in the interference pattern occur. The analysis is essentially the same for the water waves example and for the Young's twin slits because both can be treated in two dimensions. (It is assumed that the Young's slits are very long compared with their width and separation.)

The wave amplitude at some point P depends on the phase difference between the two interfering waves. If the waves are in phase at their sources (the slits in the case of Young's experiment), then the phase difference at P is determined by the difference in times taken for the light to get from the sources to P. That time difference, in turn, depends on the speed of the waves and the the difference in the distances, called the **path difference**, travelled by the two waves. In the case of light, we can say that the phase difference is proportional to the **optical path difference**, which is the product of the actual path difference and the refractive index of the medium.

$$\text{Optical path length} = n \times (\text{geometrical path length}).$$

Since we usually consider Young's experiment in air, the optical and geometrical path differences are essentially equal. The path difference for Young's slits is labelled D in figure 4.7. It is very important to note that this figure is not to scale. This analysis is valid only if the screen is a long way from the slits and if the point P is close the central axis. If those conditions are satisfied rays from each of the slits to P are almost parallel and it is said that the experiment satisfies Fraunhofer conditions. (The general case in which angles and distances are not so small, Fresnel conditions, is very difficult to analyse.)

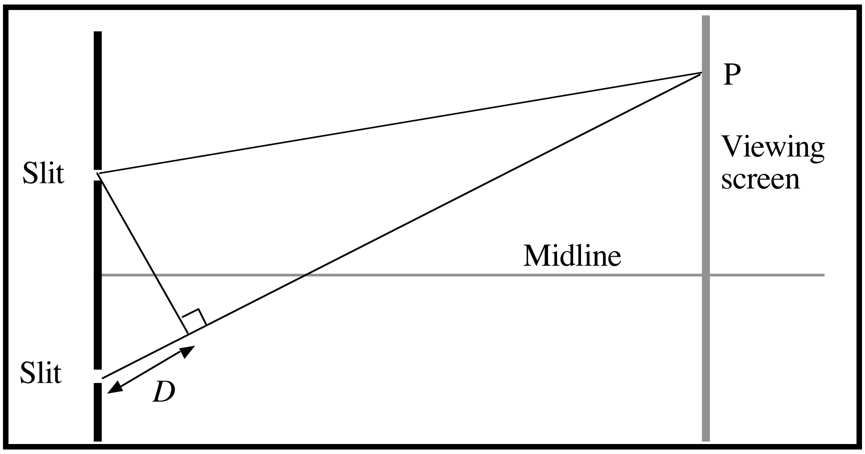


Figure 4.7. Path difference for Young's twin slits
Not to scale: In reality the two rays are almost parallel.

If the optical path difference, D, is equal to a whole number (n) of wavelengths then the phase difference will be n times 2π (corresponding to n wave cycles) and the two waves will be exactly in phase. That produces a maximum in the amplitude of the resultant wave.

For a maximum: $D = m\lambda$ (for $m = 0, \pm1, \pm2, \pm3, ...$). ... (4.1a)

The value of m is called the **order** of the bright fringe; the fringe in the middle is the zero-order fringe.

A minimum in the amplitude will occur if the optical path difference is equal to an odd number of half-wavelengths.

For a minimum: $D = (m + \frac{1}{2})\lambda$ (for $m = 0, \pm1, \pm2, \pm3, ..$). ...(4.1b)

Since the irradiance ("intensity") in the interference pattern is proportional to the square of the wave amplitude, maxima and minima in the intensity occur at the same places as the maxima and minima in the wave amplitude. The relation between phase difference ϕ and path difference D which applies at all points (including those between the intensity maxima and minima) is

$$\frac{\phi}{2\pi} = \frac{D}{\lambda} . \qquad ... (4.2)$$

These results apply to all kinds of interference between two elementary waves, not just the Young's slits experiment.

Location of maxima and minima

The location of points in the interference pattern is most conveniently specified in terms of the angular position θ of the point P (see figure 4.8). The **angular position** of P is measured from the midline between the slits.

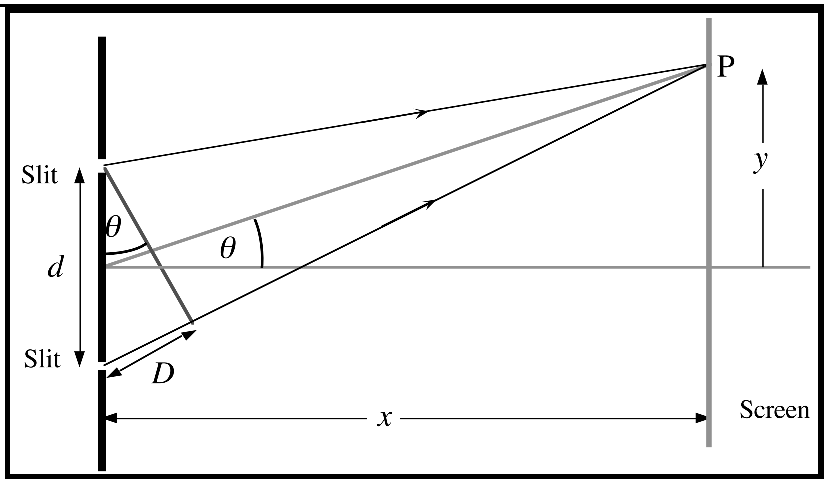


Figure 4.8. Geometry of the twin slits arrangement
Not to scale: distances along the y-axis are grossly exaggerated.

It can be seen from the diagram that $D \approx d\sin\theta$ so the conditions become:

for a maximum $$m\lambda \approx d\sin\theta \qquad \text{... (4.3a)}$$

and for a minimum $$(m + \tfrac{1}{2})\lambda \approx d\sin\theta\ . \qquad \text{... (4.3b)}$$

These formulas give the angular positions of the bright and dark fringes in the space beyond the slits. They are perfectly accurate only for small values of the angle θ.

If the fringes are viewed on a screen at a long distance from the slits these formulas (4.3) can be rewritten approximately in terms of the linear position y of the point P on the screen. First note that the position coordinate, y, of P can be written as $y = x\tan\theta$. For small values of the angle, as considered here, $\tan\theta \approx \sin\theta$, so the conditions can be rewritten:

for a maximum $$y \approx \frac{m\lambda x}{d} \qquad \text{... (4.4a)}$$

and for a minimum $$y \approx \frac{(m + \frac{1}{2})\lambda x}{d}\ . \qquad \text{... (4.4b)}$$

It follows from these equations that the fringe spacing, the distance between two successive light fringes or two successive dark lines on the screen, is given by

$$\Delta y \approx \frac{\lambda x}{d}\ . \qquad \text{... (4.4c)}$$

These results illustrate why you can see good fringe patterns only if the range of wavelengths in the light is small. The angular positions of the maxima and minima depend on the wavelength, so if the wavelength doesn't have a well defined value then the fringes are not well-defined either.

Coherence of light sources

In the water wave experiment there is no problem with the coherence of the two sources. Both sets of waves are produced by splitting one continuous wave. On the other hand light from an ordinary source can normally be described as a superposition of a vast number of elementary waves, which have a continuous range of frequencies and wavelengths. These elementary waves can be related to photons emitted by atoms or molecules in the light source. Although each elementary wave has a fairly well-defined frequency, it does not last for long. Since the emission of elementary waves from the light source is entirely uncorrelated, the source is said to be incoherent. If the Young's slits were illuminated directly by an ordinary lamp, coherence between individual waves arriving at the two slits would exist only for extremely short times, and the interference fringe pattern would jump around very rapidly. Although an interference pattern would exist it would not stay in one place long enough to be seen. The important feature of Young's experiment is the production of long-term coherence by splitting the wavefronts of each and every elementary wave so that the same phase relationship between waves from the two slits persists for a relatively long time. The coherence is achieved by using the single narrow slit as a common source which illuminates both the twin slits. Each wave from the first slit produces two coherent parts at the twin slits, so that whatever phase fluctuations there are among the elementary waves, exactly the same fluctuations occur at both slits.

Laser light

One of the special features of light from a laser is that it is highly coherent. Therefore Young's experiment can be done by sending light from a laser directly onto the twin slits, without using the single slit. Another advantage of laser light is that it is highly monochromatic - the spread of wave frequencies (the bandwidth) is much smaller than anything that can be obtained from a lamp and coloured filters. In these respects the water wave interference experiment is analogous to Young's experiment with a laser.

Behaviour of the interference pattern

If the fringes are observed on a screen a long way from the slits (x » d) the irradiance of the fringes is fairly uniform (figure 4.9). In that case the maximum irradiance is about four times the irradiance due to one source alone. The "missing" intensity from the dark fringes has gone into the extra intensity in the bright fringes, so that there is no violation of energy conservation. On a distant screen the fringes are uniformly spaced and the separation is approximately equal to $\lambda x/d$.

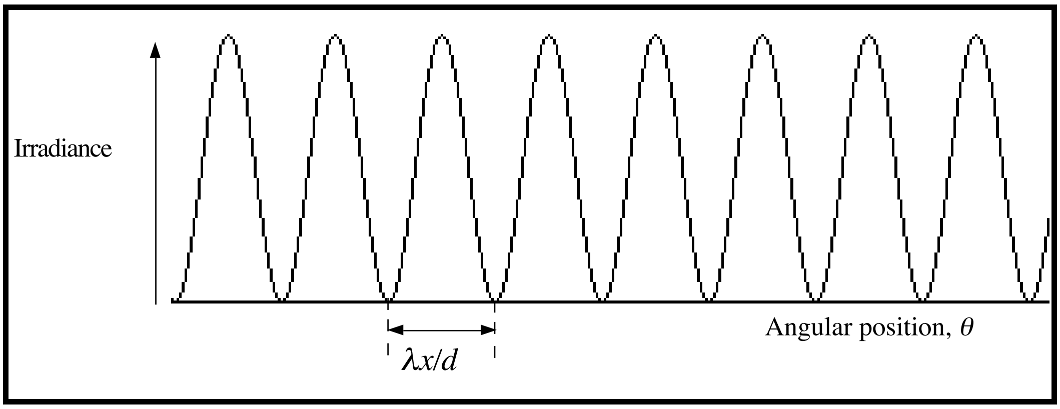


Figure 4.9. Idealised intensity pattern for Young's twin slits

In this example the screen is a long way from the slits and the width of a slit is much less than a wavelength.

Points to note

- For a fixed wavelength the fringe spacing varies inversely with the separation of the slits. If the slits are moved further apart then the fringes get closer together.
- For a fixed separation of the slits, the fringe spacing is proportional to the wavelength (provided that the approximations stated above are satisfied).
- There is always a bright fringe (order zero) on the central axis. So when white light is used the only bright fringe which shows up strongly is the central one because its location does not depend on wavelength. Since the fringe spacing depends on wavelength, the total pattern for white light is a continuous mess of overlapping fringe patterns. The overall effect is white light in the middle and "washed-out" coloured fringes on the sides.

4-5 THIN FILM INTERFERENCE

Interference patterns can be observed whenever waves from two or more coherent sources come together. In Young's experiment the waves came from two separate sources but in **thin film interference**, the waves come from one source. One wavefront is split into two parts which are recombined after traversing different paths. Examples of thin film interference occur in oil slicks, soap bubbles and the thin layer of air trapped between two glass slabs. Here thin film means a layer of transparent material no thicker than several wavelengths of light.

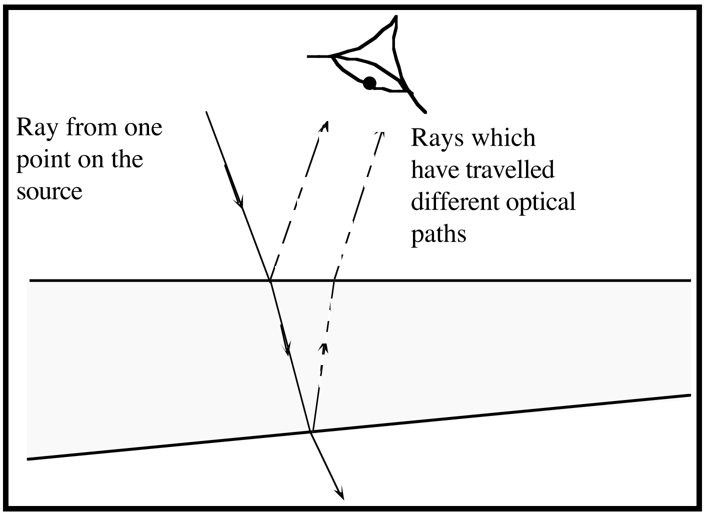


Figure 4.10

Thin film interference

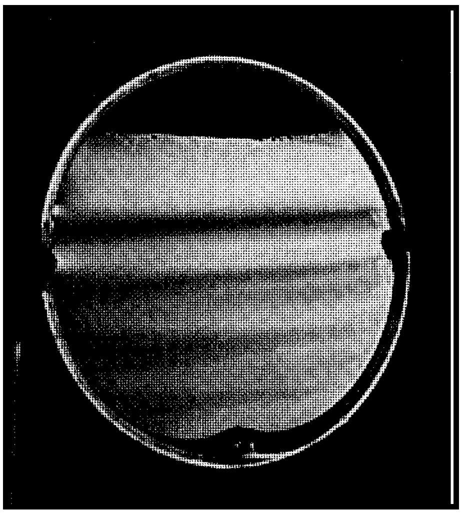


Figure 4.11

Interference fringes in a soap film

When light strikes one boundary of the film, some of it will be reflected and some will be transmitted through the film to the second boundary where another partial reflection will occur (figure 4.10). This process, partial reflection back and forth within the film and partial transmission, continues until the reflected portion of the light gets too weak to be noticed. The interference effects come about when parts of the light which have travelled through different optical paths come together again. Usually that will happen when the light enters the eye.* Thus for example, light

* When light rays are brought to a focus either by the eye or a lens, there is no extra optical path difference introduced so the focussing has no effect on the conditions for the location of the interference fringes.

reflected back from the top surface of the film can interfere with light which has been reflected once from the bottom surface and is refracted at the top surface.

The interference effect for monochromatic light, light or dark or somewhere in between, is determined by the amplitudes of the interfering waves and their phase difference. The conditions for a maximum or minimum in the irradiance are the same as before: a phase difference of $m(2\pi)$ gives a maximum and a phase difference of $(m + \frac{1}{2})(2\pi)$ produces a minimum.

Change of phase at reflection

A new phenomenon reveals itself here. A straightforward interpretation of the conditions for interference maxima and minima solely in terms of optical path difference gives the wrong answer! Two examples illustrate this point. In a very thin soap film it is possible to get a film thickness which is much less than one wavelength. So the path difference between light reflected from the two surfaces of the film is much less than a wavelength and the corresponding phase difference will be almost zero. A zero phase difference should produce brightness, but the opposite is observed - when the film is very thin there is no reflection at all! The explanation is that whenever a light wave is reflected at a boundary where the refractive index increases, its phase jumps by π or half a cycle. In the case of the soap film, the light reflected from the first surface, air to soapy water, suffers a phase change, but light reflected at the water-air boundary has no phase change. You can observe this effect yourself in soap bubbles. Carefully watch the top of a bubble as the water drains away. As the film gets thinner you will see a changing pattern of coloured fringes. Just before the bubble breaks, the thinnest part of the film looks black - indicating no net reflection.

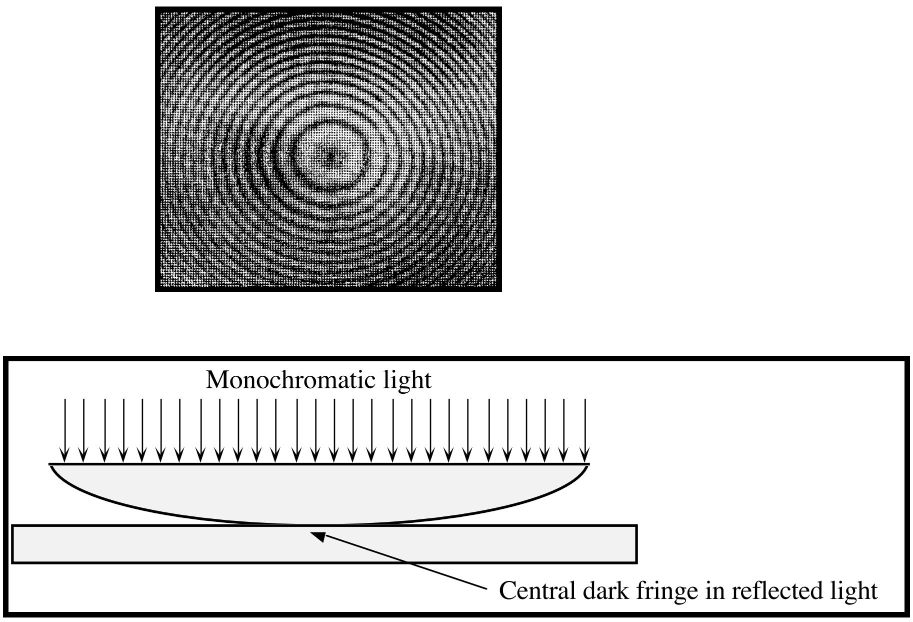


Figure 4.12. Newton's rings

The other example is a thin film interference pattern called Newton's rings which are formed using a curved glass lens resting on a flat glass slab (figure 4.12). The thin film is the air between the lens and the slab. The important feature is that where the optical path difference is zero, right in the middle of the pattern where the lens actually touches the slab, there is darkness instead of a bright fringe. The dark spot can be explained by saying that there is a phase change of π in the light reflected at the boundary between air and glass.

Analysis of thin film interference

The conditions for finding bright or dark fringes in a thin film clearly depend on the angle of incidence of the light, but a useful approximation can be worked out assuming that the incident light rays are normal to the surface, or almost so. In that case the optical path difference between parts of an elementary wave reflected from the top and bottom surfaces of a film is just $2nb$, where b is the thickness and n is the refractive index (figure 4.13).

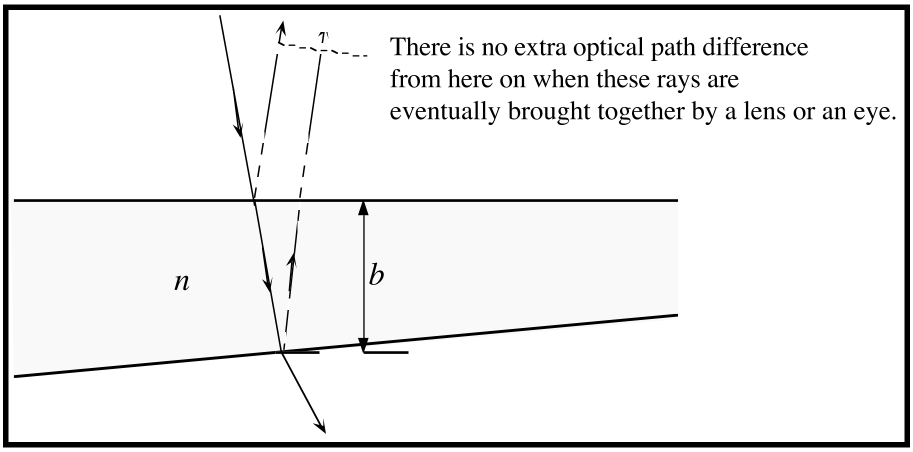


Figure 4.13. Calculating the optical path difference
For near normal incidence, $D = 2nb$.

To work out the conditions for bright and dark fringes you have to include the effect of phase changes at reflection. Each phase change of π has the same effect as the addition of an extra half wavelength of optical path.

No net phase change at reflection

If there is no phase change at either boundary or a phase change at both boundaries (for example: a film of water on glass), the conditions for maxima and minima are

for a bright fringe: $$2nb = m\lambda \qquad \text{... (4.5a)}$$

and for a dark fringe: $$2nb = \left(m + \tfrac{1}{2}\right)\lambda \qquad (m = 0, 1, 2, 3, ...). \qquad \text{... (4.5b)}$$

Phase change at one boundary

Where there is a phase change at only one boundary (for example an air film trapped between two glass plates or a soap bubble) the interference conditions depend on both the thickness and the phase change at reflection. The conditions are simply interchanged:

for a bright fringe: $$2nb = \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{... (4.6a)}$$

and for a dark fringe: $$2nb = m\lambda \qquad (m = 0, 1, 2, 3, ...). \qquad \text{...(4.6b)}$$

Notes

- There is no point in trying to memorise these equations. It is better to work them out when you need them by combining the conditions expressed in terms of phase difference (equations 4.1a and 4.1b) with the phase changes at reflection and the relation between optical path and phase difference.
- It is important to remember that the value of wavelength to be used in these relations is the wavelength in vacuum (or air). If you need to know the value of the wavelength, λ_m, in the medium with refractive index n it can be calculated using the relation

$$\frac{\lambda}{\lambda_m} = n \, .$$

Example: fringe patterns in wedges

If two flat glass plates are allowed to touch at one edge and are separated by a small object such as a thin wire at the opposite edge, the space between the plates contains a wedge-shaped thin film of air (figure 4.14).

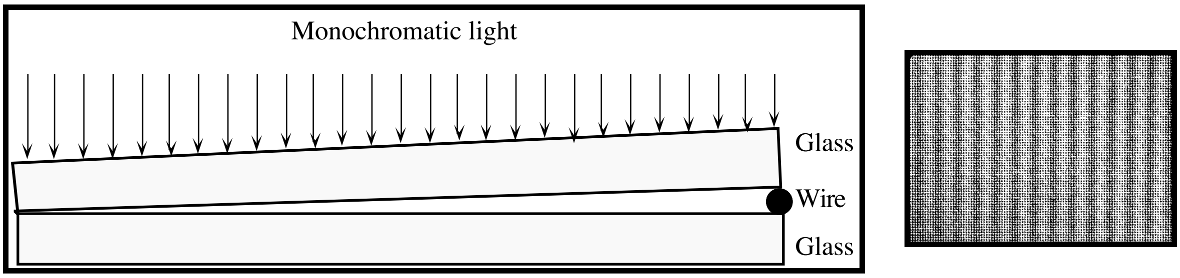


Figure 4.14. Interference fringes in a wedge of air
The vertical scale is greatly exaggerated.

When monochromatic light is shone down on to this arrangement, interference fringes will be observed in the reflected light. Since the existence of a bright or dark fringe depends on the thickness of the film at a particular place, fringes will be seen at various places across the air wedge. The analysis above shows that the spacing of the fringes is proportional to the wavelength. For a given wavelength each fringe follows a line or contour of constant thickness in the air film. If you follow across the fringe pattern, the thickness of the film will change by $\lambda/2n$ as you go from one fringe to the next. If the medium in the wedge is air then $n = 1.000$, so the fringe spacing corresponds to a change in thickness of $\lambda/2$. This gives a way of measuring the thickness of the thin object used to prop the plates apart if you already know the wavelength: just count the total number of fringes across the whole wedge and multiply by $\lambda/2$. The resolution in this measurement is about half a wavelength, or better, depending on how well you can estimate fractions of a fringe. Alternatively, you could use this method and a wire of known diameter to find the wavelength.

Localisation of the fringes

Although a narrow light source (the single slit) is needed to produce coherence in Young's experiment, thin film fringes can be formed using extended light sources, even daylight from the sky. The difference is that in thin film interference every incident wavefront, no matter where it comes from, is split into two wavefronts when it meets the first surface of the film. When the two waves meet again they have a definite phase relationship so that interference is seen to occur. The phase difference between the waves is locally constant and the fringes are said to be localised. You can see that when you look at thin film fringes - they appear to be located in (or just behind) the film.

Coloured fringes

If a thin film is illuminated with white light the reflected light will contain a continuous range of fringe patterns corresponding to the spectrum of wavelengths in the light. You do not, however, see the same colours as the pure spectrum like a rainbow. Instead the colours are formed by subtraction from the white light. For example, at a place where the film thickness is just right for a dark fringe in the green you will see white light minus green, which leaves the red end and the blue end of the spectrum; the resulting visual sensation is purple.

Where does the energy go?

There is a puzzle that needs to be answered: what happens to the energy of the light when the reflected light is removed by interference? The energy cannot be destroyed so it must go somewhere else - it is transmitted through the film instead of being reflected. As in the case of Young's experiment, the energy is rearranged in space but it is never destroyed. If you are used to thinking of energy as a kind of fluid, then that idea may be hard to understand. However experimental evidence supports the wave theory, so the "fluid" model of energy needs to be abandoned. Energy is *not* like matter, it does not have to flow continuously through space. Another way of resolving the problem is to say that the principle of superposition (just adding things up) works for electric fields but it does not apply to energy or wave intensity.

4-6 APPLICATIONS OF THIN FILM INTERFERENCE

Testing for flatness

Given a slab with a very accurately flat surface, thin film interference can be used to test the flatness of the another surface. (At least one of the two objects needs to be transparent.) Interference fringes formed by the thin film of air between the surfaces gives a contour map of variations in the height of the surface being tested. The contour interval is equal to half a wavelength of the light in the gap.

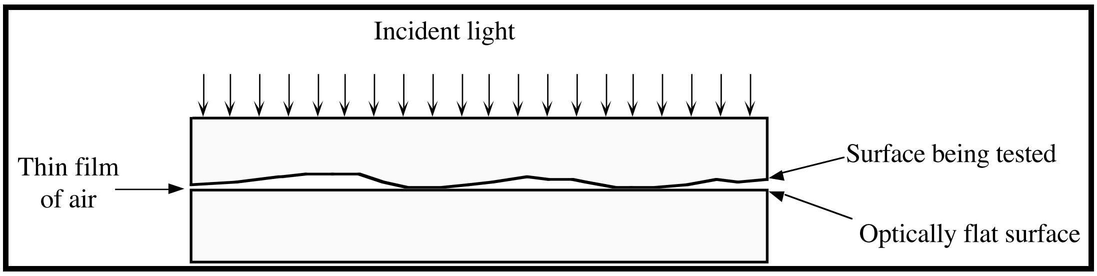


Figure 4.15. Testing for flatness

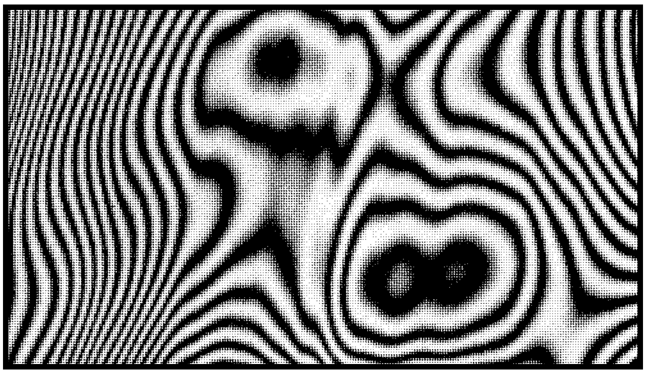


Figure 4.16.
Thin film contour fringes

Blooming of lenses

A common application of thin film interference is in anti-reflection coatings on lenses that are used in cameras, microscopes and other optical instruments. A modern lens system may have as many as ten glass surfaces each with a reflectivity of about 5%. Without some kind of treatment about half the light entering such a lens system would be reflected instead of going on to form the final image. Apart from the loss of brightness involved, multiple reflections in an optical system can also degrade the quality of an image.

The amount of light reflected from each surface can be greatly reduced using the technique of **blooming**, that is the deposition of an anti-reflection coating. Interference in the reflected light means that light is transmitted instead of being reflected. The choice of material for the coating is important. Clearly it must be transparent, but it should also result in approximately equal reflectivities at both surfaces, so that the reflected waves (at a chosen wavelength) can completely cancel each other. Cancellation is achieved exactly when the refractive index of the coating is equal to the geometric mean of the refractive indices of the air and the glass: $n_2 = \sqrt{n_1 n_3}$. See figure 4.17. However it is not easy to find materials with exactly the right properties, so in practice a compromise is needed. Magnesium fluoride, which has a refractive index of 1.38, is often used.

The thickness of the coating is chosen to work best for light of a wavelength near the middle of the visible spectrum, for example a wavelength of 500 nm corresponding to yellow-green light. In that case the lens still reflects some light in the blue and red so it looks purple in reflected light. The refractive index of the coating is between that of air and glass so there is a phase change at both reflections. At the chosen wavelength we require $2n_2 b = \left(m + \frac{1}{2}\right)\lambda$ for no reflection. With $m = 0$, the film thickness is a quarter of a wavelength.

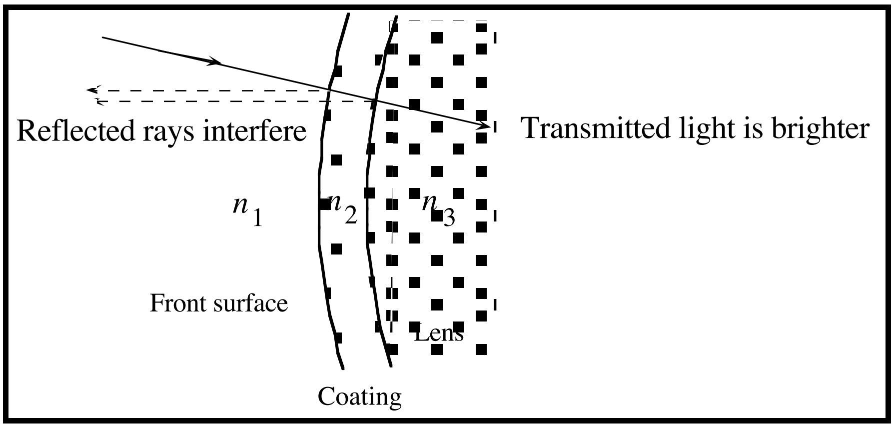


Figure 4.17. Anti-reflection lens coating

THINGS TO DO

Look for examples of interference in your environment. The colours in oil slicks are an example of thin-film interference. Next time that you see one make a note of the colours and their sequence. Are they the same as the colours of the rainbow? Can you explain the differences or similarities? Other examples of thin-film interference may be found in soap bubbles, the feathers of some birds and opals.

You can make a thin film using two sheets of transparency film like that used on overhead projectors. Just place the sheets together and look at the reflected light. A dark background behind the sheets will help. You should be able to see coloured contour fringes which map the thickness of the air between the sheets. To enhance the effect place the two sheets on a hard surface and by rubbing something like a handkerchief over them, try to squeeze the air out of the gap. What do you see now? See what happens when you press your finger on one part of the top sheet. Does the angle at which you look make any difference? Does the angle of the incident light matter? Look through the sheets and try to see the interference in the transmitted light; why is that harder to see?

Observe the colour of the light reflected from various camera lenses. Can you explain the colour? Is the colour the same for all lenses?

QUESTIONS

Q4.1 Two coherent monochromatic sources produce an interference pattern on a screen. What happens to the pattern if
a) the wavelength is doubled,
b) the distance between the two sources is doubled?

Q4.2 A Young's double slit experiment consists of two slits 0.10 mm apart and a screen at a distance of 1.0 m.
Calculate the separation of blue light (λ = 400 nm) fringes.
Calculate the separation of red light (λ = 600 nm) fringes.
Sketch the pattern near the centre of the screen.
Can you deduce anything about interference patterns in white light?

Q4.3 Suppose that two coherent sources have a constant phase difference ϕ which is not equal to zero. How do the conditions for interference maxima and minima change?

Q4.4 If you dip a wire frame into soapy water and take it out, you will see colours in the thin film of water. If the light source is behind you, you will see thin film interference.
As the water drains away from the top, the colours there will disappear. Can you explain why and estimate the thickness of the soap film at the top?

Q4.5 The edges of an oil patch on the road appear coloured. Can you explain why and estimate the thickness of the film there?
Why is it not possible to see fringes over the whole of the film?

Q4.6 The wavelength of a spectral line was measured using a Young's twin slit set-up with a suitable filter to select the appropriate line. The separation of the twin slits was 0.523 mm and the screen was placed 1.22 m from the slits. The distance between the two second-order bright fringes on the screen was measured as 5.50 mm.
Calculate the wavelength of the spectral line. What would happen to the separation of the fringes if the distance to the screen were doubled?

Q4.7 A lens with refractive index 1.53 is to be coated with magnesium fluoride (refractive index 1.38) in order to eliminate reflections at the peak sensitivity of the human eye. (See chapter L1.) How thick should the coating be?

Discussion questions

Q4.8 Could Young's twin slit experiment be done with sound waves? Discuss.

Q4.9 What would happen to the fringe separation if Young's experiment were done entirely inside a big tank of water? Explain.

Q4.10 Is it possible that you could observe interference fringes in the light from the two headlamps of a distant car? Explain.

Q4.11 Suppose that instead of putting a filter between the lamp and the single slit in Young's experiment, a red filter were put over one of the twin slits and a blue filter over the other. What effect would that have on the fringes?

Q4.12 What happens if you remove the screen with the single slit in figure 4.2 ?

Q4.13 When an oil slick spreads out on water, reflections are brightest where the oil is thinnest. What can you deduce from that?

Q4.14 Bloomed lenses look coloured. Does that mean that the lens coating is made of a coloured material?

L5 DIFFRACTION

OBJECTIVES

Aims

From this chapter you should gain an understanding of the process of diffraction and its role in producing distinctive interference patterns. You should also aim to understand how the effects of diffraction can affect and limit the formation of images. As the classical example of diffraction, you should be able to describe and explain the Rayleigh criterion and apply it to simple examples.

Minimum learning goals

1. Explain, interpret and use the terms:

   diffraction, diffraction pattern, Fresnel diffraction, Fraunhofer diffraction, angular resolution, Rayleigh criterion, diffraction envelope, principal maximum, secondary maxima, double slit, diffraction grating.
2. Describe qualitatively the diffraction patterns produced in monochromatic light by single slits, rectangular apertures, circular apertures, double slits and diffraction gratings.
3. Describe the interference pattern produced by monochromatic light and a grating in terms of the interference pattern produced by a set of line sources modulated by a diffraction envelope.
4. State and apply the formulas for the angular widths of the central maxima in the Fraunhofer diffraction patterns of a single slit and a circular hole.
5. State the Rayleigh criterion, explain its purpose and apply it to simple examples.
6. Describe how wavelength, slit width, slit spacing and number of slits affect the Fraunhofer diffraction patterns produced by multiple slits and gratings.
7. State and apply the formula for the angular positions of maxima in the Fraunhofer diffraction patterns produced by multiple slits and gratings.

PRE-LECTURE

5-1 SHADOWS

In the ray model we suppose that when light travels through a homogeneous medium it moves along straight lines. That observation is often called the law of rectilinear propagation. The existence of shadows is good evidence for the ray model of light. When light from a small (or 'point') source goes past the edges of an opaque object it keeps going in a straight line, leaving the space behind the object dark (figure 5.1).

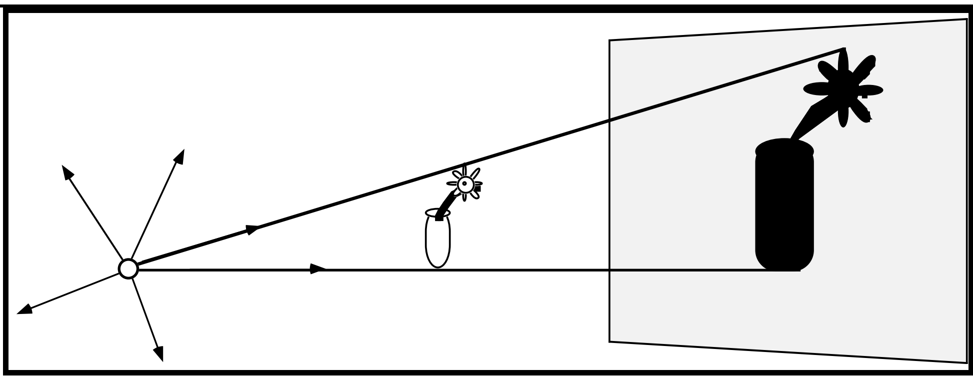


Figure 5.1. Straight line propagation of light
A small source of light produces sharp shadows.

When the light reaches some other surface the boundary between light and dark is quite sharp. (On the other hand if the light comes from an extended source the shadow is not so sharp - there is a region of partial shadow surrounding the total shadow.)

It was not until the about the beginning of the nineteenth century that scientists noticed that shadows are not really perfectly sharp. Looked at on a small enough scale the edge of a shadow is not just fuzzy, as you might expect for an extended source of light, but there are also light and dark striations or fringes around the edge of the shadow. Even more remarkable is the slightly later discovery that there is always a tiny bright spot right in the middle of the shadow cast by a circular object (figure 5.2). The fringes and the bright spot cannot be understood in terms of the ray model - the explanation lies in the wave model. According to the wave theory, the fringes are formed by the **diffraction** or bending of light waves around the edges of objects and the subsequent interference of the diffracted waves. The diffraction of water waves at a hole in a barrier was shown in the video lecture on interference (L4) and it is sketched in figure 4.4.

One of the effects of diffraction is the production of interference or fringe patterns. Although these patterns are actually interference patterns in the same sense as those we have already discussed, when they are produced by the bending of light around obstacles or apertures (holes) they are called **diffraction patterns**.

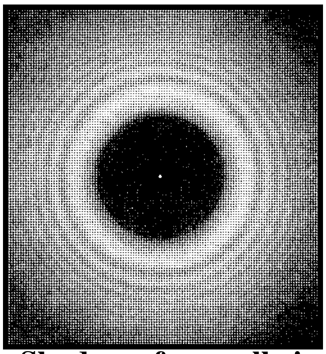


Figure 5.2. Shadow of a small circular object
The edges of the shadow contain fringes and there is a small bright spot in the middle.

Young's twin slits experiment shows that light does not always travel along straight lines. You can see that it must bend somewhere by looking at figure 5.3. Since there is light at the middle of the screen but no straight through path from the source to the screen, the light which gets there must somehow be going around corners. What must be happening is that after the light reaches the first slit it then spreads out so that some of it reaches the other two slits. Then, having passed through those slits it spreads out again in many directions until it reaches the screen. This behaviour is typical of waves, but not of particles.

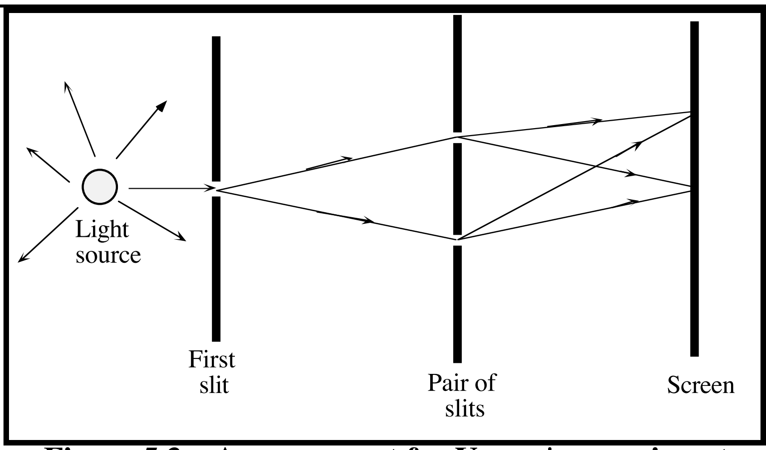


Figure 5.3. Arrangement for Young's experiment
Light cannot be travelling in a straight line all the way from the source to the screen.

The connection between waves and diffraction is much more noticeable for sound than it is for light. The observation that sound easily travels around corners is strong evidence for the wave nature of sound.

TEXT & LECTURE

5-2 HUYGENS' CONSTRUCTION

A way of describing how diffraction occurs was invented by Christian Huygens (1629-1695) in about 1679 and was modified much later into the form we now use by Augustin Fresnel (1788 - 1827). Huygens' construction is a method for locating the new position of a wave front. Starting from a known wavefront, we imagine each point on the wavefront to be a new source of secondary wavelets. The wavelet from each point then spreads out as a sphere (which appears as a circle in two-dimensional diagrams). After a certain time the new position of the original wavefront is defined by the boundary or envelope of all the secondary wavelets. Huygens construction for a plane wave going through a slit is shown in figure 5.4; after it has passed through the hole, the wavefront is no longer plane, but has curved edges. The result of Huygens' construction is significantly different from the ray model in that it shows light waves spreading into the region of the geometrical shadow. You should notice that opposite the slit the wavefront is still plane; it bends only at the edges. This bending effect is noticeable only on a scale comparable with the wavelength - for a very wide hole comparatively little of the wavefront bends around the edges.

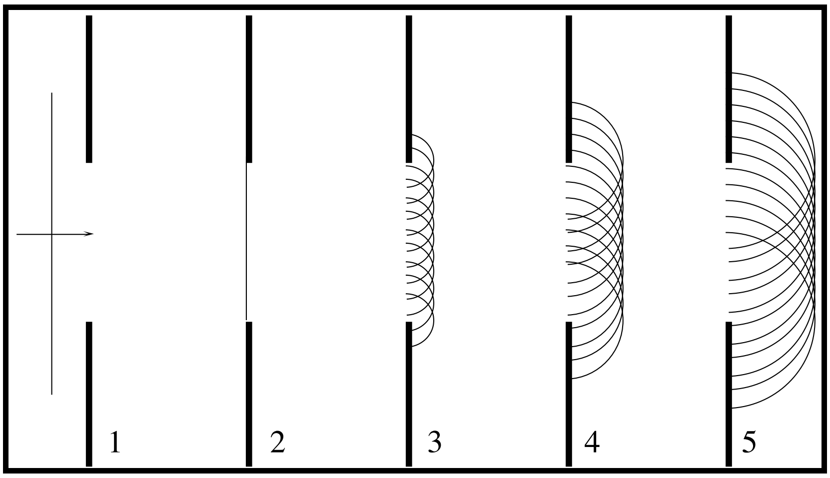


Figure 5.4. Huygens' construction

Five stages in the progress of a wavefront through a slit. The diagrams show the construction for three equal time intervals after a plane wave reaches the slit, travelling left to right. At any stage the new wavefront is the boundary of all the wavelets used in the construction.

Although the Huygens construction 'explains' how new wavefronts are formed, it does not predict the wave's amplitude; other methods are needed for that. The construction does however contain a clue about the strength of the waves. If you look at the straight-through wave in figure 5.4

you will see that there are many wavelets, but on the sides relatively few wavelets appear, which would seem to suggest, correctly, that the diffracted wave is weaker on the sides.

5-3 PRODUCING DIFFRACTION PATTERNS

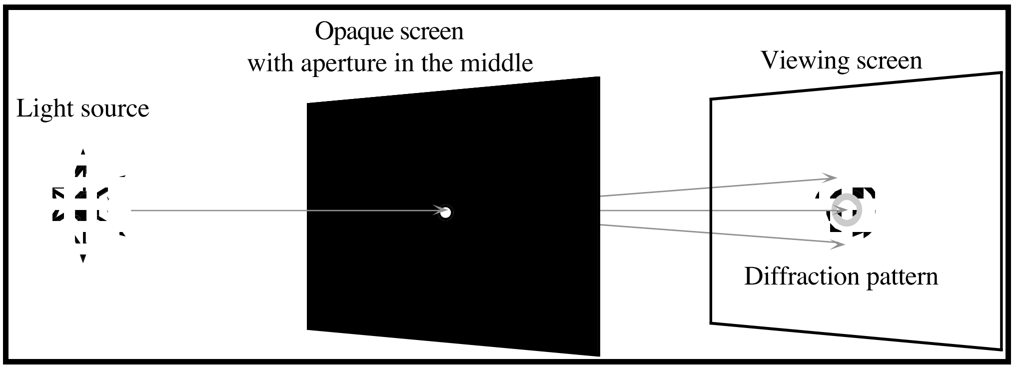


Figure 5.5. Producing a diffraction pattern (Not to scale)

Details of the distribution of light over the diffraction pattern depend on the distances of both the source and the screen from the diffracting aperture. The general situation (figure 5.5), known as **Fresnel diffraction**, can be mathematically very complex. However, calculations are greatly simplified if both the source and the screen are at very large distances from the aperture (i.e. if those distances are much greater than the diameter of the aperture). We will deal quantitatively only with this situation, which is known as **Fraunhofer diffraction.**

Fraunhofer conditions can be achieved in practice by using two lenses (figure 5.6). The first lens ensures that the wavefronts arriving at the aperture will be plane (with parallel rays) and the second lens brings beams of diffracted light together to form an interference pattern on the screen.

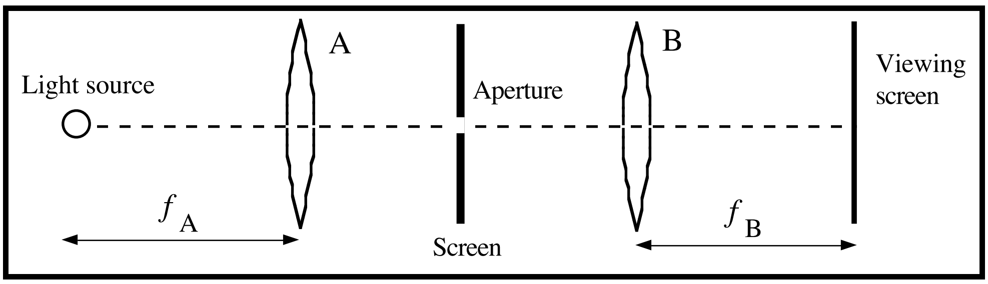


Figure 5.6. Producing Fraunhofer diffraction

Lens A ensures that the wavefronts which arrive at the aperture are plane. In order to see what the diffraction pattern at infinity would be like, lens B is used to produce its image on the screen.

5-4 DIFFRACTION AT A SINGLE SLIT

You will recall that light consists of a superposition of many elementary waves, with a wide range of frequencies. In the simple theory of diffraction we deal with one frequency at a time. The diffraction of a complex beam of light is then described in terms of what happens to each of the component waves with different frequencies.

The simplest case of Fraunhofer diffraction is that for a long narrow slit in an opaque screen. The interference pattern consists of a set of light and dark parallel fringes (figure 5.7).

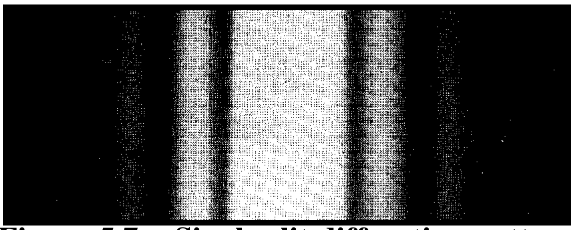


Figure 5.7. Single slit diffraction pattern

Figure 5.8 shows how the diffraction by a slit is studied using Fraunhofer conditions. The first lens ensures that the wavefronts arriving at the aperture will be plane and the second lens focusses the light onto the screen.

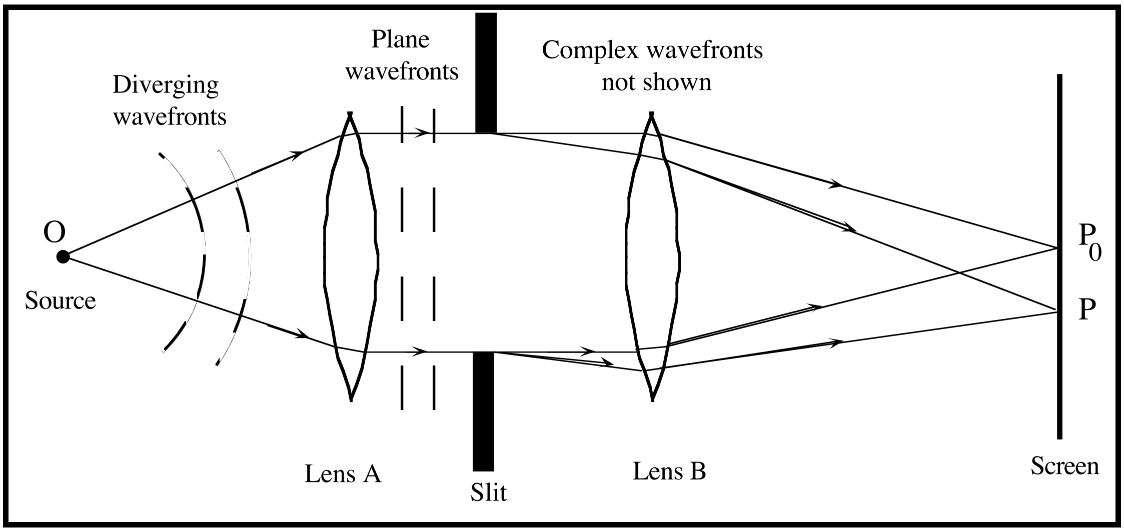


Figure 5.8. Fraunhofer diffraction by a single slit

After passing through the slit light waves spread out in all directions. The second lens brings parallel groups of rays to a focus on the screen. If there were no diffraction there would be only one such focus (P_0) for each source point O.

The wavefronts arriving at the slit are plane, but because of diffraction, the wavefronts on the other side will not be plane. Light arriving at any point in the aperture has a fixed phase relationship to light from the same part of the source arriving at any other point in the aperture. The simplest case is to imagine an elementary plane wavefront arriving parallel to the slit as shown in figure 5.8. Since all points on a wavefront have the same phase we can imagine the aperture filled with many tiny coherent sources. When the light from all of these sources comes together at various points on the screen an interference pattern will be seen. The brightness at any point will depend on the phase differences among all the secondary waves arriving there and those phase differences will depend on the optical paths travelled by the different waves. Path differences can be calculated using rays which leave the slit parallel to each other, so that they would eventually meet at infinity were it not for the presence of the second lens. Since the lens itself introduces no additional optical path difference, calculations can be done on the assumption that the rays meet at infinity.

The following argument, given in the video lecture, shows how to work out the condition for a minimum in the interference pattern. Consider rays coming from various coherent points spread across the slit (figure 5.9).

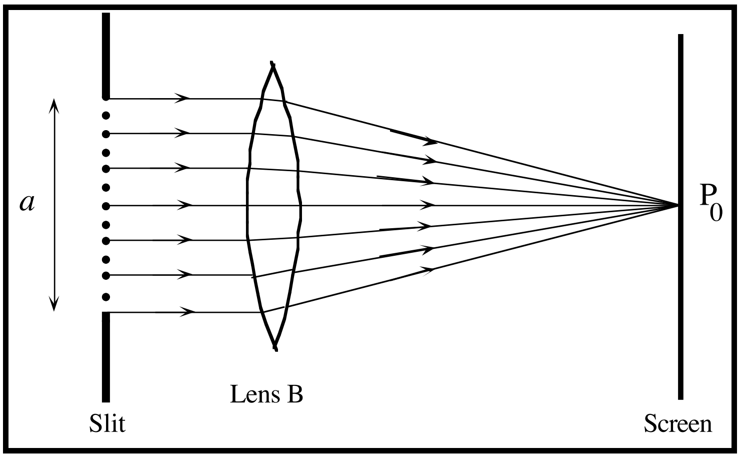


Figure 5.9. Formation of the central bright fringe
All these rays have the same optical path length.

All the rays parallel to the axis will be focussed at P_0. Although the geometrical path lengths of the rays are obviously different, the optical paths from different points across the slit to the point P_0 are all equal. That is so because the longer paths outside the lens are compensated by shorter optical paths of the rays within the lens. Rays from the outside of the slit go through more air but less glass than rays near the middle of the lens. So all the light arriving at P_0 is in phase, giving a bright region there.

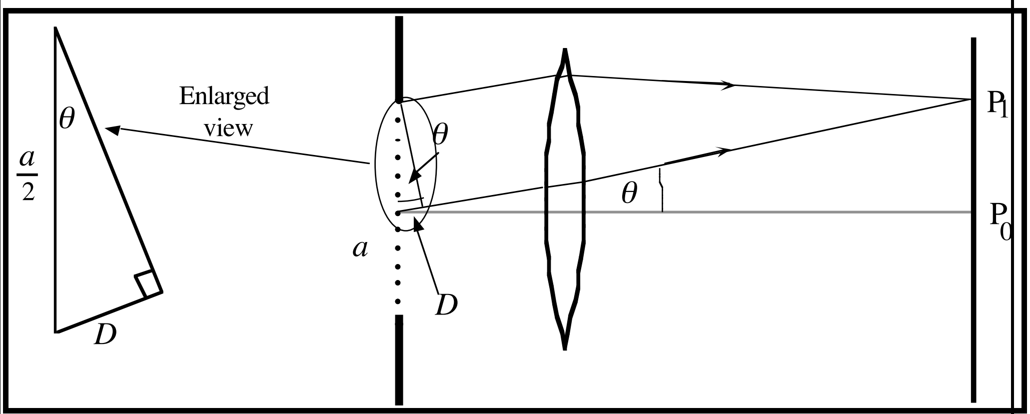


Figure 5.10. Condition for a minimum - Fraunhofer diffraction

For other points on the screen the phases for light arriving along a parallel group of rays are all different, but it is fairly straightforward to work out where there will be complete cancellation. Pick two parallel rays, one leaving from the top of the slit and the other from a point halfway across it (figure 5.10). Both rays leave the slit at an angle θ to the axis. If the angle θ is chosen so that the path difference between these two rays is some odd multiple of half a wavelength ($D = \lambda/2$, $3\lambda/2$,.. etc) they will interfere to produce a minimum in the irradiance at the screen.

Any other pair of rays half a slit width apart and leaving at angle θ will also interfere to give a minimum. We can choose similar pairs of rays until the whole aperture has been covered. Since all pairs produce a minimum there will be a dark region at P_1. From the small triangle in the diagram it can be seen that the path difference for every pair of rays is equal to $(a/2)\sin\theta$. Finally equate this expression with the value of D (above) needed for a minimum.

The condition for an interference minimum is that

$$\sin\theta = \frac{m\lambda}{a} \quad \text{for} \quad m = \pm1, \pm2, \pm3 \ \ldots \ . \qquad \ldots (5.1)$$

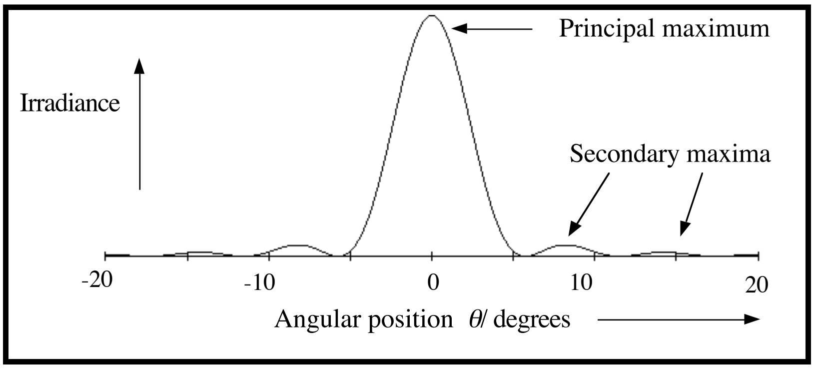


Figure 5.11. Single slit diffraction pattern - Fraunhofer conditions

A plot of the intensity of the light against position on the screen is shown in figure 5.11. Several features of the diffraction pattern are worth noting.

- The angle between the central peak and the first minimum ($m = 1$) is given by

$$\sin\theta = \frac{\lambda}{a} \ .$$

  So the widest diffraction patterns are produced by the narrowest slits. If the slit is only one wavelength wide the angular position of the first minimum would be 90° giving an angular separation of 180° between the zero intensity positions - so the diffracted light spreads out in all directions. On the other hand if the slit is wide, the angle of the first minimum is small and diffraction effects may be hard to see.
- The principal maximum is roughly twice the width of the secondary maxima. You can see that by putting $\sin\theta \approx \theta$ in equation 5.1 and comparing the change in angular position going from $m = 1$ to $m = -1$ with the change in going from $m = 1$ to $m = 2$, for example.
- The width of the diffraction pattern depends on wavelength. The pattern for red light is wider than that for blue light.
- The intensities of the secondary maxima are very much less than the intensity of the principal maximum.

5-5 DIFFRACTION BY A CIRCULAR APERTURE

In practice the most important example of diffraction produced by a single aperture is that for a circular hole. Most optical instruments have circular apertures or lenses that act as circular apertures, so whether we wish it or not we will get diffraction effects.

For Fraunhofer diffraction at a circular hole, the plot of intensity against position on the screen (figure 5.12) is similar in its general shape to that for a single rectangular slit. However it differs from the single slit pattern in the following ways.

- The diffraction pattern is a circular patch of light (called the Airy disc) surrounded by rings of light.
- The angle between the principal maximum and the first minimum is about 20% greater than that for a slit of the same width:

$$\sin\theta = 1.22\frac{\lambda}{a} \ . \qquad \ldots (5.2)$$

- The spacings between adjacent minima are not as uniform as those for the slit.
- The intensities of the secondary maxima are smaller.

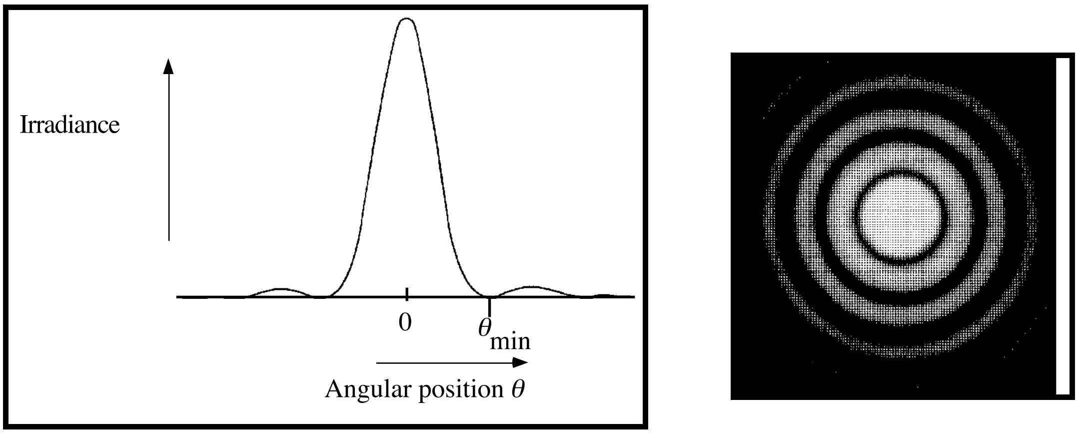


Figure 5.12. Diffraction pattern for a circular aperture
At a large distance from the hole (Fraunhofer conditions) the first minimum in the irradiance occurs at an angular position $\sin\theta_{min} = 1.22\ \lambda/D$. The secondary maxima are less bright than those for a slit.

Example

It is important to realise the size of the diffraction pattern for an average sort of lens, say 50 mm in diameter, which acts as a circular aperture whenever it is used to form images. Consider a typical visible wavelength of 500 nm and calculate the angle between the central maximum and the first minimum.

$$\sin\theta = 1.2\frac{\lambda}{D}$$

$$= \frac{1.2 \times 500 \times 10^{-9}\ \text{m}}{50 \times 10^{-3}\ \text{m}} = 1.2 \times 10^{-5}.$$

Since θ is small, $\sin\theta \approx \theta$ and $\theta = 1.2 \times 10^{-5}$ rad or 0.0007°.

5-6 YOUNG'S DOUBLE SLIT

Young's classic interference experiment with the two slits has already been described (§§4-2, 4-4) and quoted as evidence for the diffraction of light (§5-2). Detailed descriptions of the two-slit interference pattern involve the width of a slit, a, and the separation between the slits, d (figure 5.13). Note that $d > a$.

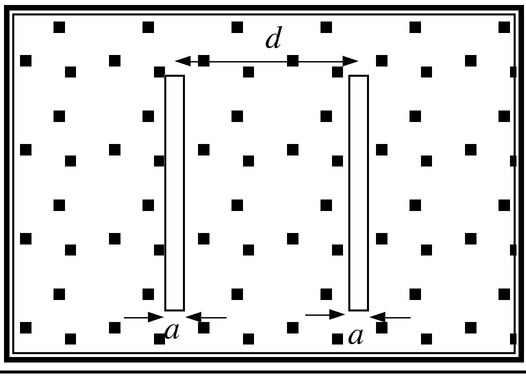


Figure 5.13

Dimensions of a double slit

A simple analysis of the experiment assumed that both the sources were very narrow - much less than a wavelength wide ($a << \lambda$). In practice however real slits are not that narrow and the interference pattern depends on the width of the slits as well as their separation. The kind of interference pattern produced is illustrated in figures 5.14 and 5.15.

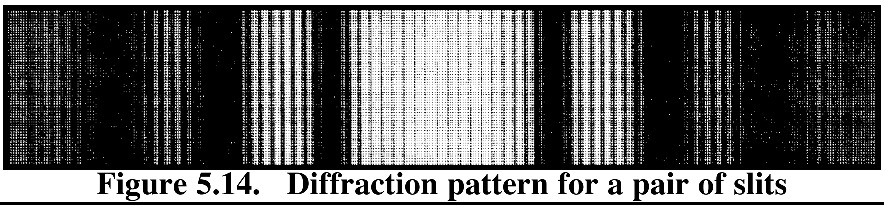


Figure 5.14. Diffraction pattern for a pair of slits

For Fraunhofer conditions the plot of irradiance against angular position on the screen shows a pattern of almost equally spaced fringes with different brightnesses (figure 5.15). The maximum irradiance occurs in the central bright fringe, at position $\theta = 0$. Notice how the fringes occur in groups, with the brightest fringes in the middle group.

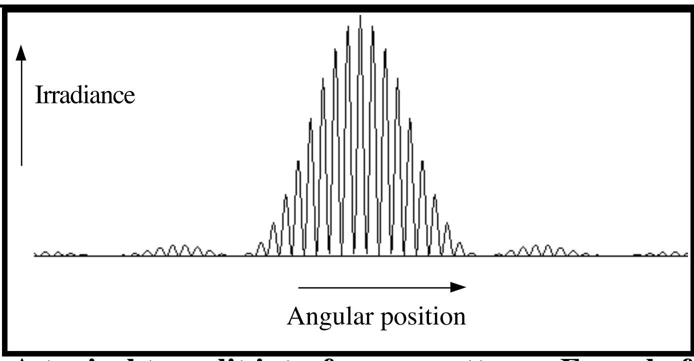


Figure 5.15. A typical two-slit interference pattern - Fraunhofer conditions

Figure 5.16 shows a comparison between the interference patterns for one and two slits. You may recall from chapter L1 that the maximum intensity for two identical coherent sources should be four times (not twice) the intensity for one source. That result posed a puzzle about conservation of energy. Energy or power is distributed differently in the two cases. Although the peak of the two-slit curve is four times higher than that of the single-slit curve, there are places where there is no light at all. The total power is represented by the areas under the graphs and, as expected from conservation of energy, the area under the two-slit curve is twice the area under the one-slit curve.

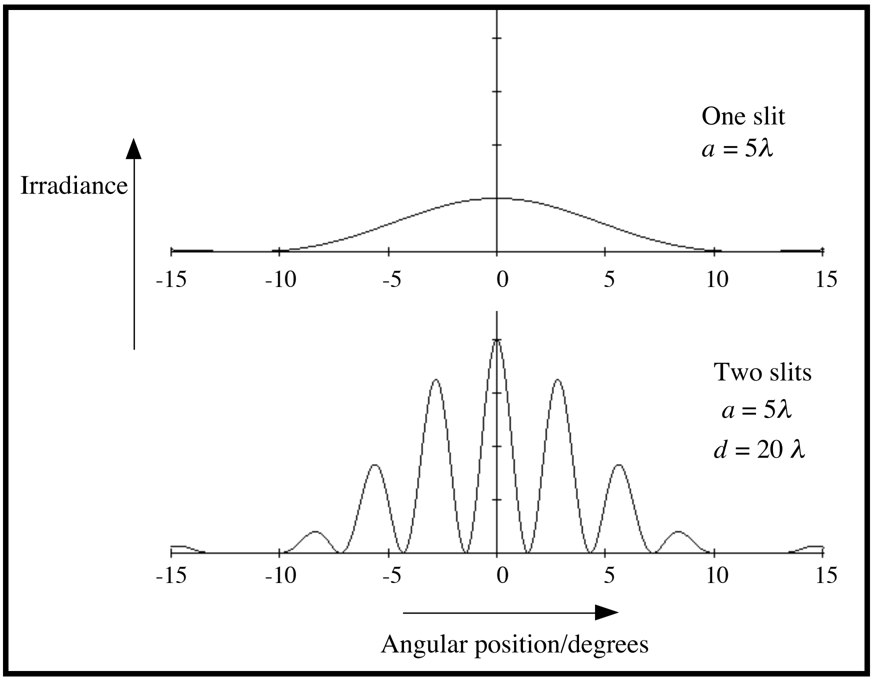


Figure 5.16. Irradiance and total power for one and two slits
Each slit is 5 wavelengths wide. These graphs are plotted on the same scale. Two identical slits give a maximum irradiance equal to 4 times the maximum irradiance from one slit. The total power, represented by the area under the curve, is only twice as large.

The interference pattern in Young's experiment can be described as a combination of the interference pattern for two very narrow slits and the pattern produced by one slit of finite width. When the mathematical forms of these two patterns are multiplied together we get the shape of the observed two-slit pattern. The spacing of the fringes is determined by the separation of the slits but their brightness is influenced by the width of a slit. Figure 5.17 shows how the interference pattern of two slits with zero width is modulated by the pattern of a single slit.

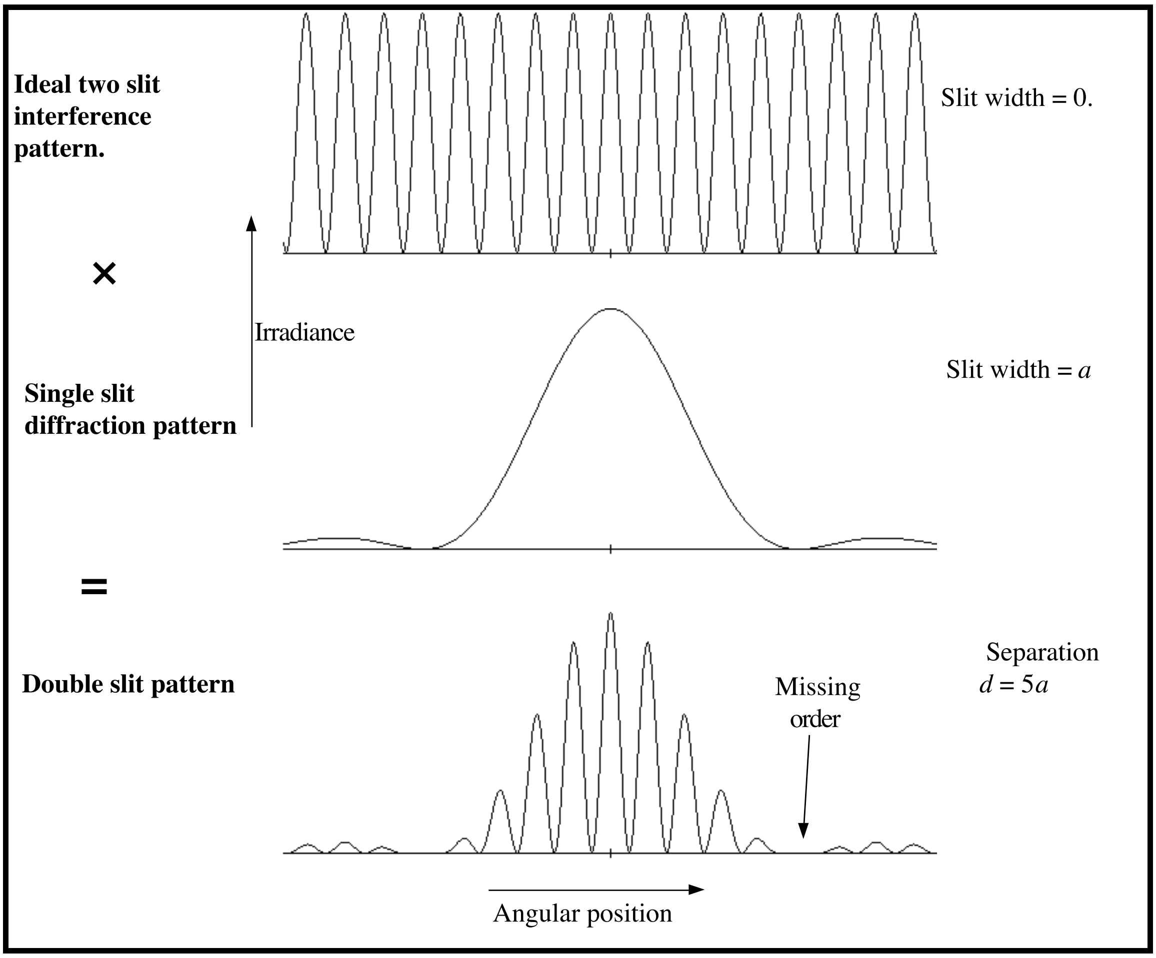


Figure 5.17. Young's slits pattern as a combination of two patterns
The ideal two-slit interference pattern (top graph) multiplied by the diffraction pattern for one slit (middle) yields the double slit diffraction pattern (bottom). The vertical scales in these graphs are not the same.

Since the description of the two-slit interference is just a combination of two simpler cases that we have already considered, it is to be expected that the results we got earlier should carry over. The locations of the interference maxima are still given by equation 4.3. For the

*m*th order maximum: $$\sin\theta = \frac{m\lambda}{d} \quad ; \quad m = 0, \pm 1, \pm 2, \ldots \; . \qquad \ldots (5.3)$$

And the places where there would be a zero in the single slit pattern are still dark; for a non-zero integer value of n , there is a minimum when

$$\sin\theta = \frac{n\lambda}{a} \quad ; \quad n = \pm 1, \pm 2, \ldots \; . \qquad \ldots (5.4)$$

If you look again at figure 5.17 you can see that in the place where the fifth order fringe ($m = 5$) ought to be there is nothing (remember to count the central fringe as number 0). That happened because the condition for the first zero ($n = 1$) in the single slit pattern coincided with the position of the fifth bright fringe in the two-slit pattern. By comparing the two conditions (equations 5.3 and 5.4) you can see that the missing fringe was caused by the fact that the slit spacing in the example was chosen to be exactly five times the slit width. Other missing orders may be caused by similar coincidences.

Figure 5.18 shows how the single slit pattern forms an **envelope** to the double slit pattern for some different values of the slit width but the same slit spacing. You could say that the two-slit interference pattern has to be squeezed inside the single slit pattern (figure 5.18).

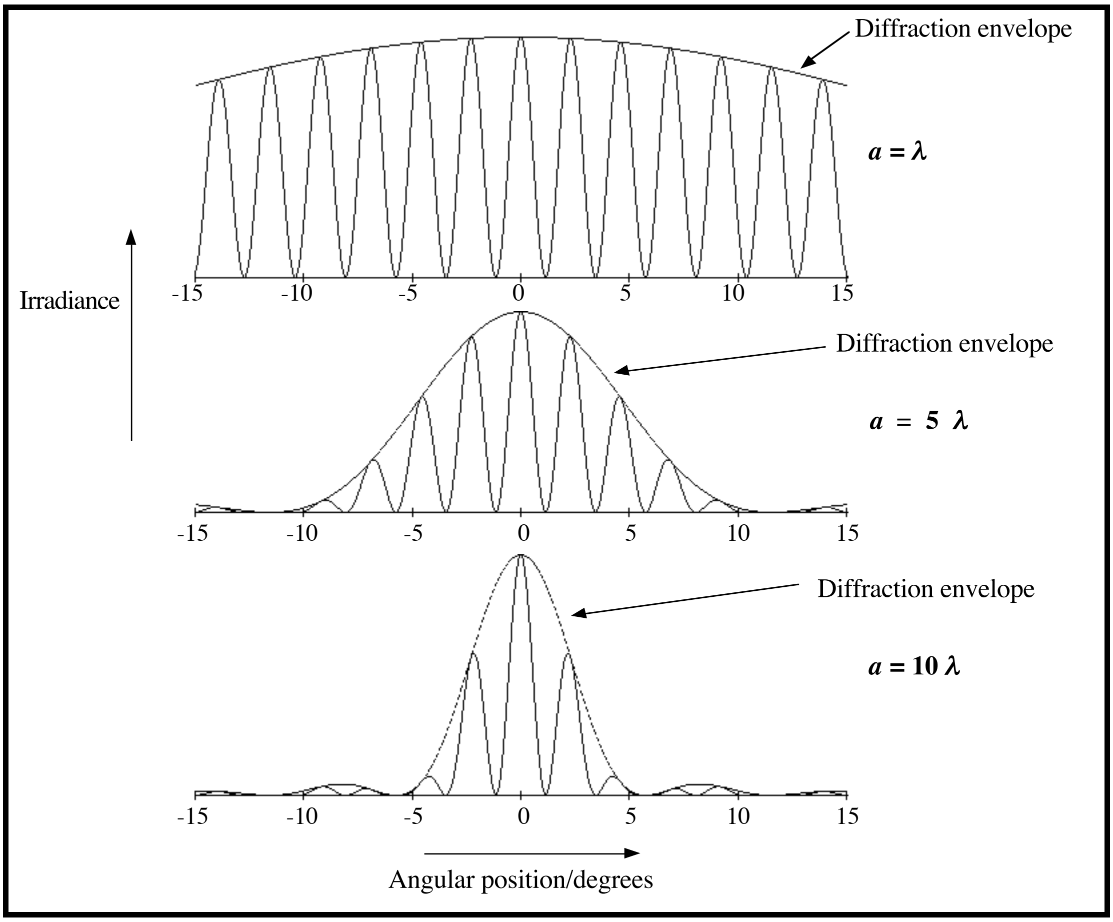


Figure 5.18 Effect of slit width on the fringe pattern
Although widening the slits lets more light through and increases the overall brightness, these graphs have been normalised to have the same maximum irradiance. Making both slits wider reduces the width of the diffraction envelope and the relative brightness of the outer fringes.

5-7 RESOLUTION OF IMAGES

The theory of diffraction shows that a point source of light or an object point cannot possibly produce a perfect point image even in an optical system which is free of aberrations. The wave nature of light places a fundamental limitation on the quality of an image. Resolving power is a somewhat loose term which refers to the ability of an optical system to distinguish fine detail in an image. It can be illustrated by considering how one might distinguish between the images of two stars in a telescope. Stars are so far away that they can be considered as point objects, and the fact that they often appear to have finite sizes is due to diffraction. Although a refracting telescope normally contains several lenses, the concept of resolving power can be understood by representing the system as one lens (the objective lens), and an aperture as shown in figure 5.19. The centres of the real images of the stars are formed in the focal plane of the objective lens and are separated by an angle θ which is equal to the angle subtended by the two stars at the objective. (The images are viewed using the telescope's eyepiece.)

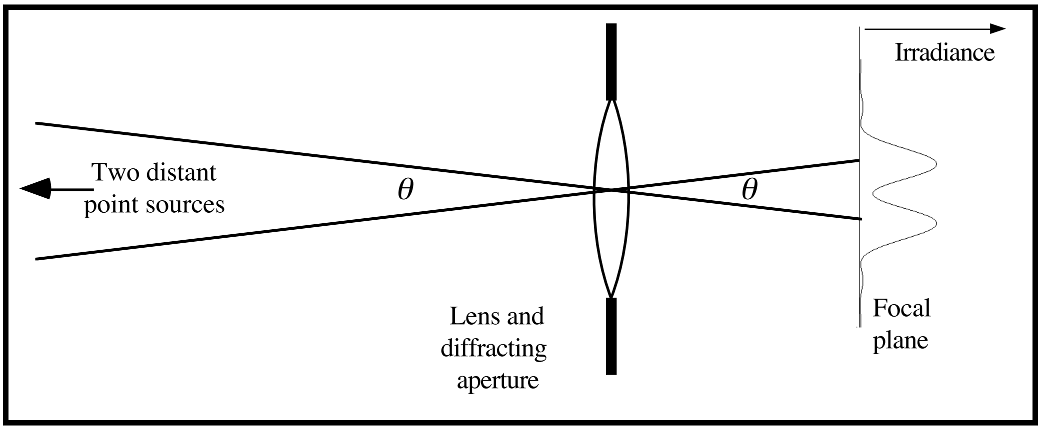


Figure 5.19. Resolving the images of two point sources
The sources are incoherent so the images are two separate diffraction patterns. (Note: the figure is not a ray tracing diagram.)

If the angular separation θ is so small that the images are effectively on top of each other, it will not be possible to recognise that there are two separate images. The **angular resolution** of an instrument is the smallest angular separation that can be distinguished. Although the decision about the actual value of an angular resolution may depend on the skill of the observer, the relative brightness of the sources and other factors, it is useful to have a generally agreed definition of when the two images of two incoherent point sources can be resolved. The criterion generally used was proposed by Lord Rayleigh (1821-1894): the two point sources are just resolved if the central maximum of one image coincides with the first minimum of the other.

For a circular aperture and Fraunhofer diffraction, you can see from equation 5.2 that the **Rayleigh criterion** is satisfied when the angular separation of the two point sources has the value θ_{min} given by

$$\sin\theta_{\text{min}} = \frac{1.22\lambda}{a}$$

where a is the diameter of the aperture. In practice, since the angles involved here are always very small we can put $\sin\theta = \theta$ and since 1.22 is near enough to 1, the Rayleigh criterion for a circular aperture is that

$$\text{angular resolution} = \frac{\lambda}{a} \quad .$$

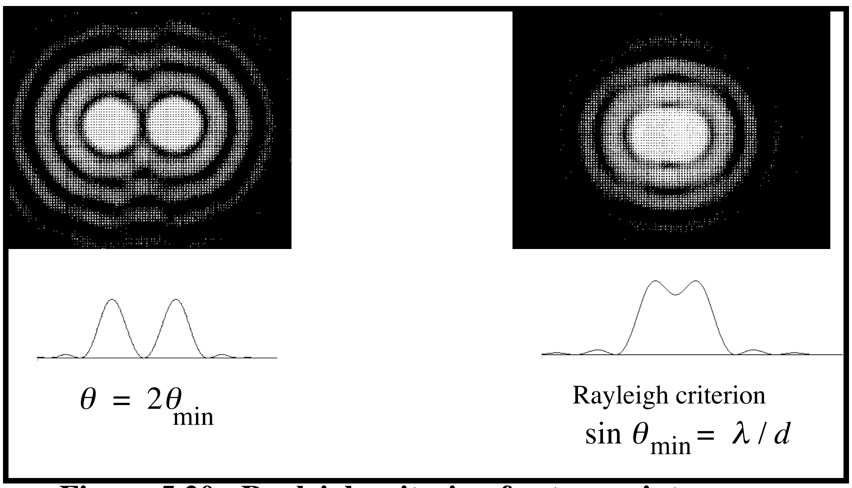


Figure 5.20. Rayleigh criterion for two point sources
The pair of images on the left is easily resolved. The two images on the right are just resolved according to Rayleigh's criterion.

It follows that to improve resolving power one can use a shorter wavelength, which is usually not possible, or a larger lens which gives a larger aperture. That is one of the reasons that modern astronomers need large telescopes.

The discussion above applies directly to telescopes and other optical instruments which are used to look at incoherent sources. However the details do not necessarily carry over to all kinds of microscopy, because the illumination of adjacent parts of a specimen may be at least partially coherent. The analysis of such cases is more difficult, but the general idea that resolution can be improved by using a larger aperture or a shorter wavelength remains valid.

5-8 DIFFRACTION GRATINGS

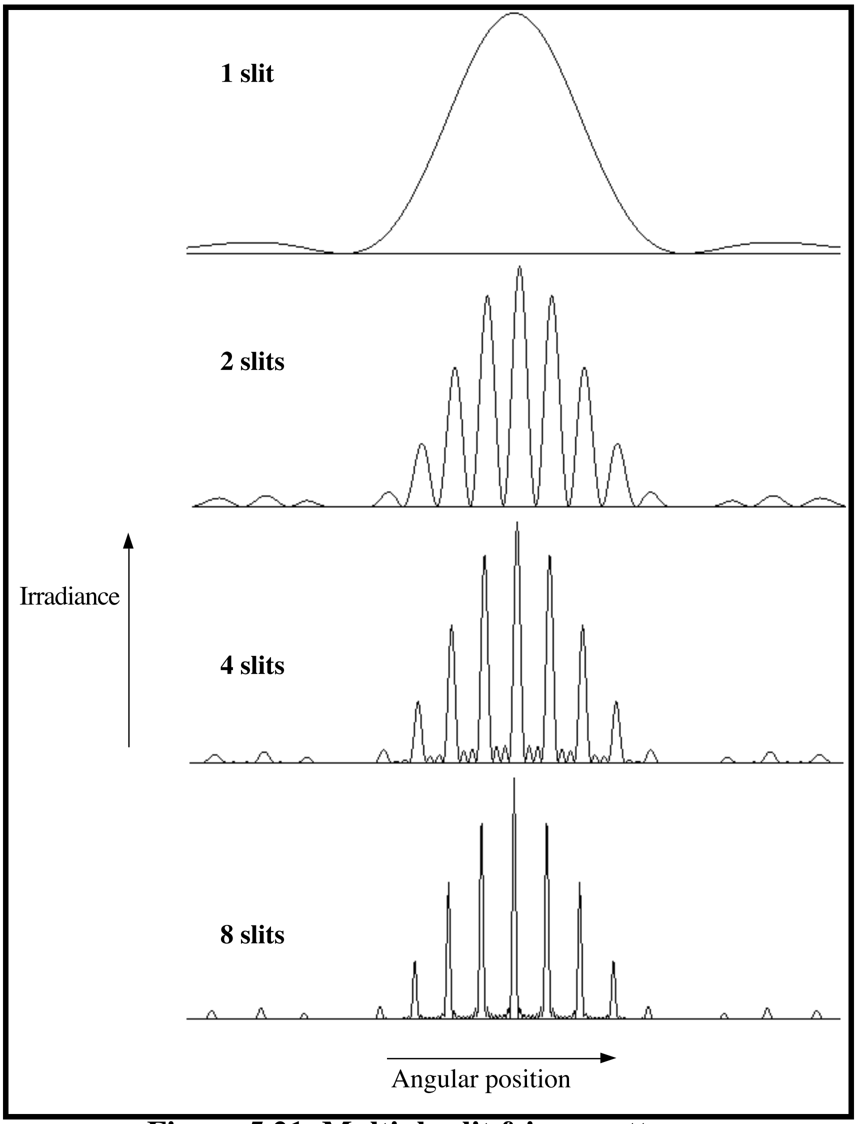


Figure 5.21. Multiple slit fringe patterns

These graphs have been normalised to have the same value of the central maximum in order to emphasise the effect on the fringe width. In fact using *N* slits instead of one slit makes the central maximum N^2 times brighter. In the examples above the separation of the slits is 5 times the slit width so that the fifth order fringe is missing.

Now that you have seen how a double slit pattern is formed, it is interesting to ask what happens if the number of equally spaced slits in a Young's experiment is increased. There are two main effects. Firstly, and fairly obviously, the whole diffraction pattern becomes brighter because more light gets through. The second effect is more surprising - the bright fringes get sharper! Although the positions of maximum irradiance remain unaltered the width of each bright fringe decreases as the number of slits increases. (A third, but less noticeable, effect is that some new weak fringes appear in between the principal maxima.)

The reason that the principal fringes stay in the same locations is that the conditions for maximum irradiance are still essentially the same - the optical path difference from adjacent slits to the point on the screen needs to be a whole number of wavelengths and that condition does not depend on how many pairs of slits you have. On the other hand, if there are many slits then the condition that all possible adjacent pairs of slits must give the same path difference is much tighter. Consider a point on the screen just a little bit away from a peak. The corresponding points on the wavefront at two adjacent slits do not have the exactly right phase difference for a maximum and the phases of the waves from other slits will be even more out of step. The more slits you have the worse the matching of the phases will be.

Although the fringes are much sharper their positions are still described by the same Young's slits equation (5.3).

The sharpening of the fringes is exploited in the **diffraction grating**, a device which consists of a very large number of narrow, uniformly spaced, parallel slits (typically 1200 slits per millimetre). There are two kinds of grating. A **transmission grating** is made by cutting grooves in a material such as glass; the grooves are effectively opaque strips and the unruled portions are the slits. A **reflection grating** works by reflecting light from many parallel mirror-like strips.

Since the fringe spacing depends on wavelength and because the bright fringes are very sharply defined, the diffraction grating can be used in a spectroscope or spectrograph to spread a beam of light into components with different frequencies. The grating is far superior to the prism because it can be made to give a much greater angular separation of the spectrum. If the slit separation is very small (say 10^{-6} m) and if there are many slits (say 10^6), then a line spectrum will have very sharp, clearly-separated interference maxima.

THINGS TO DO

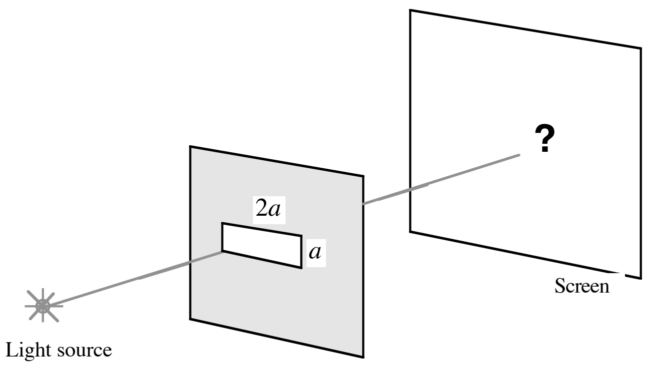
- The grooves on an LP record or (better) a compact disc can function as a reflection diffraction grating. Look at various sources of light reflected from the surface of a disc. How many orders of the diffraction pattern can you see? Can you estimate the spacing of the grooves?
- You can use a piece of finely woven cloth as a kind of two-dimensional transmission grating. Look through the cloth at a mercury or sodium street light.

QUESTIONS

Exercises

Q5.1 Estimate the width of the central maximum of a single slit diffraction pattern which appears on a screen 1.0 m away from a slit of width 0.10 mm, illuminated with light of $\lambda = 500$ nm.

Q5.2



Using the result that the angle between the centre and the first minimum of a single long slit diffraction pattern is $\sin\theta = \dfrac{\lambda}{a}$, can you predict what the diffraction pattern of the rectangular aperture in the figure will look like?

Q5.3 Estimate the angular width of the central maximum of the diffraction pattern produced by a circular aperture 2.0 mm in diameter. Take $\lambda = 500$ nm.

Q5.4 The two headlights of a distant approaching car are l m apart. Make a rough estimate of the distance at which the eye can resolve them. Take a pupil diameter of 5 mm and use a typical visible wavelength, say $\lambda = 500$ nm.

Q5.5 A grating with 400 grooves per millimetre is used to examine the spectrum of a source of light.

a) Calculate the angle between the first order maxima for red light with $\lambda = 700$ nm and violet light with $\lambda = 400$ nm.

b) Do the same for the two components of the sodium-D line which have wavelengths 589.0 nm and 589.6 nm.

Discussion questions

Q5.6 Can you explain why the centre of the diffraction pattern for a circular obstacle always contains a bright spot?

Q5.7 A grating will not be of much use for producing a clear spectrum if the first order fringes of some part of the spectrum overlap the second order fringes of some other part. Discuss. How should the grating be constructed in order to avoid the problem?

Q5.8 Why is diffraction more noticeable for sound than for light?

Q5.9 Light waves do not bend noticeably around buildings, but radio waves which are also electromagnetic waves do diffract around buildings. Discuss.

Q5.10 What do you think of the claim that you can't see an interference pattern in a Young's experiment if the separation of the slits is less than half a wavelength?

Q5.11 What would you see in a Young's two-slit experiment with white light?

Q5.12 Why is a grating better than just two slits in an experiment to measure wavelength?

APPENDIX

MATHEMATICAL DESCRIPTION OF THE DIFFRACTION PATTERNS

The shape of the Fraunhofer intensity distribution curve for a grating with N slits is described by the equation:

$$I = A\left(\frac{\sin\alpha}{\alpha}\right)^2\left(\frac{\sin(N\beta)}{\sin\beta}\right)^2$$

where $$\alpha = \frac{\pi a}{\lambda}\sin\theta$$

which is half the phase difference between one edge and the middle of a slit

and $$\beta = \frac{\pi d}{\lambda}\sin\theta$$

which is half the phase difference between corresponding points on adjacent slits. A is a constant which represents the amplitude of the incoming wave - assumed to be the same at all slits.

The single slit diffraction pattern ($N = 1$) and the Young's double slit case ($N = 2$) are also described by the same equation.

The envelope (single slit pattern) is described by the term $\left(\frac{\sin\alpha}{\alpha}\right)^2$ while the idealised pattern for N slits of zero width is given by the term $\left(\frac{\sin(N\beta)}{\sin\beta}\right)^2$.

The results for the zeros in the diffraction envelope (equation 5.1) and the condition for the maxima (equation 5.3) can be obtained from the equation above.

L6 POLARISATION

OBJECTIVES

Aims

Once you have studied this chapter you should understand the concepts of transverse waves, plane polarisation, circular polarisation and elliptical polarisation. You should be able to relate this understanding to a knowledge of methods for producing the different types of polarised light in sufficient detail so that you can explain the basic principles of those methods.

Minimum learning goals

1. Explain, interpret and use the terms:

   polarised light, unpolarised light, randomly polarised light, linear polarisation (plane polarisation), partially polarised light, polarising axis, polariser, ideal polariser, analyser, crossed polarisers, Malus's law, circular polarisation, elliptical polarisation, birefringence (double refraction), dichroic material, dichroism, optical activity, quarter-wave plate, polarising angle (Brewster angle).
2. Describe how plane polarised light can be produced by dichroic materials, by birefringent materials, by reflection and by scattering.
3. State and apply Malus's law.
4. Explain how circularly or elliptically polarised light can be regarded as a superposition of plane polarisations.
5. Describe how circularly polarised light can be produced from unpolarised or plane polarised light.
6. Describe the phenomenon of optical activity and describe one example of its application.

Extra Goals

7. Describe and discuss various applications of polarised light and explain how they work.

TEXT

6-1 PLANE OR LINEAR POLARISATION

In light and all other kinds of electromagnetic waves, the oscillating electric and magnetic fields are always directed at right angles to each other and to the direction of propagation of the wave. In other words the fields are transverse, and light is described as a **transverse wave**. (By contrast sound waves are said to be longitudinal, because the oscillations of the particles are parallel to the direction of propagation.) Since both the directions and the magnitudes of the electric and magnetic fields in a light wave are related in a fixed manner, it is sufficient to talk about only one of them, the usual choice being the electric field. Now although the electric field at any point in space must be perpendicular to the wave velocity, it can still have many different directions; it can point in any direction in the plane perpendicular to the wave's direction of travel.

Any beam of light can be thought of as a huge collection of elementary waves with a range of different frequencies. Each elementary wave has its own unique orientation of its electric field; it is **polarised** (figure 6.1). If the polarisations of all the elementary waves in a complex beam can be made to have the same orientation all the time then the light beam is also said to be polarised. Since there is then a unique plane containing all the electric field directions as well as the direction of the light ray, this kind of polarisation is also called **plane polarisation**. It is also known as **linear polarisation**. However, the usual situation is that the directions of the electric fields of the component wavelets are randomly distributed; in that case the resultant wave is said to be **randomly polarised** or **unpolarised**.

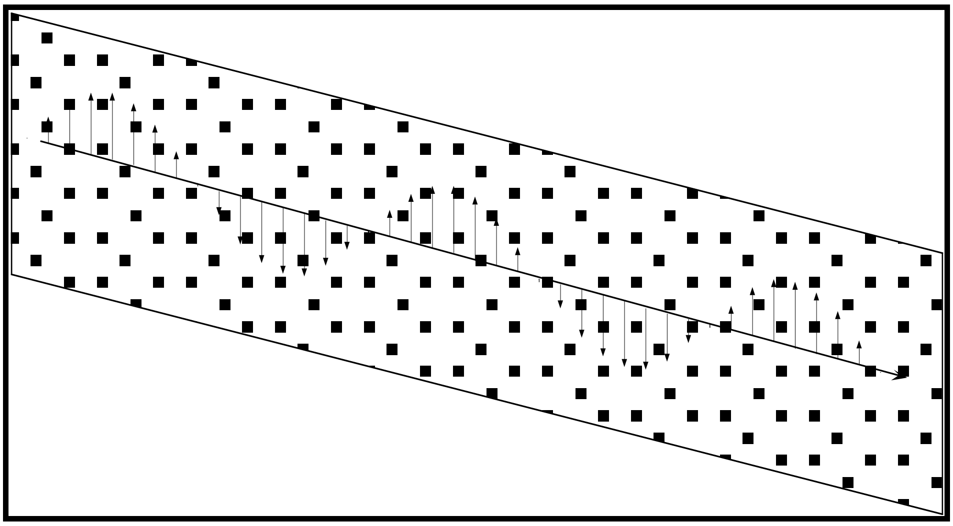


Figure 6.1. A polarised elementary wave

The picture shows a perspective plot of the instantaneous electric field vectors which all lie in the same plane (shaded). Every such elementary harmonic wave is plane polarised.

It is quite common to find **partially polarised light** which is a mixture of unpolarised (completely random polarisations) and plane polarised waves, in which a significant fraction of the elementary waves have their electric fields oriented the same way.

Components of polarisation

Since electric field is a vector quantity it can be described in terms of components referred to a set of coordinate directions. In the case of polarised waves we can take *any* two perpendicular directions in a plane perpendicular to the wave's direction of travel. An electric field $\boldsymbol{E}$ which makes an angle α with one of these directions can then be described completely as two components with values $E\cos\alpha$ and $E\sin\alpha$. We can think of these components as two independent electric fields, each with its own magnitude and direction, which are together equivalent in every respect to the original field. So any elementary wave can be regarded as a superposition of two elementary waves with perpendicular polarisations.

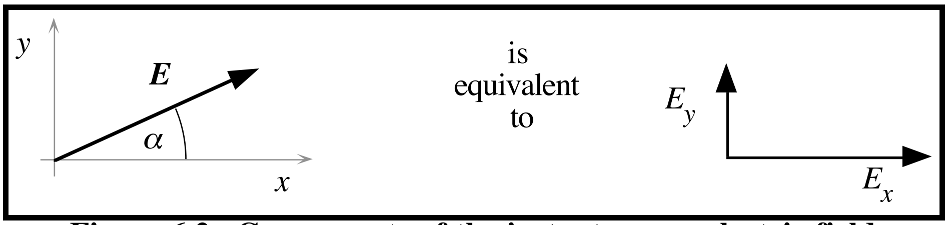


Figure 6.2. Components of the instantaneous electric field

In just the same way, any plane polarisation can be described in terms of two mutually perpendicular component polarisations. In the schematic diagrams that we use to represent polarisation such a line can be drawn as a double-headed arrow, representing the two opposite directions that a plane polarised wave can have at any point. The instantaneous value of an electric field (which has a unique direction at any instant of time) will be shown as a single-headed arrow.

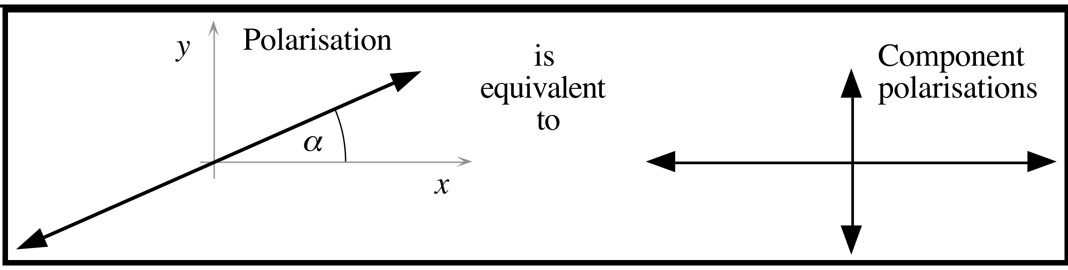


Figure 6.3 Components of the polarisation

Polarisers and Malus's law

A **ideal polariser**, or polarising filter, turns unpolarised light into completely plane polarised light. Its action can be described in terms of its effect on elementary waves with different polarisations; waves whose polarisation is parallel to an axis in the polariser, called its **polarising axis**, are transmitted without any absorption but waves whose polarisation is perpendicular to the polarising axis are completely absorbed. An elementary wave whose polarisation is at some other angle to the polarising axis is partly transmitted and partly absorbed but it emerges from the other side of the polariser with a new polarisation, which is parallel to the polariser's axis. This can be described in terms of components of the original wave. Since we can use any reference directions for taking components we choose one direction parallel to the polariser's axis and the other one perpendicular to it. If the angle between the original polarisation and the polariser's axis is θ, then the component parallel the the polariser's axis, which gets through, has an amplitude of $E_0 \cos\theta$. Since the other component is absorbed, the wave which emerges has a new amplitude $E_0 \cos\theta$ and a new polarisation. Since the irradiance or "intensity" of light is proportional to the square of the electric field's amplitude,

$$I_{out} = I_{in} \cos^2\theta. \qquad ...(6.1)$$

This result is known as **Malus's law**.

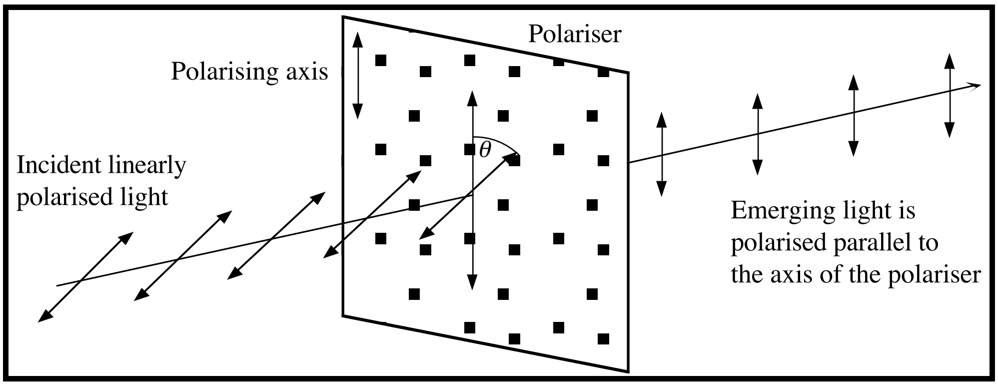


Figure 6.4. Effect of a polariser on plane polarised light

Many practical polarisers do not obey Malus's law exactly, firstly because they absorb some of the component with polarisation parallel to the polarising axis and secondly because some of the component polarised perpendicular to the axis is not completely absorbed.

Malus's law also describes the action of an ideal polariser on unpolarised light. Unpolarised light is really a vast collection of polarised elementary waves whose polarisations are randomly spread over all directions perpendicular to the wave velocity. Since these elementary waves are not coherent, their intensities, rather than their amplitudes, can be added, so Malus's law works for each elementary wave. To work out the effect of the polariser on the whole beam of unpolarised light we take the average value of $I_{in} \cos^2\theta$ over all possible angles, which gives

$$I_{out} = \frac{1}{2} I_{in} .$$

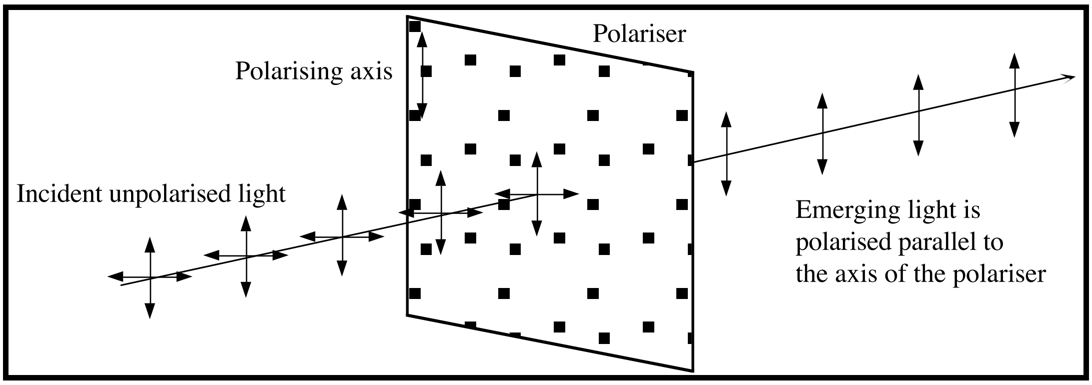


Figure 6.5. Effect of an ideal polariser on unpolarised light

If we send initially unpolarised light through two successive polarisers, the irradiance (intensity) of the light which comes out depends on the angle between the axes of the two polarisers. If one polariser is kept fixed and the axis of the other is rotated, the irradiance of the transmitted light will vary. Maximum transmission occurs when the two polarising axes are parallel. When the polarising axes are at right angles to each other the polarisers are said to be **crossed** and the transmitted intensity is a minimum. A pair of crossed ideal polarisers will completely absorb any light which is directed through them (figure 6.6). Note that the polarisation of the light which comes out is always parallel to the polarising axis of the last polariser.

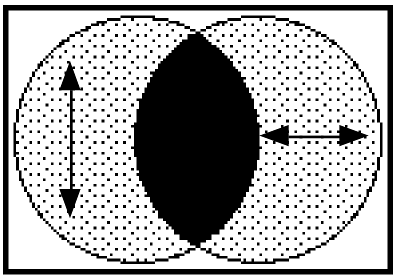


Figure 6.6. Crossed polarisers

Each polariser on its own transmits half the incident irradiance of the unpolarised light.

So far we have considered a polariser as something which produces polarised light. It can also be considered as a device for detecting polarised light. When it is used that way it may be called an **analyser**. For example, in the case of crossed polarising filters above, you can think of the first filter as the polariser, which makes the polarised light, and the second filter as the analyser which reveals the existence of the polarised light as it is rotated.

6-2 CIRCULAR POLARISATION

Plane polarisation is not the only way that a transverse wave can be polarised. In **circular polarisation** the electric field vector at a point in space rotates in the plane perpendicular to the direction of propagation, instead of oscillating in a fixed orientation, and the magnitude of the electric field vector remains constant.

Looking into the oncoming wave the electric field vector can rotate in one of two ways. If it rotates clockwise the wave is said to be right-circularly polarised and if it rotates anticlockwise the light is left-circularly polarised.

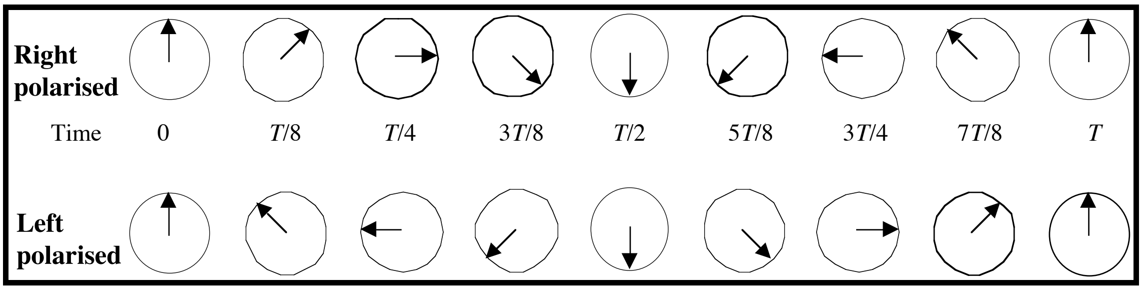


Figure 6.7. Circularly polarised waves

The diagrams show the electric field vector of an elementary wave at successive time intervals of 1/8 of a wave period, as the wave comes towards you.

Actually circular polarisation is not anything new. A circularly polarised elementary wave can be described as the superposition of two plane polarised waves with the same amplitude which are out of phase by a quarter of a cycle ($\pi/2$) or three quarters of a cycle ($3\pi/2$). Figure 6.8 shows how.

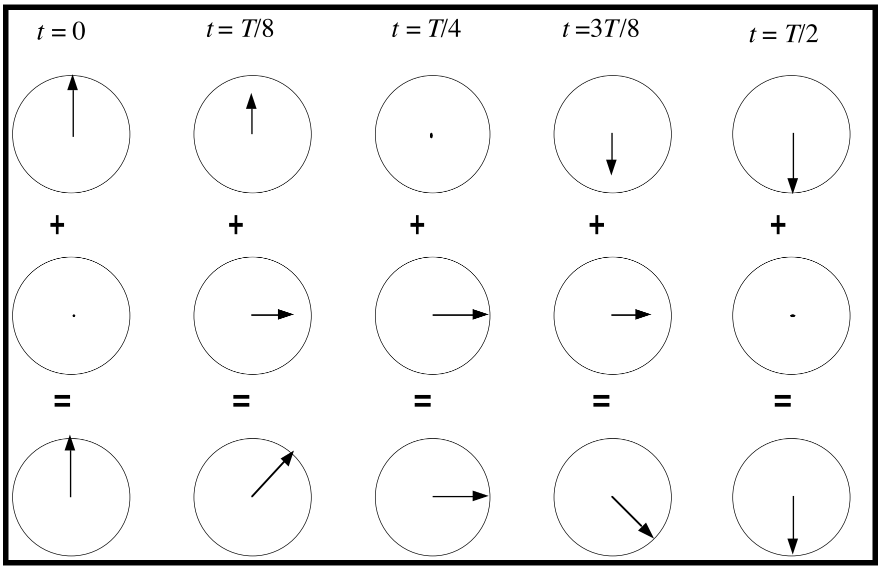


Figure 6.8 Circular polarisation as the superposition of two linear polarisations

The illustrations show the two linearly polarised electric fields with the same amplitude plotted at intervals of one eighth of a wave period. When these are combined the resultant electric field vector always has the same magnitude, but its direction rotates. Note that the amplitude of the circularly polarised wave is equal to the amplitude of each of its linearly polarised components. Its period and frequency are also identical with those of the component waves.

There is an interesting symmetry between the concepts of linear and circular polarisation. Not only can circular polarisation be described in terms of linear polarisation, but linear polarisation can be described as the superposition of two circular polarisations! In figure 6.9, left and right circularly polarised waves with equal amplitudes are added to produce one linearly polarised wave. Note that in this case the amplitude of the linearly polarised wave is the sum of the component amplitudes.

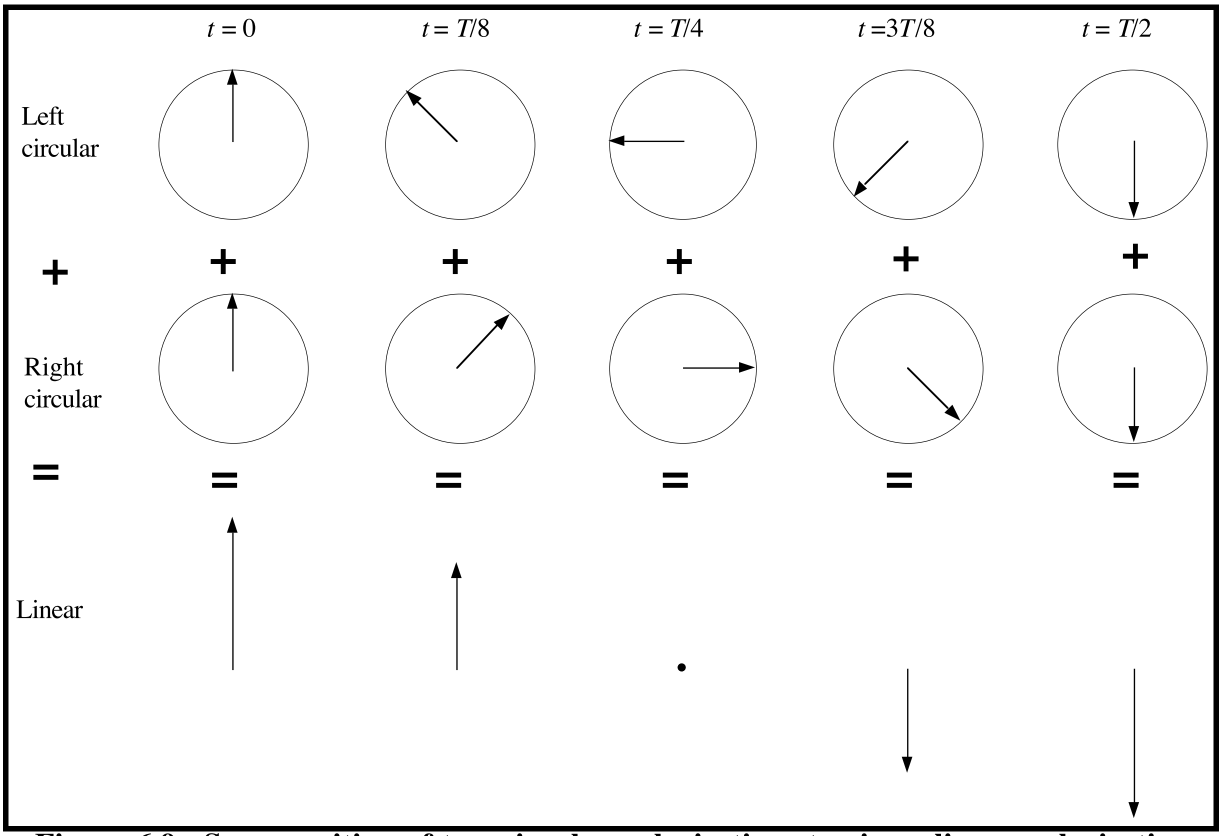


Figure 6.9. Superposition of two circular polarisations to give a linear polarisation

Elliptical polarisation

Circular polarisation can be regarded as a superposition of two linear polarisations with the same amplitude and just the right phase difference, $\pi/2$, $3\pi/2$ etc. In general the combination of two linearly polarised elementary waves with the same frequency but having unequal amplitudes and an arbitrary value of the phase difference, produces a resultant wave whose electric vector both rotates and changes its magnitude. The tip of the electric field vector traces out an ellipse so the result is called **elliptical polarisation** (figure 6.10). Circular polarisation is thus a special case of elliptical polarisation.

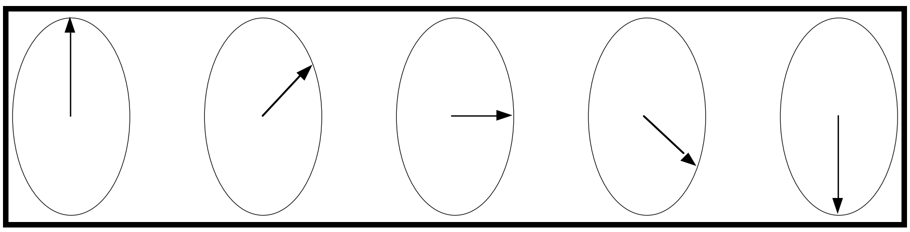


Figure 6.10. Elliptical polarisation

In this example two component waves have a phase difference of a quarter cycle and different amplitudes. The total electric field vector changes size as it rotates.

We have already seen that the resultant of two linear polarisations with zero phase difference is also a linear polarisation. Another special case is the combination of two elementary linearly polarised waves whose phase difference is exactly π. The resultant is a linear polarisation but its orientation is perpendicular to the linear polarisation when the component waves have no phase difference.

6-3 PRODUCTION OF POLARISED LIGHT

When an elementary light wave interacts with matter, its electric field causes electrons within the substance to vibrate at the wave's frequency. These vibrating electrons then re-radiate the absorbed energy as new electromagnetic waves in all directions. Although this **scattered light** has the same frequency as the incident wave its polarisation depends on the new direction of propagation.

In general, therefore, when light interacts with matter its polarisation may be changed. The main mechanisms by which this happens are :

1. by passing through **dichroic materials**;
2. by passing through **birefringent materials**;
3. by **scattering;**
4. by **reflection**;
5. by passing through **optically active materials**.

6-4 DICHROIC MATERIALS

In some crystalline materials, which are described as **dichroic**, the absorption of light depends on the orientation of its polarisation relative to the polarising axis of the crystal. Light whose plane of polarisation is perpendicular to the polarising axis is absorbed more than that which is parallel to it. The most common example is a group of materials sold under the trade name Polaroid which are used, for example, in sunglasses and photographic filters. One variety of Polaroid contains long molecules of the polymer polyvinyl alcohol (PVA) that have been aligned and stained with iodine. The best known example of a crystalline dichroic material is the mineral tourmaline.

Polarisers made from dichroic materials differ from an ideal polariser in the following ways. Firstly, if the polariser is thin, the emerging light may not be completely plane polarised. Secondly there is some absorption of the transmitted polarisation component. Thirdly the amount of absorption usually depends on the frequency of the light, so that the light which comes out may appear to be coloured.

6-5 BIREFRINGENCE

In some materials light with different polarisations travels at different speeds. Since we can regard any wave as the superposition of two plane polarised waves, this is equivalent to saying that one beam of light travels at different speeds in the material, that is the material has different refractive indices for light of the same frequency. Such materials are said to be **doubly refracting** or **birefringent**. Examples are crystals such as the minerals calcite (calcium carbonate) and quartz (silicon dioxide) or materials like Cellophane when it is placed under stress.

The speed of light in a birefringent crystal depends, not only on the polarisation, but also on the direction of travel of the light. As usual we can regard any beam of light as a superposition of two linearly polarised components at right angles to each other. By choosing suitable directions for the polarisation components it is found that one component wave, called the **ordinary wave**, travels at the same speed in all directions through the crystal, but the speed of the other polarisation component, called the **extraordinary wave**, depends on its direction of travel. There are some propagation directions in which all polarisations of light travel at the same speed and a line within the crystal parallel to one of those directions is called an **optic axis**. Some crystals, called **uniaxial crystals** have only one optic axis, while others, the **biaxial crystals**, have two.

Figure 6.11. shows what would happen to light starting out from some point inside a calcite crystal. (This is not as silly as it may seem; Huygens' construction regards each point on a wavefront

as a source of new waves. So the 'point source' considered here could be a point on a wavefront which originated outside the crystal.)

Calcite has one optic axis, along which the ordinary and extraordinary waves travel at the same speed, and the plane of the diagram has been chosen to include that axis. Two wavefronts are shown. Since the ordinary wave travels at the same speed (v_o) in all directions its wavefronts (for light coming from a point source) are spherical, and the section of the wavefront in the diagram is therefore circular. On the other hand, the speed (v_e) of the extraordinary wave depends on the direction of travel and the section of the wave front shown is elliptical. In calcite the speed of the extraordinary wave is always greater than or equal to the speed of the ordinary wave, so the extraordinary wavefront encloses the ordinary wavefront. In a crystal where the extraordinary wave is the slower of the two, its wavefront would stay inside the spherical wavefront of the ordinary wave. In figure 6.12 the polarisations are shown. The ordinary wave is polarised perpendicular to the plane of the diagram and the polarisation of the extraordinary wave is parallel to the plane of the diagram.

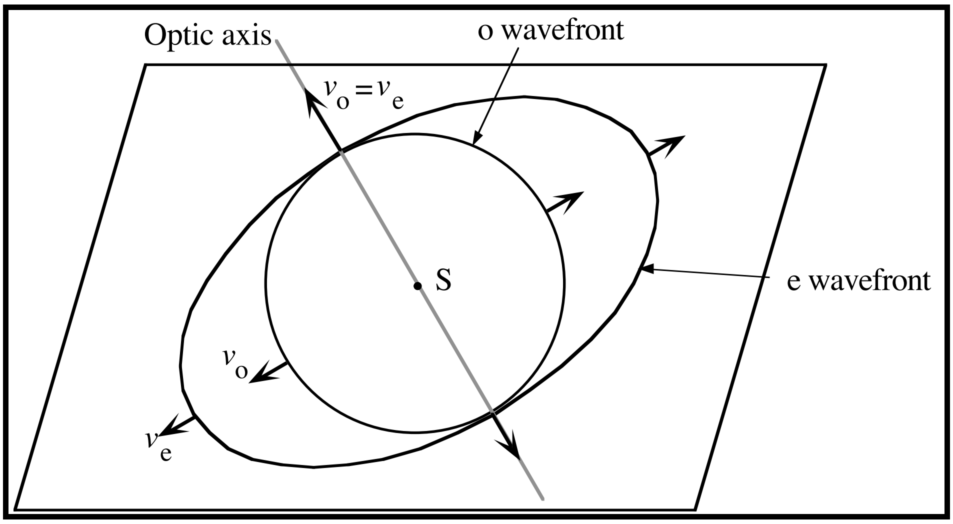


Figure 6.11. Ordinary and extraordinary waves

In this diagram the uniaxial crystal has been sliced so that the section contains the optic axis, which is defined as the orientation in which the e and o waves travel at the same speed. The speed of the extraordinary wave depends on direction. The diagram shows wavefronts for e and o waves which started from the point S at the same time.

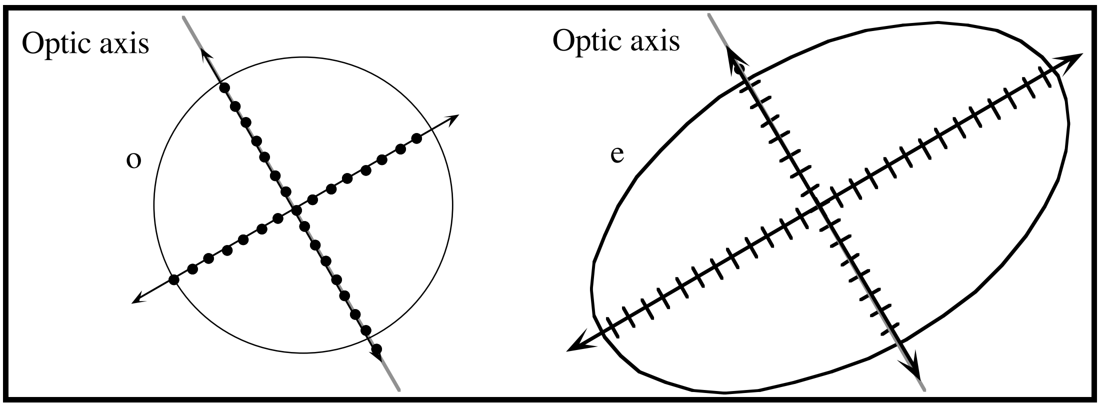


Figure 6.12. Polarisation of e and o waves

The wavefronts from figure 6.11 have been drawn separately. The ordinary wave is polarised perpendicular to the plane containing the optic axis. Note that in this and other diagrams, polarisations perpendicular to the page are shown as dots while polarisations in the plane of the page are represented by short lines.

Birefringence and circular polarisation

In any direction other than along an optic axis the wave speed depends on the direction of travel and the polarisation. In the following discussion we consider only light travelling in a plane perpendicular to the optic axis (figure 6.13). In this case the polarisation of the ordinary wave is perpendicular to the optic axis. The speed of the ordinary wave does not depend on direction. The component with polarisation parallel to the optic axis is an extraordinary wave. It can be faster or slower than the ordinary wave.

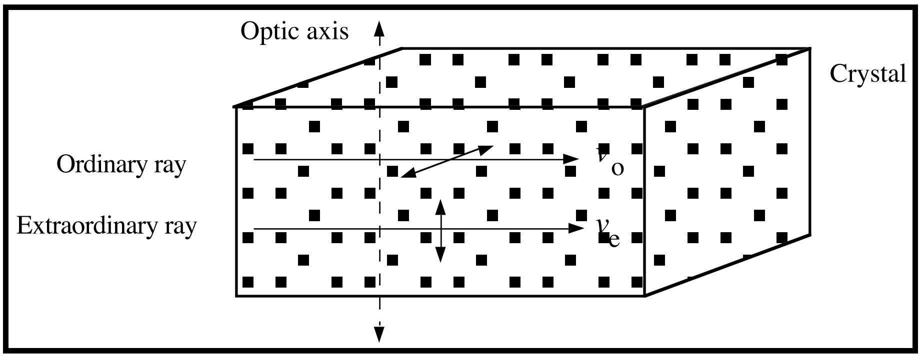


Figure 6.13. Ordinary and extraordinary rays in a uniaxial crystal
The faces of this crystal are not natural; they have been cut so that one pair of opposite faces is perpendicular to the optic axis while the other faces are parallel to the optic axis.

Birefringence can be exploited to produce circular polarisation. This can be achieved by letting a beam of plane polarised monochromatic light strike a specially prepared slab of birefringent material with faces cut like the crystal shown in figure 6.13. Light enters the crystal normal to a surface which contains the optic axis, with its polarisation at 45° to the optic axis. In order to analyse what happens, the electric field of the incident light can be resolved into ordinary (o) and extraordinary (e) components, perpendicular and parallel to the optic axis (figure 6.14). Since the angle of incidence is 90° both ordinary and extraordinary waves travel inside the crystal in the same direction (there is no refraction), but with different speeds.

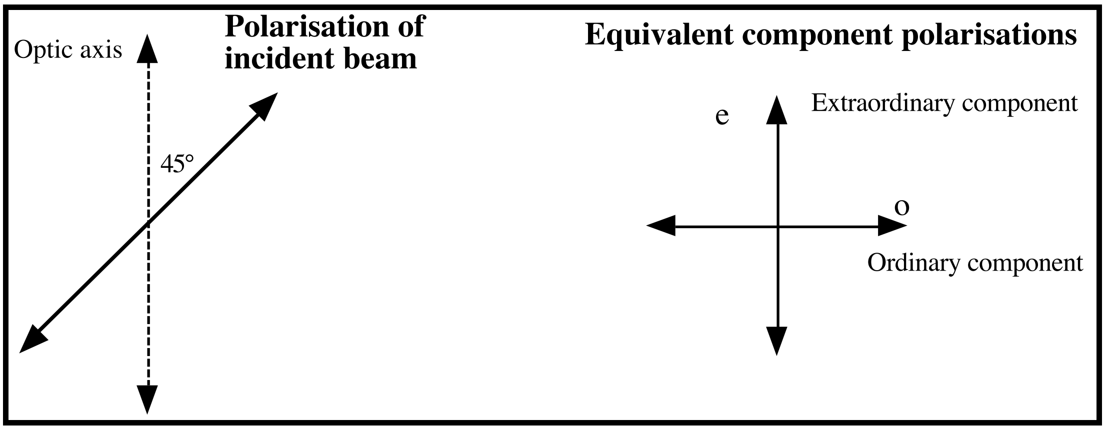


Figure 6.14 Resolving the plane polarisation into e and o components
A plane polarised wave enters a crystal with its polarisation at 45° to the crystal's optic axis. The plane polarisation can be regarded two perpendicular plane polarisations with equal amplitudes. Each component is at 45° to the original polarisation. What has happened to the light by the time it comes out the other side of the crystal depends on the thickness of the crystal and is shown in figure 6.15.

Since the ordinary and extraordinary waves travel at different speeds through the crystal, their phase difference and the polarisation of the emerging light will depend on the thickness of the crystal. If the extraordinary wave is faster, it will progressively move ahead of the ordinary wave.

To find the polarisation of the wave that comes out at the second boundary we can add the extraordinary and ordinary components together again. Depending on the thickness of the crystal, any of the following can happen.

- If the extraordinary wave has gained one complete wavelength (figure 6.15d) the phase difference between the ordinary and extraordinary components will be effectively the same as it was originally, so the emerging wave is linearly polarised with the same plane of polarisation as the incident wave.
- If the extraordinary wave has gained exactly half a wavelength (figure 6.15b) the two components will be out of phase by π. This phase relation is maintained at all times. The resultant wave is linearly polarised with its polarisation perpendicular to that of the original wave, i.e. the plane of polarisation has been rotated through an angle of 90°. A slab of birefringent material which produces this effect is called a **half wave plate**.
- If the extraordinary wave has gained a quarter wavelength (figure 6.15a) there is a phase difference of $\pi/2$ between the e and o waves so the light becomes circularly polarised. (Have another look at figure 6.8.)

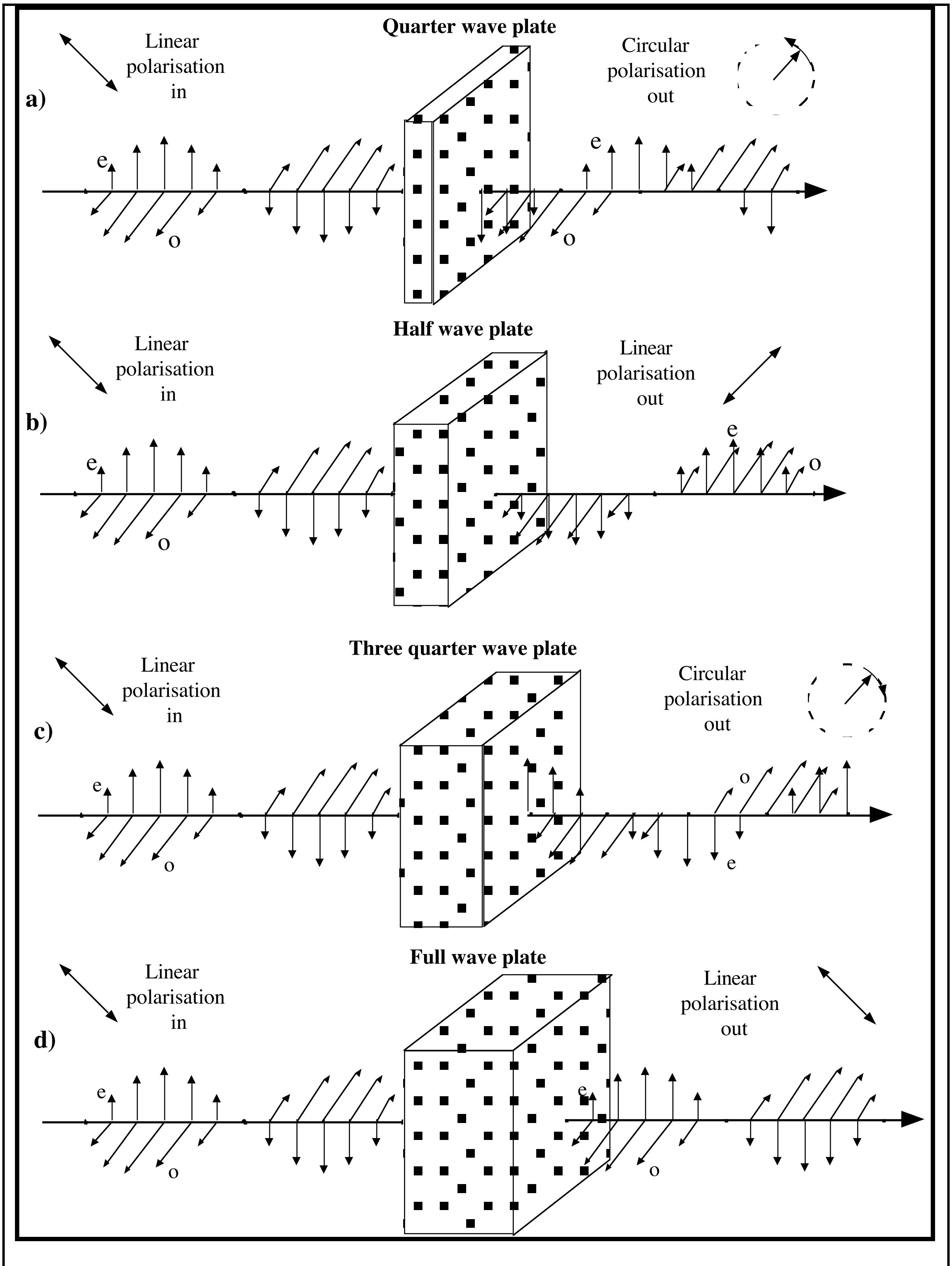


Figure 6.15 Action of wave plates

- If the extraordinary wave has gained three quarters of a wavelength (figure 6.15c) the phase difference is $3\pi/2$ and the light is circularly polarised with the resultant electric field rotating the other way.

One can use slabs of birefringent material where the extraordinary wave gains $\frac{1}{4}$ or $\frac{3}{4}$ wavelength to produce circularly polarised light from linearly polarised light or vice versa. Such slabs are called **quarter wave plates**. Note that to get circularly polarised light, the incident light must be polarised at 45° to the optic axis; other angles will give unequal e and o components so the light which comes out will be elliptically polarised.

Double images

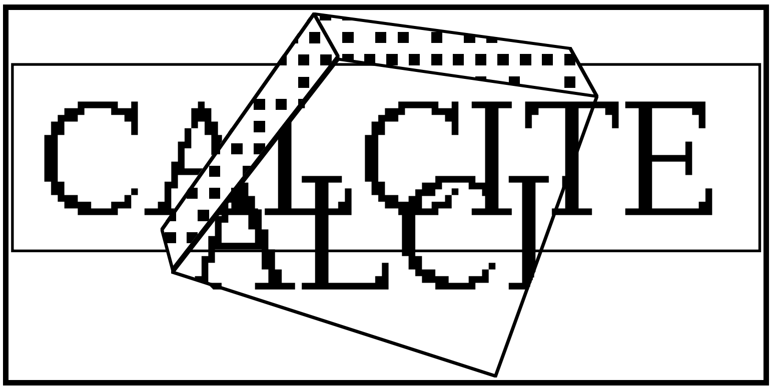


Figure 6.16. Double image in a calcite crystal
The viewing angle has been chosen so that the ordinary image appears to be undisplaced.

Suppose that unpolarised light propagating in the plane perpendicular to the optic axis of a slab of birefringent material does not strike the surface of the slab at right angles. When the incident beam enters the birefringent material, it separates into two. One beam is polarised perpendicular to the optic axis (ordinary) and the other is polarised parallel to the optic axis (extraordinary). The two polarisations travel in different directions because they have different speeds, and hence different refractive indices. So they are refracted along different paths. One consequence of this is that a single object viewed through a birefringent material will produce a double image (figure 6.16).

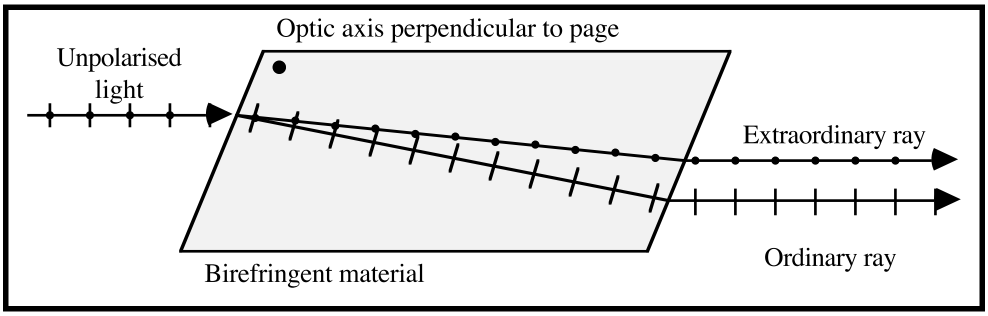


Figure 6.17. How a double image is formed
The transmitted rays seem to come from different places.

Many birefringent crystals have refractive indices which are very similar but the mineral calcite, one of the crystalline forms of calcium carbonate has noticeably different refractive indices for the ordinary and extraordinary rays.

Crystal	n_o	n_e
ice	1.309	1.313
quartz	1.544	1.553
calcite	1.658	1.486

Table 6.1. Refractive indices for some uniaxial crystals

The Nicol prism

Since calcite is colourless and absorbs very little of either extraordinary or ordinary light, very pure calcite (called 'Iceland spar') was once used to make a very good kind of polariser, called a Nicol prism. A crystal of calcite is carefully shaped and cut in two. The two parts are then rejoined using a thin layer of transparent glue whose refractive index lies between those for the e and o rays. For a suitable direction of incident unpolarised light, the ordinary rays are totally internally reflected at the boundary with the glue, while the extraordinary rays pass through. This gives a separation of the light into two components with different polarisations, travelling in quite different directions. A Nicol prism has the advantage that the light coming out is completely plane polarised and it is not tinted.

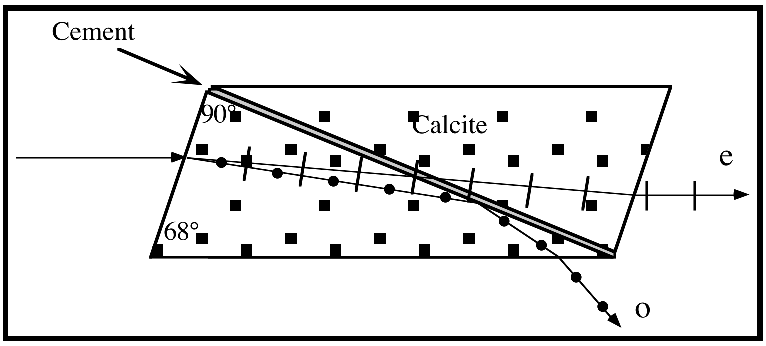


Figure 6.18. A Nicol prism
The ordinary ray is totally internally reflected at the cemented joint, leaving the completely plane polarised extraordinary ray.

6-6 POLARISATION BY SCATTERING

Light from the sky is sunlight scattered by air molecules; the scattered light propagates from the scattering molecules to the observer. If you look at the sky through a piece of Polaroid in a direction perpendicular to the sun's rays you will observe that the scattered light is polarised with its direction of polarisation perpendicular to the plane containing your line of sight and the sun. In interpreting this diagram you should remember that light is a transverse wave; it cannot have electric field oscillations with components in the direction of propagation.

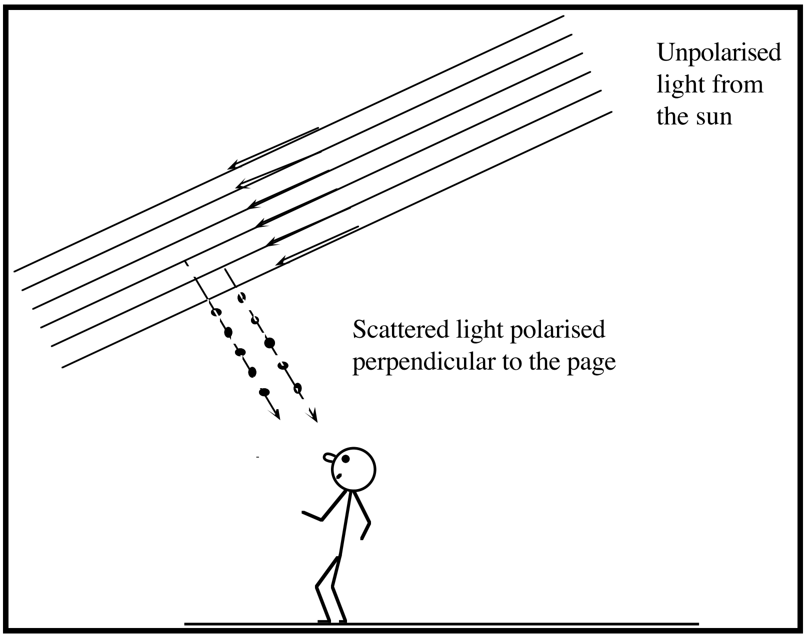


Figure 6.19. Polarisation by scattering in the atmosphere

Only light scattered through 90° is completely plane polarised. Scattering at other angles produces partially polarised light. However, when you look at the sky in a direction perpendicular to the direction of the sun, the light that you see is only weakly polarised because most of it has been scattered many times and the polarisation by scattering tends to be randomised.

6-7 POLARISATION BY REFLECTION

At boundaries between materials of different refractive index the reflectivity depends on the polarisation of the incident light beam. We can think of incident light as made up of two components, one with its *E* field parallel to the surface (in the diagram, normal to the page) and the other with its *E* field in a plane perpendicular to the surface (in the diagram, the plane of the page). Each of these components is reflected by different amounts as the angle of incidence is increased. In particular, at a certain angle of incidence, only the component with its *E* field parallel to the surface is reflected. This angle is called the **polarising angle** or **Brewster angle** ϕ_p and is given by

$$\tan\phi_p = \frac{n_2}{n_1} \qquad \text{... (6.2)}$$

where n_1 and n_2 are the refractive indices of the two materials. If the first medium is air, the Brewster angle is equal to $\tan^{-1}n_2$.

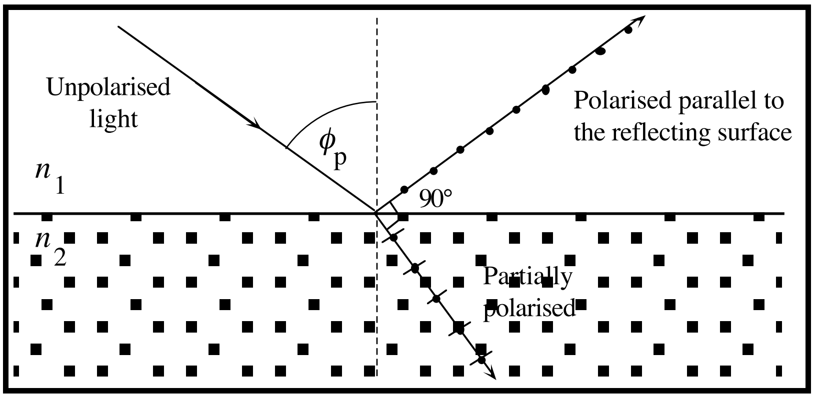


Figure 6.20. Polarisation by reflection

Note that when the reflected light is completely plane polarised, the angle between the reflected and refracted rays is 90°. At other angles of incidence the reflected light is **partially plane polarised**.

6-8 PRACTICAL AND IDEAL POLARISERS

In general, dichroic materials do not produce completely plane polarised light; the light which comes out is only partially plane polarised. Furthermore the polarised light which does get through is usually absorbed to some extent and this absorption may be greater for some some frequencies than for others. Much better, but more expensive, polarisers can be made using devices like the Nicol prism. These devices are much closer to an **ideal polariser**, which will produce completely plane polarised light, with no significant absorption of the transmitted component, for any frequency. Since each frequency component is polarised a Nicol prism can be used to polarise white light, without introducing any tinting.

Similarly polarised light can also be produced using stacks of glass plates arranged so that reflection at successive boundaries takes place at the Brewster angle.

Remember that Malus's law gives accurate results only for ideal polarisers.

6-9 OPTICAL ACTIVITY

Some materials (e.g. sugar solutions and many crystals) have different refractive indices for left and right circularly polarised light. This phenomenon is called **optical activity**. The effect of such a material on linearly polarised light can be deduced by resolving the light into left and right circularly polarised components with equal magnitudes. One of these traverses the material faster than the other so it moves ahead. When the two circularly polarised components emerge from the material their phase difference has changed. The combination of the two circularly polarised components is once again linearly polarised light with a new orientation. So optical activity is a rotation of the plane of polarisation of plane polarised light - the thicker the medium the greater the angle of polarisation.

In a technique known as saccharimetry this rotation is used to analyse sugar solutions. Some sugars rotate the plane clockwise; these are described as dextrorotatory. Other sugars which rotate the plane anticlockwise are described as levorotatory. The magnitude of the effect depends on the concentration of sugar in the solution.

6-10 PHOTOELASTICITY

Birefringence can be induced in glass and some plastics by mechanical stress. This phenomenon is called **photoelasticity**. Photoelasticity can be used to study stress patterns in loaded engineering structures and other objects. A perspex model of the object (for example an engine part, a bridge or a bone) is constructed and placed between crossed polarisers. When external forces are applied to the model, the internal strains cause birefringence, so that some of the light now gets through and the light patterns reveal the patterns of the internal strains. Since the refractive indices also depend on the frequency of the light the resulting patterns are brightly coloured when incident white light is used.

6-11 MISCELLANEOUS APPLICATIONS

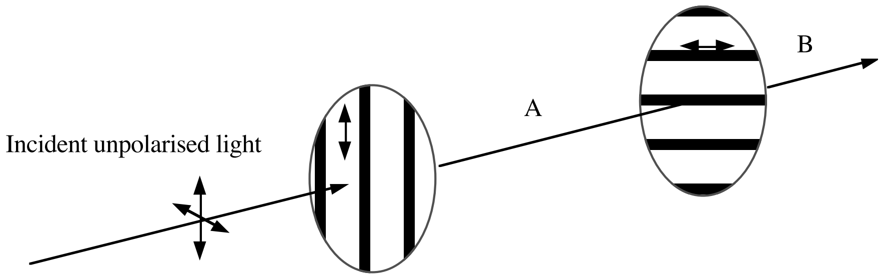
- A pair of polarisers can be used to control the intensity of light by varying the angle between their polarising axes.
- Polarising sunglasses are used to reduce glare. Since light scattered from the sky and light reflected from shiny surfaces such as water or hot roads is partially plane polarised, the appropriately oriented polarising material reduces the intensity of such light and the associated glare.
- When a thin slice of rock is placed between crossed polarisers in a petrological microscope the appearance of the mineral grains depends on their crystal shape, their light absorbing properties and birefringence. This aids in their identification.
- Birefringence can be induced in some materials by high electric fields (a phenomenon known as the Kerr effect). This effect can be used to make fast shutters for high speed photography.

THINGS TO DO

- Use a pair of polarising sunglasses to examine the polarisation of light reflected from a pane of glass or a shiny tabletop. How can you determine the polarising axis of the sunglasses? Can you measure or estimate the Brewster angle? Can you determine the refractive index of glass or furniture polish?
- Use a pair of polarising sunglasses to examine the polarisation of light from the sky. What is the orientation of the partial polarisation? From which part of the sky does the polarised light come?

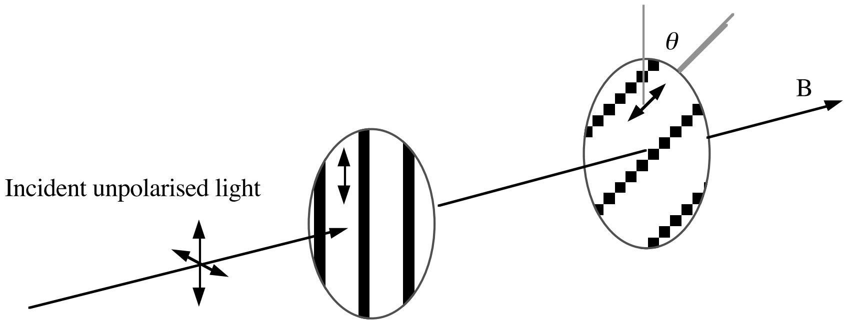
QUESTIONS

Q6.1 a)

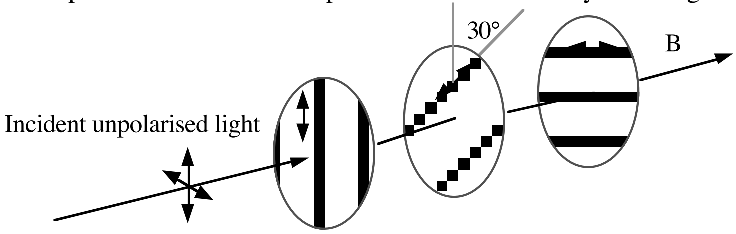


Unpolarised light of intensity I_{in} is incident on two ideal polarisers which have their polarising axes at 90° to each other. What are the polarisation and intensity of the light at A and B?

b) Suppose that the two polarising axes are at an angle θ to each other. What are the polarisation and intensity of the light at B?



c) Suppose that a third polariser is placed between the two crossed polarisers with its polarising axis at an angle of 30° to the first polariser. What are the polarisation and intensity of the light at B?



Q6.2 The refractive indices for ordinary and extraordinary waves travelling at right angles to the optic axis in quartz are $n_o = 1.544$ and $n_e = 1.553$. A quarter wave plate is one for which the two waves get exactly a quarter of a wavelength out of step after passing through it.

What is the thickness of the thinnest possible quarter wave plate for a wavelength of 600 nm?

Will such a quarter wave plate for $\lambda = 500$ nm be thicker or thinner?

Show that a much thicker piece of quartz is required if it is to act as a quarter wave plate at several visible wavelengths.

Q6.3 Draw a set of diagrams of electric field vector to show how two linearly polarised waves with perpendicular polarisations, the same frequency and phase, but different amplitudes superpose to form another linearly polarised wave. Draw another set of sketches to show what happens if the phase of one the component waves is advanced by half a cycle ($\pi/2$).

Q6.4 A material has a critical angle of 45°. What is its polarising angle?

Q6.5 How do polarising sunglasses reduce glare? Why do they have an advantage over sunglasses which rely on absorption of light only?

Q6.6 An unpolarised beam of light passes through a sheet of dichroic material which absorbs all of one polarisation component and 50% of the other (perpendicular) component. What is the intensity of the light which gets through?

Q6.7 Which would be thicker, a quarter wave plate made from calcite or one made from quartz? See table 6.1.

Discussion questions

Q6.8 How could you distinguish experimentally among beams of plane polarised light, circularly polarised light and unpolarised light?

Q6.9 Can polarisation by reflection occur at a boundary where the refractive index increases, for example with light going from water to air?

Q6.10 Ice is birefringent. (See table 6.1.) Why do you not see a double image through an ice block?

Q6.11 How could you identify the orientation of the optic axis in a quarter wave plate?

Q6.12 What happens to circularly polarised light when it goes through a quarter wave plate? What happens to it in a half wave plate?

Q6.13 One way of reducing glare from car headlights at night would be to fit polarisers to headlights and windscreens. How should the polarisers be arranged. Is this a good idea? What are the disadvantages?

Q6.14 A salesperson claims that a pair of sunglasses is polarising. How can you check the claim before leaving the shop?

Q6.15 There was once a kind of three dimensional movies based on polarised light. How might such a system work?

Q6.16 Nicol prisms, which have to be made from very pure crystals of calcite, are very expensive compared with mass-produced Polaroid sheets. What are the advantages of a Nicol prism over Polaroid?

Q6.17 What happens when circularly polarised light goes through a quarter wave plate? You can work it out by studying figure 6.15. First look at what a quarter wave plate does to linearly polarised light. What do two quarter wave plates do to linearly polarised light?

L7 OPTICAL SYSTEMS

OBJECTIVES

Aims

Your aim here should be to acquire a working knowledge of the basic components of optical systems and understand their purpose, function and limitations in terms of concepts learned from earlier chapters. You should also be able to apply your knowledge of optics to describe the structure, function and limitations of a simple camera. A long term goal is that when you encounter new or unfamiliar optical instruments in future, you will be able to understand, or figure out, their function and limitations.

Minimum learning goals

1. Explain, interpret and use the terms:

   (a) *lens system, objective, eyepiece, optical relay,*

   (b) *spherical aberration*, *chromatic aberration*, *coma*, *curvature of field*, *astigmatism*, *distortion*,

   (c) *principal planes*, *principal points*, *focal points, nodal points, cardinal points, entrance pupil, stop, aperture, focal ratio, f-number, depth of field, image brightness,*

   (d) *cornea, aqueous humour, vitreous humour, iris, retina, rods*, *cones*, *photopic vision, scotopic vision, accommodation, hyperopia, myopia, astigmatism.*
2. Describe and discuss the nature of aberrations.
3. Describe and apply ray-tracing techniques for locating images formed by lens systems whose cardinal points are given.
4. Explain how entrance pupil and aperture affect the illumination of images, and do simple calculations (photographic exposures, for example) related to these.
5. Draw a labelled diagram showing the structure of a simple camera and name its parts. Describe and discuss the function of the camera.
6. Describe the optical structure and function of the eye.

TEXT & LECTURE

7-1 OPTICAL INSTRUMENTS

Optical instruments whose function is to produce images can be divided into two groups.

- **Photographic instruments** produce real images. Examples include cameras, projectors and eyes.
- **Visual instruments** produce virtual images which can be looked at with the eye. Examples are magnifying glasses, telescopes and microscopes.

Optical instruments are made up of optical components classified as objectives, eyepieces and optical relays. The component nearest the object is called the **objective** and its purpose is to form a real (intermediate) image of the object. An **eyepiece** is used, essentially as a magnifying glass, to look at the image produced by the objective. The purpose of an **optical relay** is to transfer an intermediate image from one place to another, more convenient, location. Optical relays can also change the orientation of an image, e.g. make an inverted image upright. Prismatic binoculars (figure 7.1) are an example which uses all three types of component.

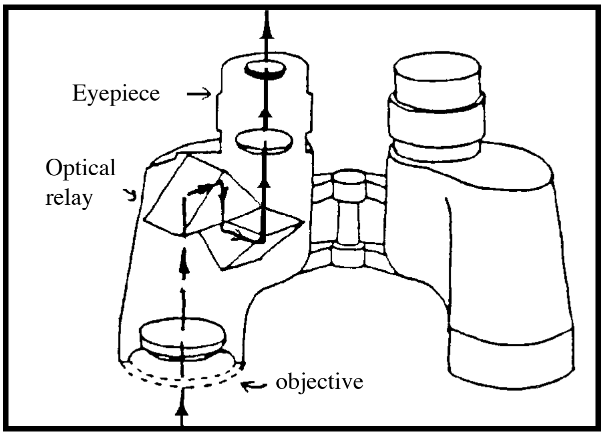


Figure 7.1. Prism binoculars

Objectives and eyepieces are designed to act like single ideal lenses, but they are usually made up of a number of lens elements. That is done in order to reduce lens aberrations and thus give clearer images.

7-2 ABERRATIONS

The equations relating image and object distances, derived earlier assuming paraxial rays, describe the performance of an "ideal" lens. If paraxial approximations are not made, the way a real lens forms images can still be calculated, although with difficulty. The differences between the performance of a real lens and an ideal lens are called **aberrations**.

There are six types of aberration. Spherical aberration and chromatic aberration were discussed in chapter L3. The remaining four are: coma, curvature of field, astigmatism, distortion.

Coma

Coma is an aberration which shows up in the images of points well away from the principal axis. The image (I_C in figure 7.2) formed by rays which pass through the central region of the lens is further from the principal axis and also further from the lens than the the image (I_E) formed by rays which go through the region near the edge of the lens. The net effect is that the image of an off-axis point has a comet-like or pear-shaped appearance.

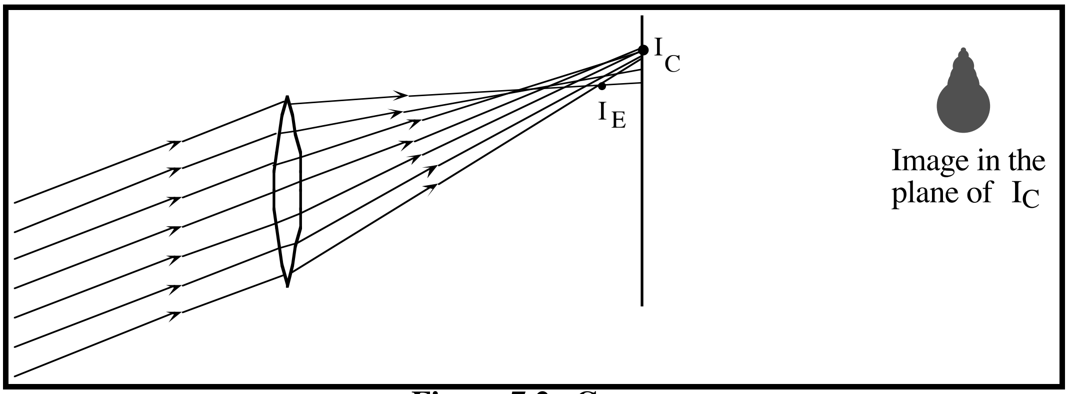


Figure 7.2. Coma

Curvature of field

The image of a plane object, perpendicular to the principal axis, is really located on a curved surface. The effect is that if we look at the images formed in a plane perpendicular to the principal axis, part of each image will be out of focus. If we adjust the part of the image near the axis for good focus then the edges are out of focus; when the edges are well-focussed the central part is fuzzy.

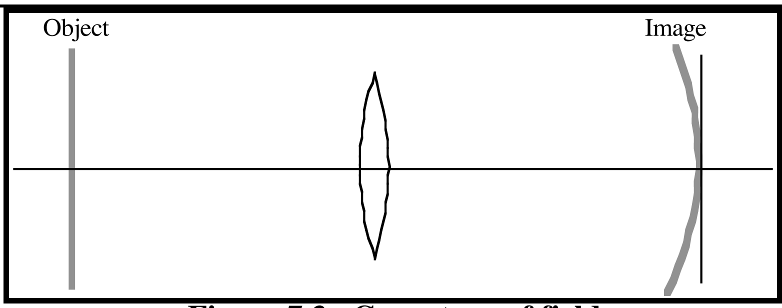


Figure 7.3. Curvature of field

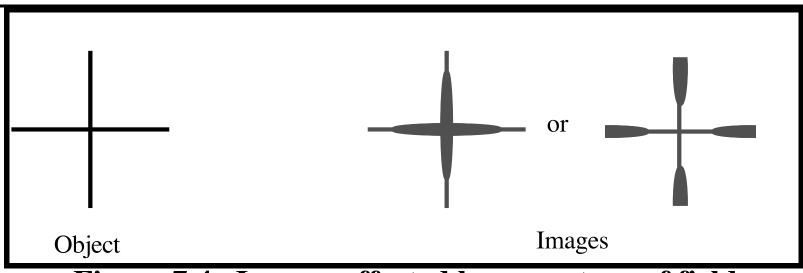


Figure 7.4. Images affected by curvature of field

Astigmatism

Astigmatism is a complex geometrical aberration associated with the fact that the images of points off the principal axis can become elongated. Although astigmatism occurs in symmetrical lenses the effect can also be produced by asymmetries in the lens. For example rays which pass through a vertical section of the lens may come to a focus closer to the lens than do the rays passing through a horizontal section.

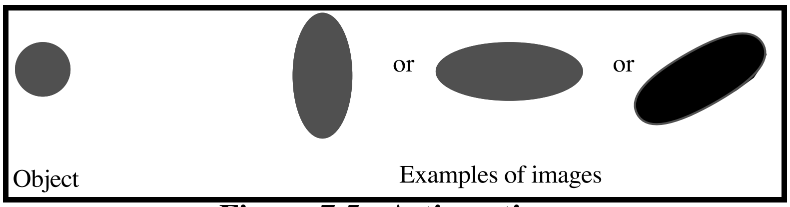


Figure 7.5. Astigmatism

Distortion

Distortion is the alteration of the shape of an image. Two common kinds of distortion are pincushion and barrel distortion. In pincushion distortion the image of a square grid has the corners pulled out whereas in barrel distortion they are pushed in.

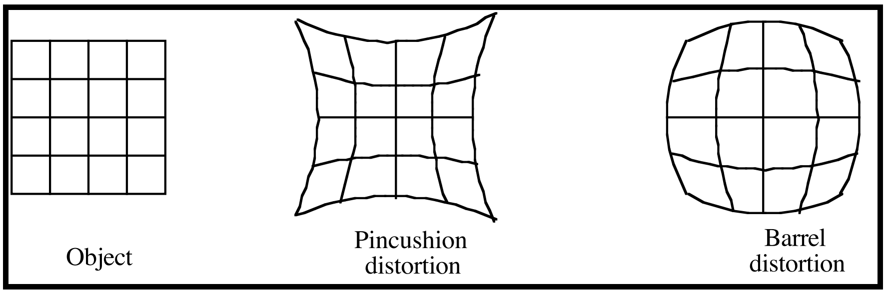


Figure 7.6. Distortion

7-3 IMAGE FORMATION BY OPTICAL SYSTEMS

Ray tracing for a single thin lens

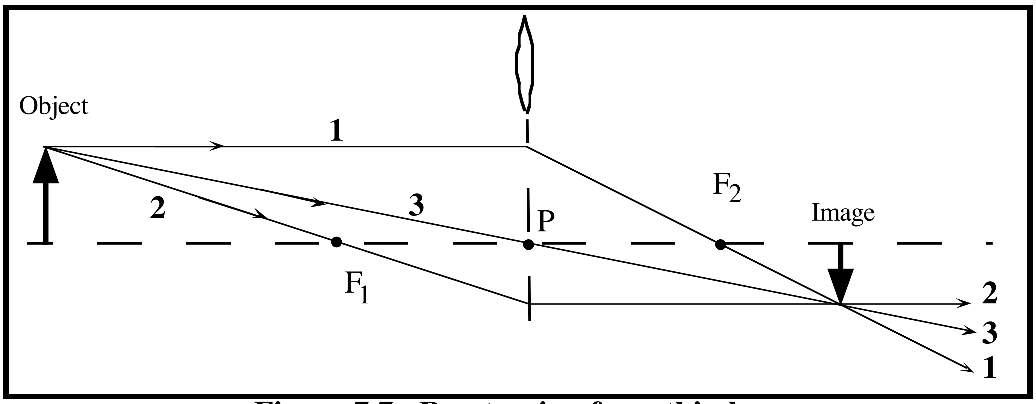


Figure 7.7. Ray tracing for a thin lens

In chapter L3 we discussed the following rules for ray tracing using the paraxial approximation for a thin lens (figure 7.7).

1. Rays incident parallel to the principal axis after passing through the lens, are deflected so that they pass through (or appear to come from) the second focal point, F_2.

2. Incident rays passing through the first focal point, F_1, are refracted so that they emerge parallel to the principal axis.

3. Rays which pass through the centre of the lens emerge in the same direction.

Since the lens is thin, all the constructions can be made by making all deflections at a plane (called the principal plane) through the centre of the lens.

Ray tracing for a thick lens or any optical component

The function of a thick lens or a system of any number of lenses can be described in a similar manner. The properties of such a system can be described in terms of two **focal points** (as for a thin lens) as well as two **principal points** and two **principal planes** (instead of one). In addition there are two new points, called **nodal points**.

The ray-tracing rules are now as follows (see figure 7.8).

1. Rays which come in parallel to the principal axis are deflected at the **second principal plane** towards the second focal point.

2. Rays which come in through the first focal point are deflected at the **first principal plane** so that they come out parallel to the principal axis.

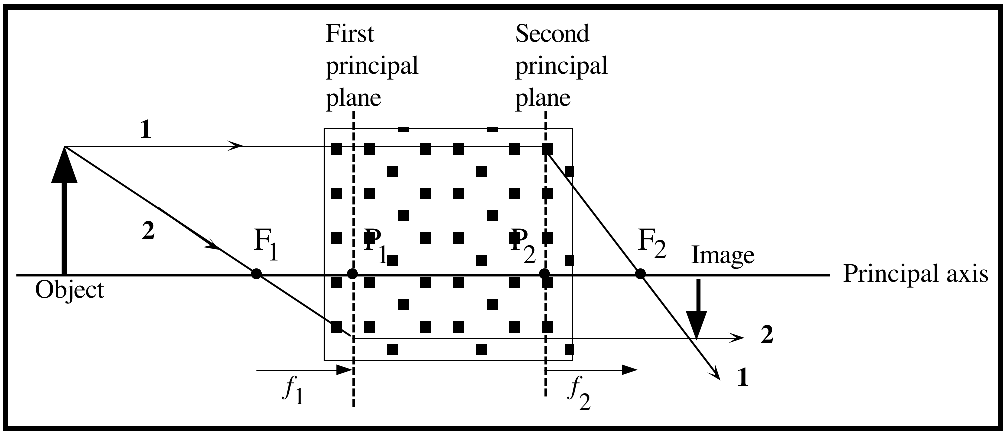


Figure 7.8. Ray tracing using principal planes

Note that these rules do not give the actual paths of the light rays while they are inside the lens system, but provided that the paraxial approximation is still good, they do give the correct paths of the rays which come out of the system. In the construction the complex set of deflections for the real rays is replaced by a single deflection for each construction ray, which takes place at one of two principal planes.

When the media on either side of the optical system have the same refractive index the distances F_1P_1, F_2P_2, which are called the first and second focal lengths (f_1,f_2) are equal. When the media on each side of the optical system are different (as in the eye or in oil-immersion microscopy) the two focal lengths have different values. For example the first focal length of a human eye is 17 mm while the second focal length is 23 mm.

Notice that for a single thin lens, the principal planes coincide, so a single principal plane suffices.

For a single thin lens we had a third ray-tracing rule: a ray passes through the centre undeflected. To get the equivalent of the rule for a lens system we need to define two more special points, the **nodal points**, on the principal axis of the system. The third rule is as follows.

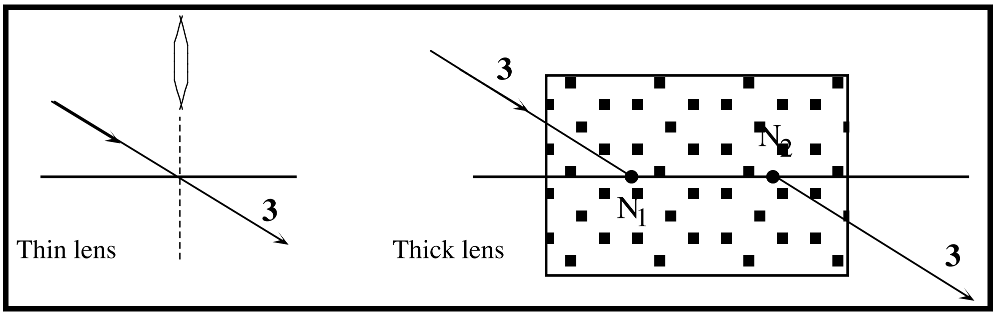


Figure 7.9. Ray tracing using nodal points

3. For a ray coming in to the first *nodal point* N_1, construct a ray from the second nodal point N_2 in the same direction.

When the medium on each side of the system is the same, the nodal points coincide with the principal points. If the media on the two sides are different the nodal points no longer coincide with the principal points.

The focal points, the principal points and the nodal points are called the **cardinal points** of a lens system.

The lens equation

With object distance defined as the distance from the object to the first principal plane and the image distance as the distance from the second principal plane to the image (figure 7.10), the lens equation (introduced in chapter L3) still works for paraxial rays.

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} . \qquad \text{... (7.1)}$$

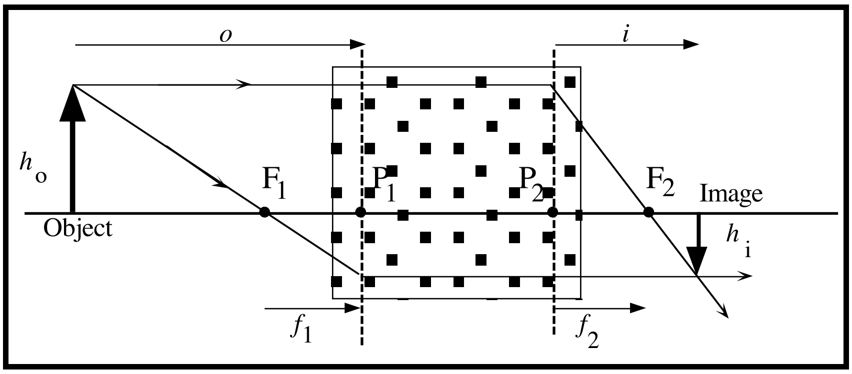


Figure 7.10. Definition of distances and lateral magnification

7-4 MAGNIFICATION

The **linear magnification** of an optical system is defined as the ratio of image size to object size. It is useful to distinguish two ways of specifying linear magnification. The first which we have already defined in chapter L3, is strictly the **lateral magnification**, defined as

$$m = \frac{\text{image height}}{\text{object height}} .$$

The magnification is still given by the formula

$$m = -\frac{i}{o} . \qquad \text{... (7.2)}$$

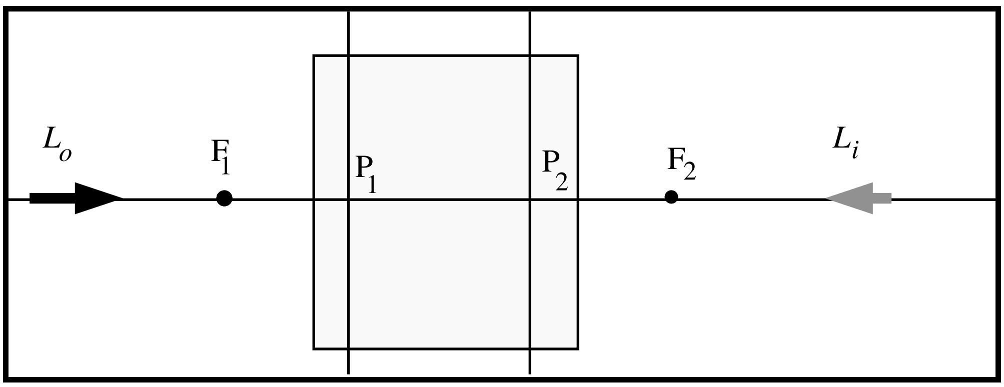


Figure 7.11. Longitudinal magnification

$$m = \frac{L_i}{L_o}$$

7-5 BRIGHTNESS OF THE IMAGE

The brightness of an image is determined by the amount of light passing through the optical system which in turn is determined by

(i) the diameter of the lenses, or

(ii) the diameter of the **apertures** (holes) in any opaque screens which are known as **stops** or **diaphragms**.

Example

For example figure 7.12 shows the formation of a point image of a distant object by a single lens. All the light collected by the lens, shown in the shaded region of the diagram, goes to form the image. The wider the lens, the more light we get, so the image is brighter.

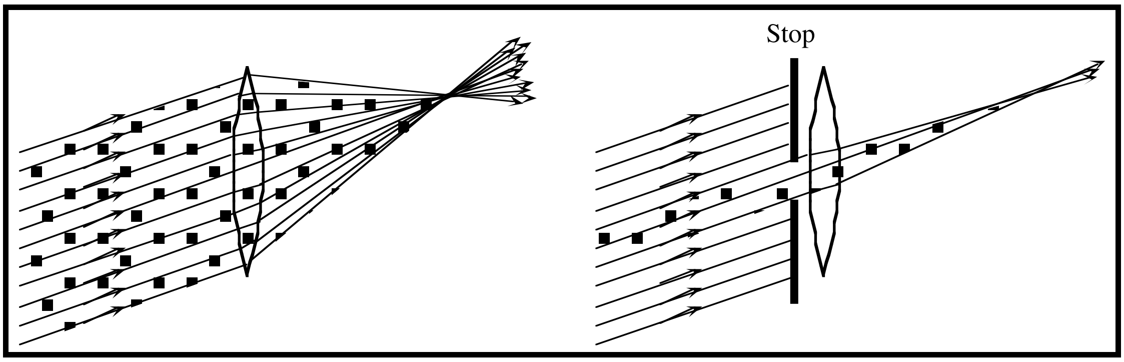


Figure 7.12. How lens diameter affects image brightness
All the light which goes through the lens goes into the image. The trade-off against less light is reduced aberrations.

On the other hand using all of a lens to form an image can result in noticeable aberrations, so we often deliberately restrict the amount of light using a screen with a hole, often called an **aperture stop**, to stop some of the light. The image in that case will not be so bright.

Example

In a two-lens system the stop is often placed between the lenses so that some of the light which enters the system does not get through to form the final image. Again the image brightness depends on the diameter of the aperture.

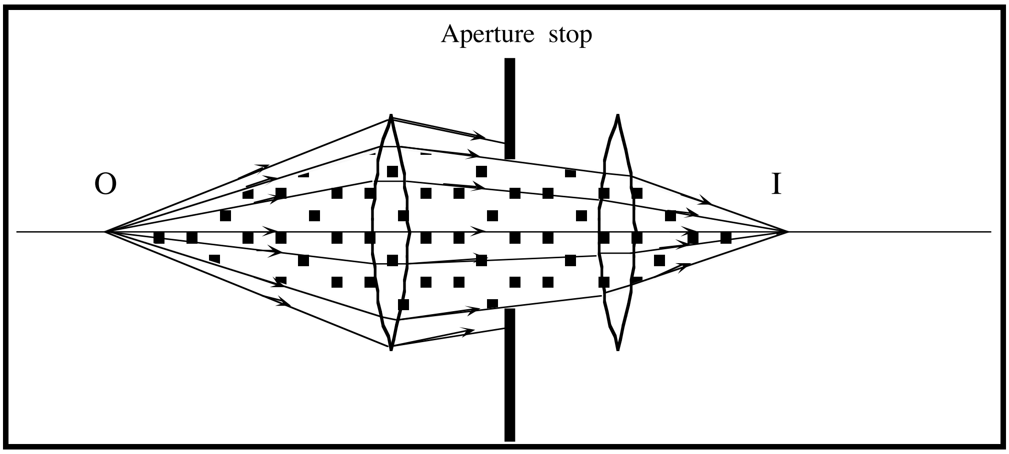


Figure 7.13. A stop used in a multiple lens system

If you look at the stop located between two lenses you can do so only by looking into the system from one side or the other. If you look in from the "object side" you will see an image of the aperture stop formed by the first lens. If the first lens is a converging lens, that image will be enlarged. The image of the hole (stop) seen from the object side is called the **entrance pupil** of the system (figures 7.14, 7.15). Similarly if you were to look into the system form the "image side" you would see a different image of the hole. That image is called the **exit pupil** (figure 7.15).

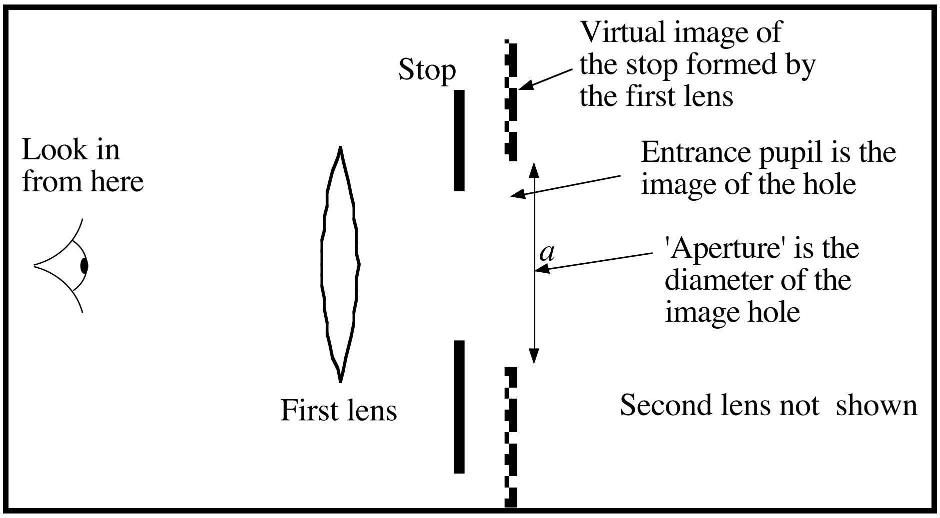


Figure 7.14. Entrance pupil of a lens system

We can describe how much light gets through the system in terms of the size of the actual aperture, or the size of the entrance pupil, or the size of the exit pupil. Figure 7.15 shows how the light from a point source is restricted in terms of these ideas. The bundle of rays which gets through is limited by the cone with the object point as apex and the entrance pupil as base. An alternative, equivalent, specification is the bundle of rays converging on the image point in the cone based on the exit pupil.

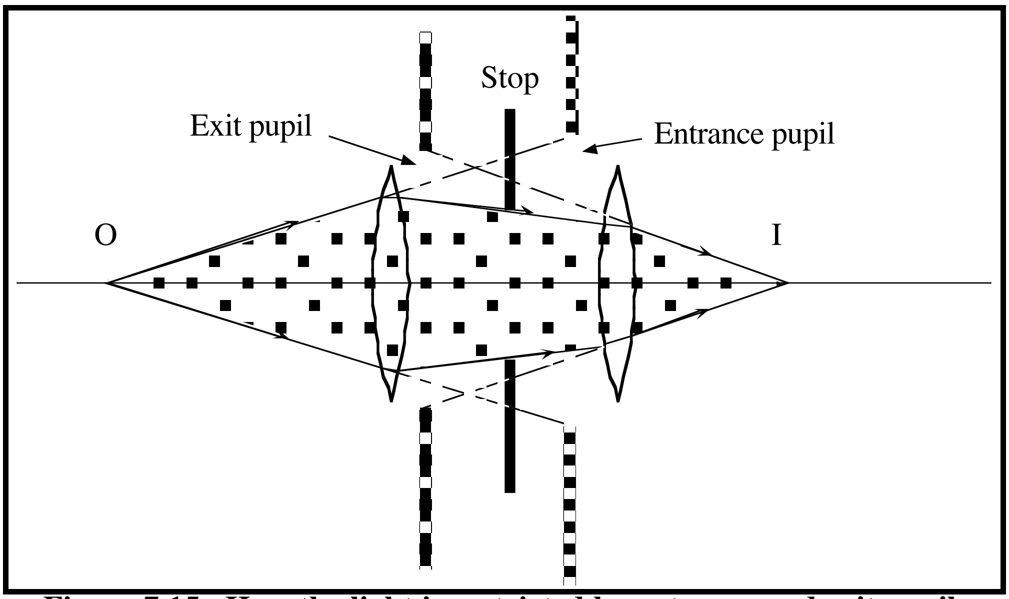


Figure 7.15. How the light is restricted by entrance and exit pupils

In order to use this approach we need to know, not only the location of the entrance pupil, but also its diameter. If we look at the stop through the first lens, it appears to have diameter a, so a is the diameter of the entrance pupil. For a single lens with no aperture stop, the diameter of the entrance pupil is just the diameter of the lens.

Aperture

The two important parameters of a lens system which affect the brightness of images are the diameter of its entrance pupil, commonly called the **aperture**, and its focal length. For a given object brightness, the image brightness is actually determined by the ratio of the aperture to the focal length. We have already seen that for a given focal length, the brightness increases with increasing aperture. But for a fixed size of aperture, a short focal length produces brighter (and smaller) images. For a given object at a reasonably large distance, systems with the same value of the ratio, aperture divided by focal length, produce images with the same image brightness. (This result breaks down at small object distances.)

Aperture is often specified as a fraction of the focal length. For example, a system which has an aperture of *f*/8 has an entrance pupil whose diameter is one eighth of its focal length. (The slash is a division sign, so the aperture is equal to *f* divided by 8.) Other terms used in connection with this idea include the following.

- The **focal ratio** (n) is the ratio of the focal length to the aperture, f/a. The terms focal ratio and **f-number** both refer to the divisor (n) in the expression f/n. Thus if the aperture (a) is equal to *f*/8 then the focal ratio and the f-number are both equal to 8.
- The **aperture ratio** is the ratio of the aperture to the focal length, i.e. the reciprocal of the focal ratio. For an aperture of *f*/8 the aperture ratio is 1/8.

Note that the aperture is a distance, whereas aperture ratio and focal ratio (f-number) are pure numbers without units.

Example

A single lens with a diameter of 10 mm and a focal length of 50 mm has an aperture of *f*/5, an f-number of 5, an aperture ratio of 0.2 and a focal ratio of 5.

7-6 RESOLUTION

The *theoretical* limit to the smallest objects that can be distinguished with an optical component is determined by diffraction. The smallest hole in the optical component, normally the stop, has the greatest diffraction effect.

From chapter L5 (equation 5.2), the angle between the central maximum and the first minimum of the diffraction pattern of a point object by a circular aperture of diameter a is given by

$$\theta \approx \sin\theta = 1.2\frac{\lambda}{a} \quad . \qquad \text{... (7.3)}$$

The ideal image of a distant point object should be a point at a distance equal to f, the focal length, from the lens system. In reality the image is a small circular diffraction pattern in which the radius of the central bright region is given approximately by

$$r = \theta f$$

$$= 1.2\,\lambda\,\frac{f}{a} \quad . \qquad \text{... (7.4)}$$

Notice that this result contains the ratio of aperture to focal length again. The larger the aperture as a fraction of the focal length, the smaller is the diameter of the diffraction pattern.

It is worth noting, however, that the practical limit to the resolution of an optical system is normally set by aberrations and other imperfections, not by diffraction.

7-7 THE CAMERA

In its simple form a camera consists of a light-tight box, a compound objective lens, a shutter and a film. A variable aperture controls the image brightness.

The objective, which forms a real image at the film plane, is usually specified in terms of its focal length, and its maximum usable aperture. For example a camera lens might be marked 70 mm, *f*/4.5. The image brightness is controlled by varying the size of the aperture which is normally described as a fraction of the focal length. Depending upon the lens, aperture settings can range from about *f*/32 to *f*/1.

Focussing is achieved by moving the objective relative to the film plane, which is fixed in the camera body. Only one object plane is in focus at any one time. The images, on the film, of point objects which are not in the object plane are small discs whose size is determined by the distance of the film plane from their true image plane and by the angle of convergence of the rays forming the image (i.e. by the focal ratio). The range of object distances for which the image discs are acceptably small is called the **depth of field**.

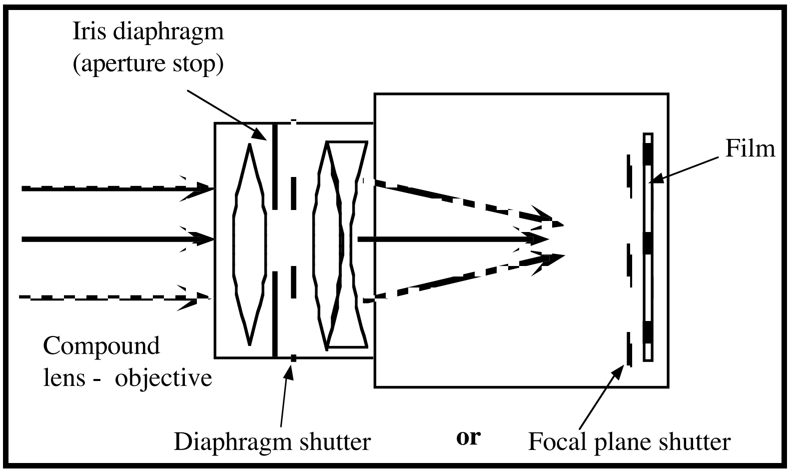


Figure 7.16. Parts of a camera (schematic)

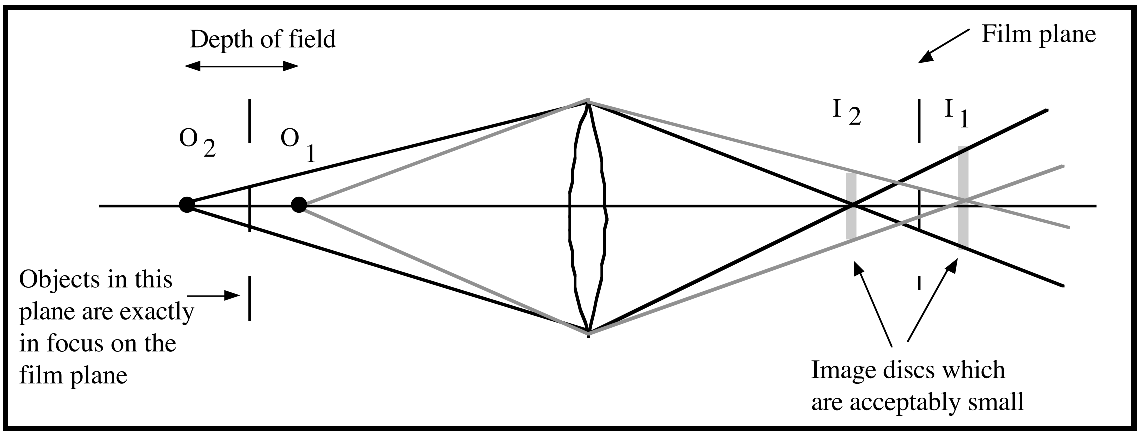


Figure 7.17. Depth of field

Note. The larger the aperture, the greater is the apex angle of the cone of rays coming to a focus at the image. Image points must then be formed closer to the film plane if the image is to remain acceptable. The depth of field is consequently reduced.

7-8 THE HUMAN EYE

The human eye is almost spherical, being about 24 mm long and 22 mm across, most of the eye being contained within a strong flexible shell called the sclera. The eye contains an optical system that produces real images on the light-sensitive **retina**. Most of the focussing is done by the outer surface of the **cornea**. The space behind the cornea, the anterior chamber, is filled with a watery liquid called the **aqueous humour** whose refractive index (1.336) is only a little less than that of the cornea (1.376), so there is very little further bending of light rays at the inside surface of the cornea. After the cornea light must pass the variable opening or **pupil** formed by the **iris** before it strikes the lens of the eye.

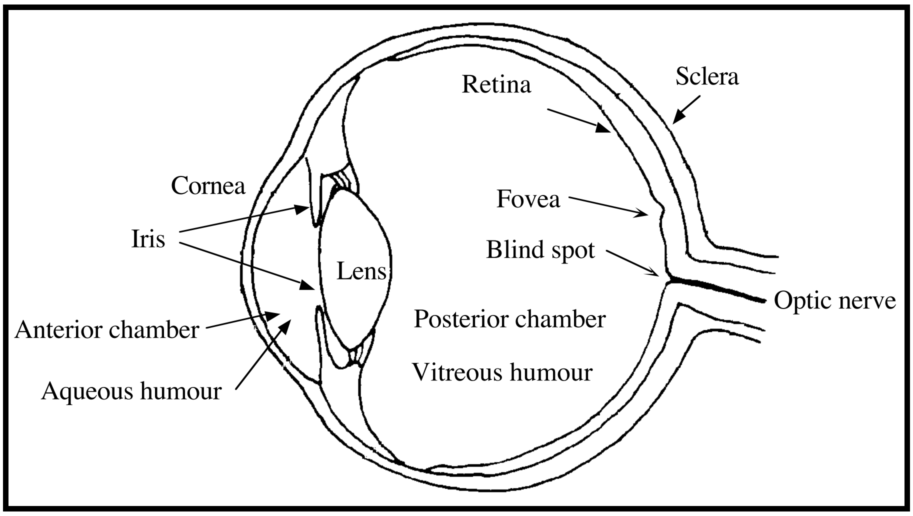


Figure 7.18. The human right eye - horizontal section

Variation in the focal length of the eye, and hence its ability to form images of objects at different distances - a process called **accommodation** - is achieved by altering the shape of the **lens** of the eye. The lens is a complex layered structure, whose refractive index varies within the lens. Light finally passes through the posterior chamber of the eye which is filled with a transparent jelly-like substance called the **vitreous humour** (refractive index 1.337). Inverted real images are formed on the retina at the back of the eye.

Normally the eye can focus on objects further than 25 cm from the eye, in fact most young people can focus much closer than that. However, instrument designers need to refer to a standard, close, distance at which most people can focus so the value of 25 cm has been chosen. That distance is often called the **least distance of distinct vision**.

Rods and cones

In the retina are two kinds of sensors: **rods** and **cones** (figure 7.19). The cones function at high levels of illumination and mediate colour vision which is called **photopic vision**. The rods function at low levels of illumination when the eye has become dark-adapted and do not give the sensation of colour, a process called is **scotopic vision**. The different spectral responses of the rods and cones are sketched in figure 7.20.

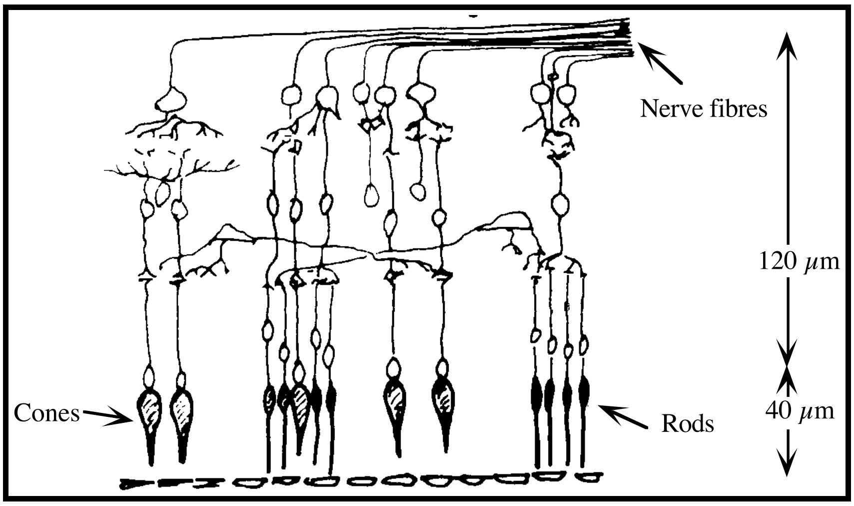


Figure 7.19. Structure of the retina

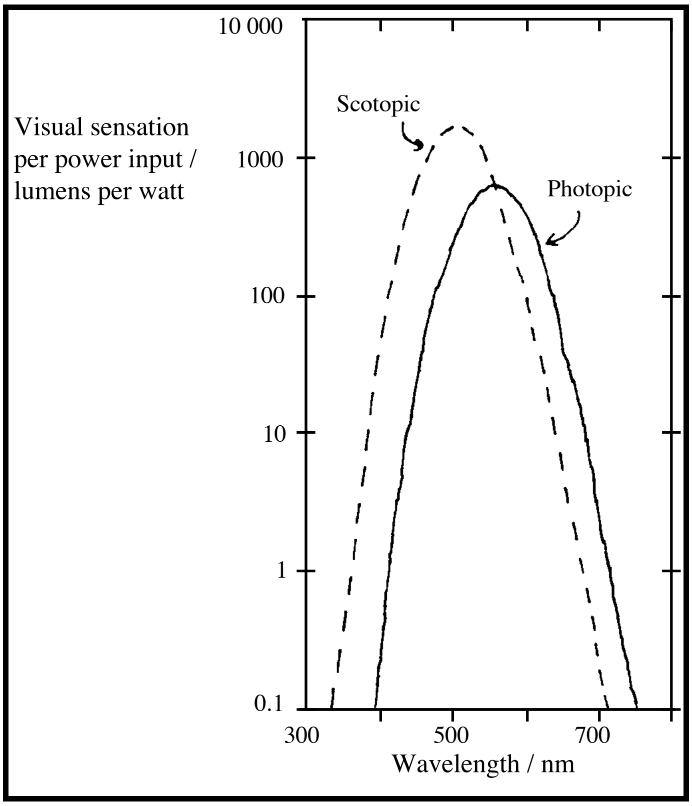


Figure 7.20. Sensitivity of the human eye

For intermediate light levels the spectral response is between the scotopic and photopic vision responses, which is is called **mesopic vision**.

Resolution

The aperture stop of the eye, called the **iris**, changes its diameter as the intensity of the incident light varies. In daylight its diameter is about 3 mm. The best resolution at the centre of the eye is about 1 minute of arc (3×10^{-4} radian). The limit is set by diffraction and by the resolving power of the retina. The centre of the eye is populated exclusively by cones which have a diameter of about 2 μm. For two point sources to be resolved their images on the retina must be separated by about 5 μm, which is comparable to the size of the diffraction pattern (about 7 μm).

Defects of the eye

The range of accommodation in eyes decreases with age. To make up for this loss and also to correct defects, spectacles are used. Common defects are as follows.

Hyperopia (or hypermetropia) is the condition in which the image of a distant object (at relaxed vision) lies behind the eye. A converging spectacle lens, which adds power to the system, is used to correct this defect.

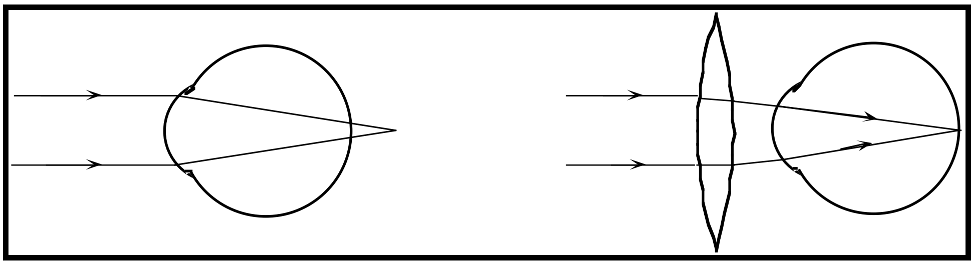


Figure 7.21. Correction of hyperopia

Myopia is the condition in which the image of a distinct object is formed in front of the retina. It is corrected using a diverging spectacle lens, i.e. by decreasing the power of the system.

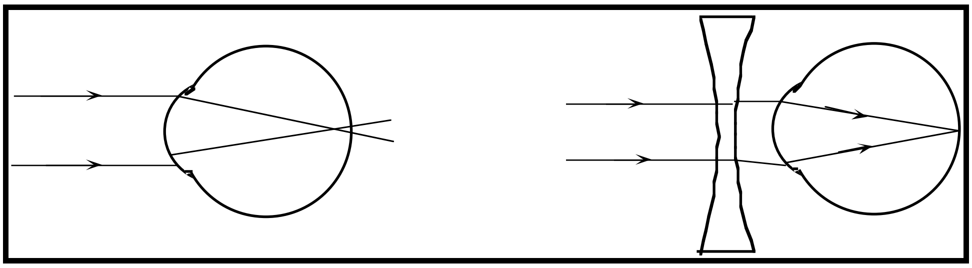
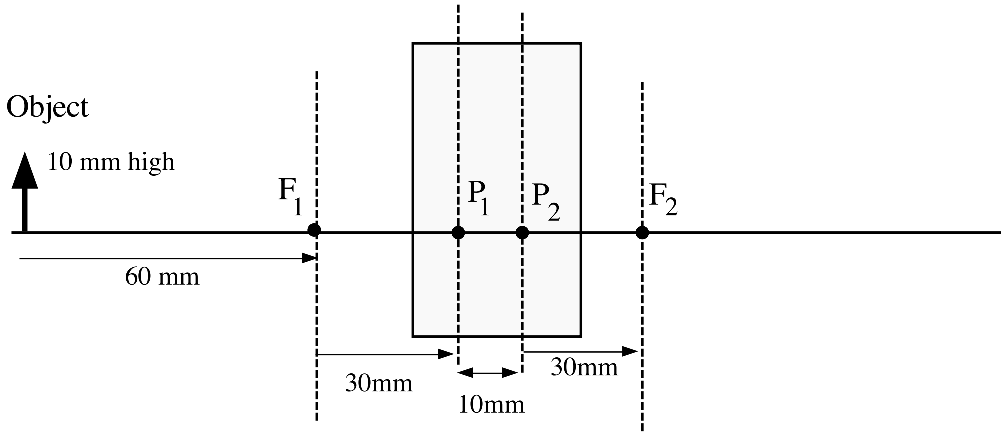


Figure 7.22. Correction of myopia

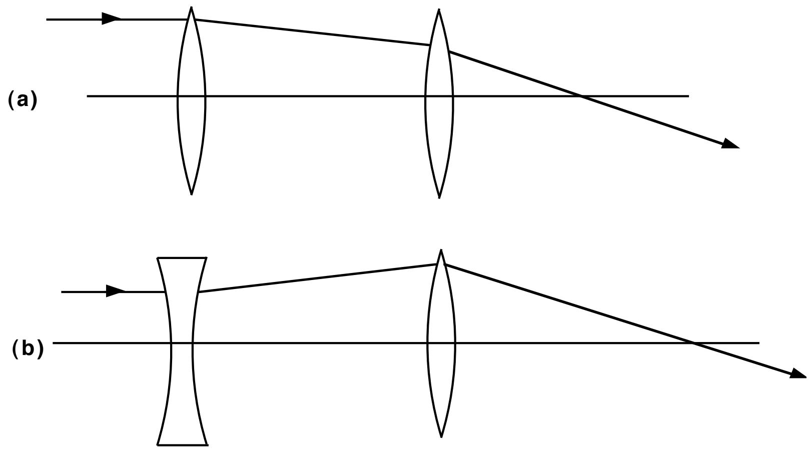
Astigmatism of the eye is a condition in which the radii of curvature of the cornea and the lens are not the same for all cross-sections containing the principal axis of the eye. A lens which has different curvatures can be used to compensate for the defect. The practical solution to correcting astigmatism depends on the other optic defects in the eye. If astigmatism is the only defect then a cylindrical lens can be used. If other defects are present as well, then the lens may be given one spherical and one cylindrical surface. It may be appropriate either to *add* power to the weaker axis or to *subtract power* from the stronger axis.

QUESTIONS

Q7.1 Use the ray tracing method described in the lecture to locate the image of the object formed by the optical component below.



Q7.2 A ray parallel to the axis enters an optical system and passes through two lenses as shown in (a) and (b). Using only a straight-edge, locate the second principal plane in each case.



Q7.3 What is the diameter of a lens, focal length +100 mm, aperture *f*/8?

Q7.4 The lens of a camera has a focal length of +50 mm. What is the tallest object standing 10 m away that can give an image fitting onto the film? (The image must be smaller than 35 mm.)

Q7.5 The lens of a camera has a focal length of 50.0 mm. Calculate how far the lens must be from the film in order to focus an object 0.20 m away. Repeat the calculation for an object at infinity. What range of travel of the lens is required to focus objects from 0.20 m to infinity.

Q7.6 A myopic eye cannot focus on objects further away from the eye than a point called the *near point*. What are the power and the focal length of a spectacle lens which enables a person whose near point is at 3 m to see distant objects (i.e. objects at infinity)?

Discussion questions

Q7.7 If you move a camera while the shutter is open you get a blurred picture, but if you move your eye you can still see clearly. Discuss.

Q7.8 In what ways are the eye and a camera similar? How to they differ?

L8 VISUAL INSTRUMENTS

OBJECTIVES

Aims

As a climax to this unit, you should end up with a good understanding of the physical principles of visual instruments. The main topic is the principles of microscopy. The section on telescopes is included for interest - it is not examinable.

Minimum learning goals

1. Explain, interpret and use the terms:

   near point, least distance of distinct vision, relaxed vision, angular size, angular magnification, visual instrument, simple magnifier, compound microscope, eyepiece, objective, optical tube length, numerical aperture, field of view, condenser, resolution of a microscope, resolving power, maximum useful magnification, dark field illumination, interference microscopy.
2. Describe, explain and discuss the operation of a simple magnifier. Solve simple quantitative problems on magnification.
3. Draw diagrams showing the essential structure and function of a compound microscope. Describe and explain how it works. Solve simple quantitative problems related to its magnifying function.
4. Describe and discuss resolution and useful magnification of microscopes.
5. Describe and discuss the brightness of images in a microscope and techniques for illuminating specimens.
6. Describe and explain the techniques of *interference microscopy*.

Extra goals

7. Describe and explain the basic principles of telescopes.

PRE-LECTURE

8-1 ANGULAR SIZE

Visual instruments are used to make an object appear larger by increasing the angle subtended at the eye by its edges. In figure 8.1, the object subtends an angle α at the eye so α is called the object's **angular size**. When α is small its value in radians is approximately equal to the ratio of the object's linear size to its distance from the eye:

$$\alpha \approx \frac{h}{d} . \quad \text{... (8.1)}$$

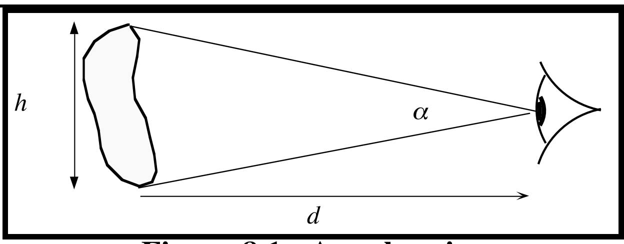


Figure 8.1. Angular size

Q8.1 Work out the following examples.

a) What angle is subtended at your eye by the width of your thumb, when you hold your arm outstretched?

b) The radius of the sun is 7.0×10^5 km and its distance form the earth is 1.5×10^8 km. What is the angle subtended by the sun at a telescope on earth?

c) What is the angle is subtended at your eye by a tiny creature 0.1 mm long at a distance of 0.25 m?

LECTURE

8-2 ANGULAR MAGNIFICATION

To see more detail in an object, we must make it look bigger. We need to make the object appear to subtend a larger angle at the eye so that the image on the retina is larger. One way of doing that is to get closer to the object. If you can't get closer you can use a **visual instrument** to achieve the magnification. For example, if the object is far away and we cannot get closer we can use a **telescope** to increase the angle. On the other hand if the object is very small we cannot bring it too close because our eyes would be unable to focus on it properly. In that case we can use a **magnifier** or **microscope** to increase the angle.

Characteristics of the unaided eye

The human eye is capable of focussing on objects close to the eye but how closely depends on a number of factors, including the age of the subject and the presence of optical defects in the eye. Try focussing on a small object as you gradually bring it closer to your eye. You will find that there is a location, called the **near point**, which is the closest you can bring the object while keeping it in sharp focus. Typically a young person (aged 20 or younger without optical defects) has a near point about 10 mm from the eye, whereas the best a normal sixty-year old subject can manage is to focus on objects 500 mm away. By convention, a comfortable close viewing distance for **near vision** is taken to be 250 mm (0.25 m). This distance is often called, inappropriately, the **least distance of distinct vision** denoted here by the symbol d_V. At the other extreme, an optical system can be arranged so that images are formed at infinity. In that case we have **relaxed** or **far vision.**

The finest detail that can be seen by the unaided eye, its **resolution**, can be calculated as follows. The minimum angle subtended by two points which can still be resolved is determined by diffraction at the pupil. This angle is about 3×10^{-4} radian. The closest distance for placing the object is about 0.25 m. Therefore, the minimum distance between two points which can still be resolved is about $3 \times 10^{-4} \times 0.25\text{ m} = 0.075\text{ mm}$.

Visual instruments

The best detail in a small object that you can see with the naked eye is obtained by putting the object at your near point. For even more detail you need to use a visual instrument which makes the object look bigger (figure 8.2). The effectiveness of a visual instrument is described by its **angular magnification**, M, defined as the ratio:

$$M = \frac{\text{angle that image subtends at eye looking through instrument}}{\text{angle that object subtends at unaided eye under the best possible conditions}}$$

$$= \frac{\beta}{\alpha}. \qquad \text{... (8.2)}$$

For far-away objects, where a telescope is used, the denominator (α) is just the angle that the object subtends at the unaided eye. For very small objects where a magnifier or a microscope is used, the denominator is found by calculating the angle α when the object is at the near point. Clearly, the magnification you get depends on how good your eyesight is. To get a nominal value for the magnification which does not depend on individual differences, the lens designer's value of magnification is obtained, by convention, by supposing that the value of α is the angle that the object subtends when it is at the standard distance of 0.25 m away from the eye.

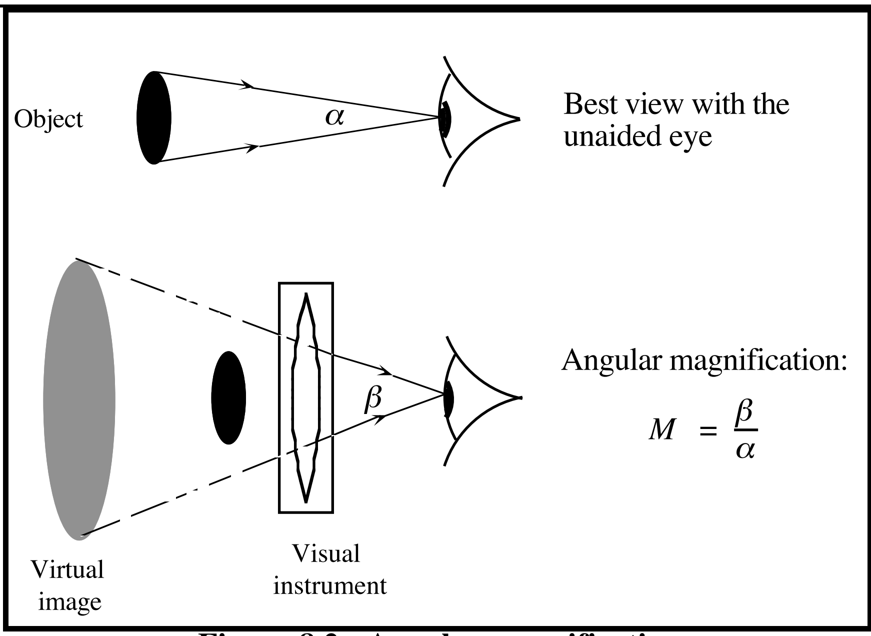


Figure 8.2. Angular magnification

To calculate angular magnification you compare the angular size of the image (β) with the best possible angular size of the object (α).

Although angular magnification is described in terms of the virtual image formed by the lens, you should remember that what you see is determined by the real image on the retina (figure 8.3). Increasing the angular size (β) of the virtual image produces a bigger final real image.

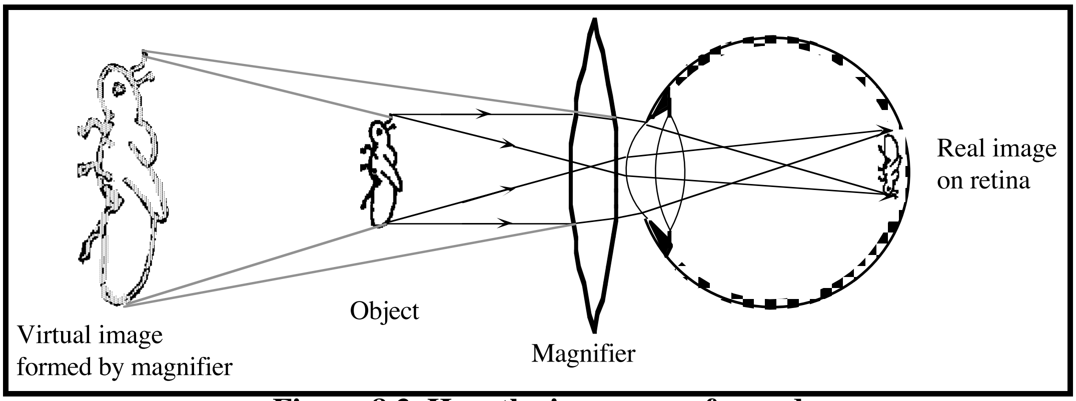


Figure 8.3 How the images are formed

8-3 SIMPLE MAGNIFIERS AND EYEPIECES

A single converging lens can be used as a visual **magnifier** which functions either on its own or as the **eyepiece** for a more complex instrument such as a microscope. The magnification is achieved by making the object seem to be closer to the eye than the near point. The lens produces a virtual image, which is larger than the object and therefore subtends a bigger angle at the eye. This virtual image can be located anywhere between the near point and infinity. The magnification actually achieved will depend on where the image and the eye are placed relative to the lens. In order to be able to compare the magnifications which can be achieved with different lenses we could just quote the focal length, but that does not give any immediate impression of the magnification. A more meaningful measure is the angular magnification achieved for some standard arrangement of image (or object), the lens and the eye. For the purpose of this calculation the eye is always placed as close as possible to the lens. Angles subtended by objects and images at the eye are then near enough to being the same as the angles subtended at the lens (figures 8.4 and 8.5).

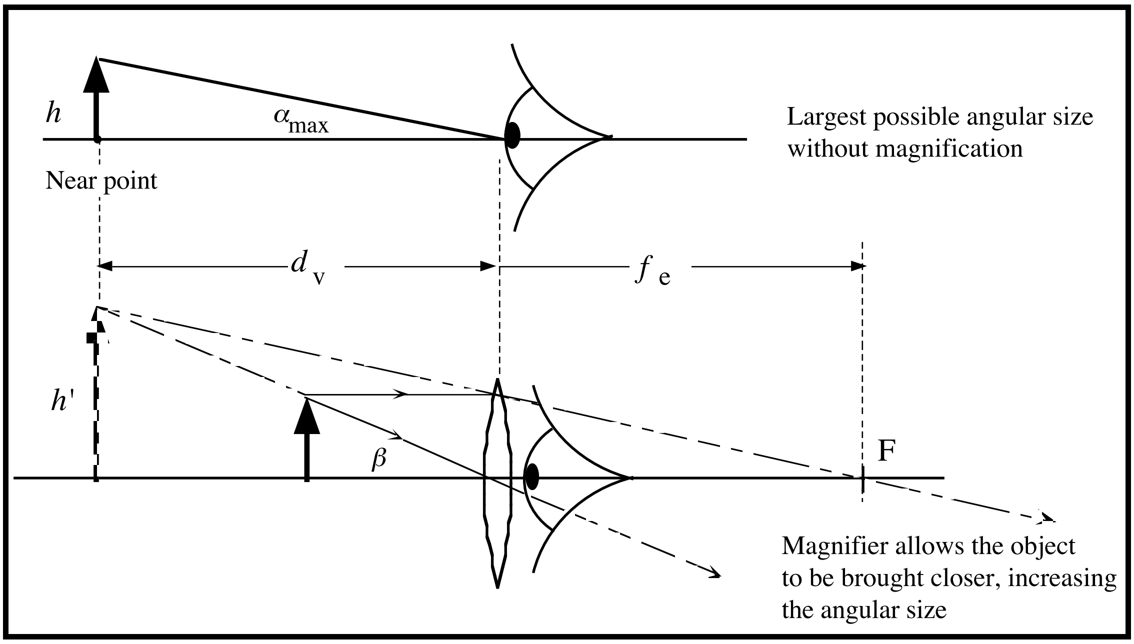


Figure 8.4. Simple magnifier or eyepiece used for near vision

For the greatest magnification the image is formed at the near point and the eye is placed close to the magnifier.

To work out the angular magnification, we compare the angular size of the image (β) with the angle (α_{max}) that the object would subtend at the eye (or lens) if it were put at the near point (distance d_v). It is fairly easy to work out (using the paraxial approximation, $\tan\alpha \approx \alpha$ etc.) that in this case the angular magnification is given by the formula:

$$M_e = 1 + \frac{d_v}{f_e} \qquad \text{... (8.3)}$$

where f_e is the focal length of the lens. (The subscript e stands for 'eyepiece'.)

Another standard way of arranging things is to have the image at infinity (figure 8.5). In that case the eye is said to be **relaxed**.

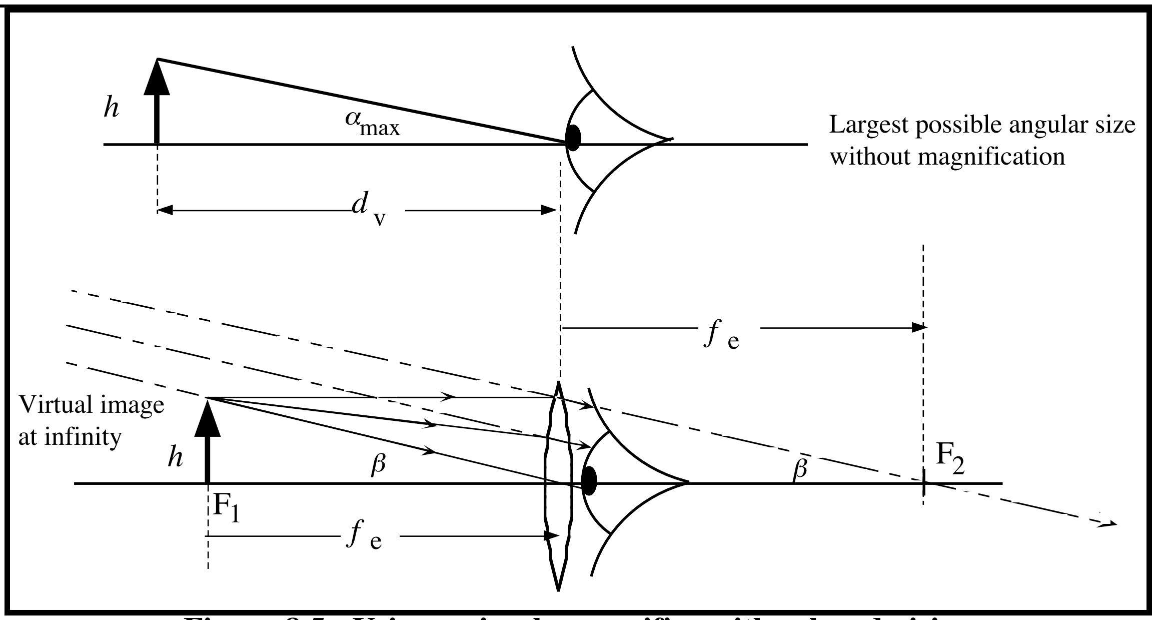


Figure 8.5. Using a simple magnifier with relaxed vision

For relaxed vision the object is at the focal point of the lens, so rays from a particular point on the object are parallel after refraction. In this example, rays from the bottom of the object enter the eye parallel to the principal axis and all the rays from the top of the object enter the eye at an angle β. Hence the image subtends an angle β at the eye. With the usual paraxial approximation,

$$\beta \approx \frac{h}{f_e} . \qquad \text{... (8.4)}$$

The object, if placed at distance d_v from the unaided eye, would subtend an angle α:

$$\alpha_{max} \approx \frac{h}{d_v} ; \qquad \text{... (8.5)}$$

so

$$M_e = \frac{\beta}{\alpha_{max}} \approx \frac{d_v}{f_e} . \qquad \text{... (8.6)}$$

For reasonably large magnifications, this formula (8.6) is not much different from the one quoted earlier (8.3) for the case with the image at distance d_v, but it does make a difference for low-power magnifiers. Not surprisingly, you get the best possible magnification by forming the virtual image at the near point.

A magnifier is usually described by its angular magnification rather than by its focal length. For example, a magnifier using a lens with a focal length of 25 mm would be described as a "10× magnifier".

8-4 TELESCOPES

A telescope, in its basic form, consists of two components, an objective and an eyepiece. Its purpose is to increase the apparent size or separation of distant objects.

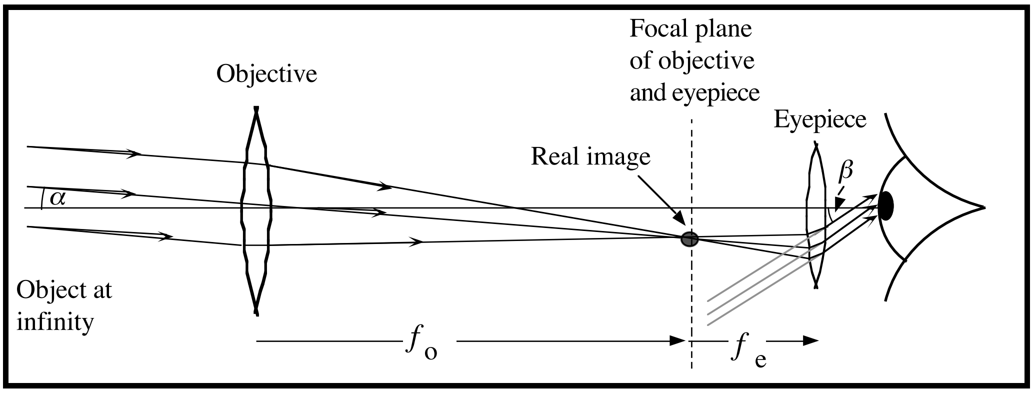


Figure 8.6. Keplerian (astronomical) telescope

In the astronomical telescope the objective produces a real image, which is then viewed with the eyepiece. The telescope produces an inverted image, but that is no problem when one is looking at astronomical objects. However, the same lens arrangement is used in prism binoculars in which the prisms restore the image to an upright position. (See figure 7.1 in chapter L7.)

Alternatively a diverging lens may be used as an eyepiece, as in the Galilean telescope.

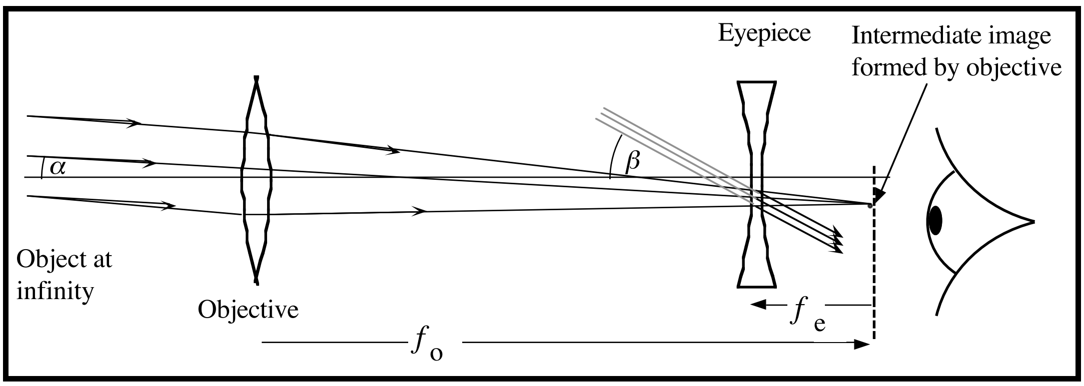


Figure 8.7. Galilean or terrestrial telescope

In both types of telescope the angular magnification is given by

$$M = \frac{f_o}{f_e} .$$

8-4 MICROSCOPES

The maximum useful magnification obtainable with a simple magnifier is about 20 times. For greater magnification we use a **compound microscope** (figure 8.8). In its basic form it consists of an **objective** and an **eyepiece** mounted in a tube. The magnification of a microscope can be worked out in terms of the focal lengths of the objective and eyepiece and the **optical tube length**, which is defined as the distance between the second focal point of the objective and the first focal point of the eyepiece. In a high power microscope the optical tube length is much larger than either focal length so it is roughly equal to the actual separation between the objective and the eyepiece. (The single lenses in figure 8.8 may in fact be optical components each made up of several lenses.)

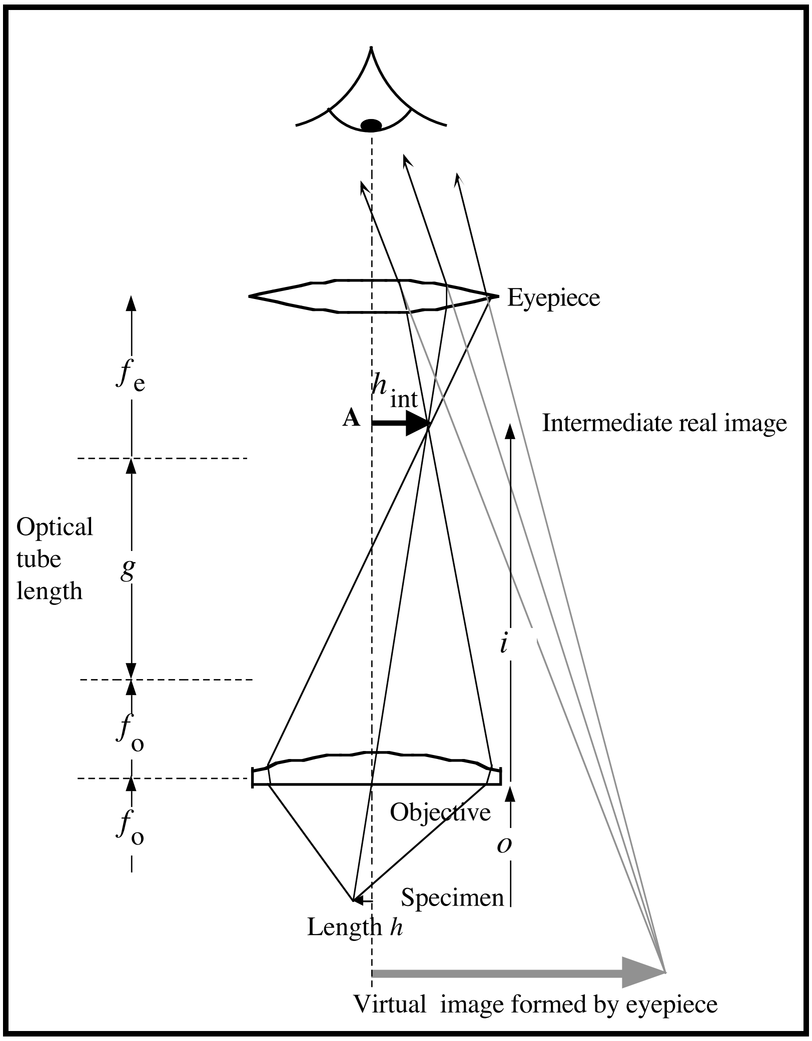


Figure 8.8. Compound microscope
Not to scale.

The specimen is placed just below the first focal point of the objective. Since the object distance is not much more than the focal length, the objective forms a much-enlarged real image, the intermediate image, at A. The eyepiece is used as a magnifier to look at the intermediate image . If that image is in the focal plane of the eyepiece, the virtual image seen by the eye will be at infinity.

The angular magnification is calculated as follows.

(i) The lateral (linear) magnification of the intermediate image can be written as the ratio of image distance to object distance, which in this case gives a magnification of

$$\frac{h_{\text{int}}}{h} = \frac{i}{o} \approx \frac{g + f_{\text{o}}}{f_{\text{o}}} .$$

Since the optical tube length g is usually very much greater than the focal length of the objective, the lateral magnification produced by the objective is

$$|m_{\text{o}}| = \frac{h_{\text{int}}}{h} \approx \frac{g}{f_{\text{o}}} . \qquad \text{... (8.7)}$$

(ii) The eyepiece acts as a magnifier so, with relaxed vision, the angle subtended by the final virtual image is

$$\beta \approx \frac{h_{\text{int}}}{f_{\text{e}}} \approx \frac{gh}{f_{\text{o}} f_{\text{e}}} .$$

(iii) Now the original object, when placed at distance d_{v} from the unaided eye, would subtend an angle

$$\alpha \approx \frac{h}{d_{\text{v}}} .$$

So the total angular magnification is

$$M = \frac{\beta}{\alpha} \approx \frac{g\, d_{\text{v}}}{f_{\text{o}} f_{\text{e}}} . \qquad \text{... (8.8)}$$

This result is just the same as saying that the total magnification is the product of the linear magnification of the objective and the angular magnification of the eyepiece:

$$M = |m_{\text{o}}|\, M_{\text{e}} \qquad \text{... (8.9)}$$

with $$|m_{\text{o}}| \approx \frac{g}{f_{\text{o}}}$$

and $$M_{\text{e}} \approx \frac{d_{\text{v}}}{f_{\text{e}}} .$$

Some typical values

Magnification of eyepieces:	up to 20×.
Optical tube length g (standard value):	160 mm.
Focal length of objectives, low power:	50 to 100 mm;
medium power:	8 to 50 mm;
high power:	4 mm;
very high power:	2 mm.

From these values, the maximum magnification of a microscope is about 1600. The maximum *useful* magnification is limited by diffraction to about 200 to 400.

Image brightness and numerical aperture

The **brightness** of the image depends on the amount of light from the specimen which enters the microscope. As we discussed in chapter L7, the brightness is determined by the entrance pupil of the objective. In microscopy another convenient way of specifying the amount of light collected is to quote the value of the angle u (figure 8.9) which describes the cone of rays collected by the objective.

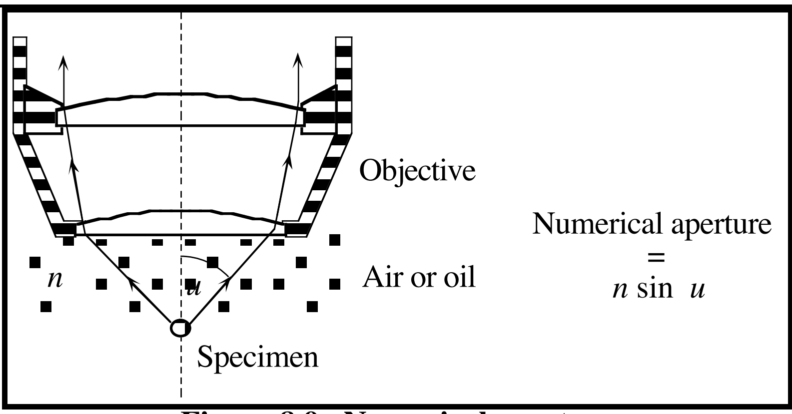


Figure 8.9. Numerical aperture
Numerical aperture depends on the angular size ($2u$) of the cone of light collected by the objective and the refractive index (n) of the material between the specimen and the lens.

However, it is not the angle u alone which matters, because the refractive index (n) of the medium between the specimen and the objective affects the refraction of the rays as they enter the objective (and hence also, the entrance pupil). The parameter which matters is called **numerical aperture** (N.A.) which is defined as $n \sin u$. It turns out that numerical aperture also determines the resolving power of an objective. Hence, microscope objectives are commonly specified in terms of their lateral magnification and numerical aperture.

Resolving power

As discussed in chapter L5, the amount of detail that can be seen in an image is limited by diffraction. Simple estimates of the limit can be made using the Rayleigh criterion. The limitations imposed by diffraction effects in a microscope which is completely free of aberrations can be described by a quantity called the **resolving power** of the microscope. Resolving power is defined as the minimum value of the distance between two points in the specimen which can just be resolved.* Its value is given by the formula, based on the Rayleigh criterion,

$$R = \frac{0.6\,\lambda}{n \sin u} = \frac{0.6\,\lambda}{\text{N.A.}} \qquad \text{... (8.10)}$$

where N.A. is the numerical aperture of the objective.

In both microscopes and telescopes, the **maximum useful magnification** is obtained when the angular separation of these two points, viewed through the microscope, is equal to the resolving power of the eye. For visible light, using a typical wavelength of 500 nm, the maximum useful magnification turns out to be about 250 times the numerical aperture.

In air the maximum practical value of the numerical aperture is about 0.85, but higher values can be obtained by filling the space between the objective and the specimen with oil. With this technique, called **oil immersion microscopy**, values of numerical aperture up to about 1.4 can be obtained. Using an oil-immersion objective can increase the maximum useful magnification to about 400×.

Field of view

The **field of view** is the region of a specimen which can be seen at any one time. It is determined by those rays from the specimen which can go through the microscope to enter the eye. Often the limiting feature is the diameter of the eyepiece. For instance the objective may produce a large intermediate image, but the rays forming the extreme points of that image may miss the eyepiece so that they will not enter the eye. In that case the observable intermediate image is about the same size

* The resolving power of a telescope is usually defined in terms of the *angle*, rather than the distance, subtended by object points at the objective. Similarly the resolving power of the eye is given as an angle. Whether distance or angle is meant can usually be determined from the context or from the unit used.

as the eyepiece, so the size of the field of view is given approximately by the diameter of the eyepiece divided by the magnification of the objective.

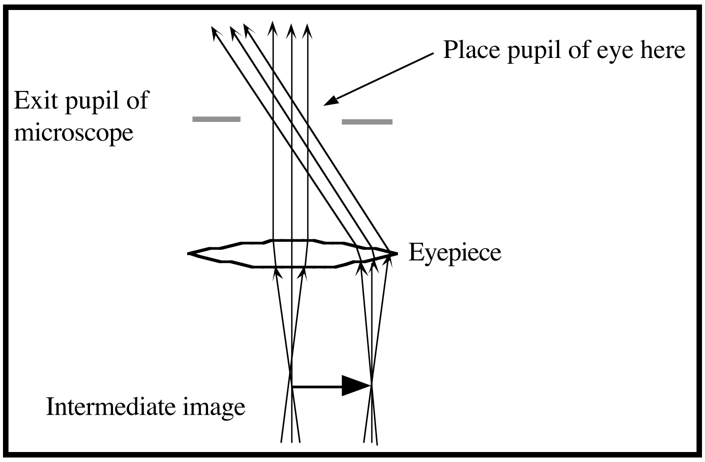


Figure 8.10. Field of view limited by the eyepiece

The field of view also depends on the position of the eye. For best viewing the pupil of the eye should coincide with the exit pupil of the microscope, since that is where the beam of light from the microscope is at its narrowest. For a microscope the exit pupil is the virtual image of the objective's aperture as seen through the eyepiece. It is a real image on the side of the eyepiece near the eye.

8-6 ILLUMINATION OF MICROSCOPE SPECIMENS

Most specimens to be viewed with a microscope are not self-luminous and so must be illuminated. For low magnification ambient lighting may be quite sufficient. At high magnifications the observed objects are small so they reflect only a small amount of light. To make such objects easily visible the intensity of light falling on them must be increased using special illuminating systems called **condensers**. Which kind of condenser is used depends on how the specimen is to be examined, which can be one of three ways:

(a) by transmitted light,
(b) by scattered light,
(c) by reflected light.

For optimum performance the condenser should be such that light from each point of the specimen fills the aperture of the objective.

Illumination by transmitted light

For specimens viewed in transmitted light, the condenser should supply a cone of light such that all the rays in the cone can enter the objective. This means that the angle of convergence of the illuminating light onto the specimen (u_c) should be equal to the acceptance angle (u_o) of the objective (figure 8.11).

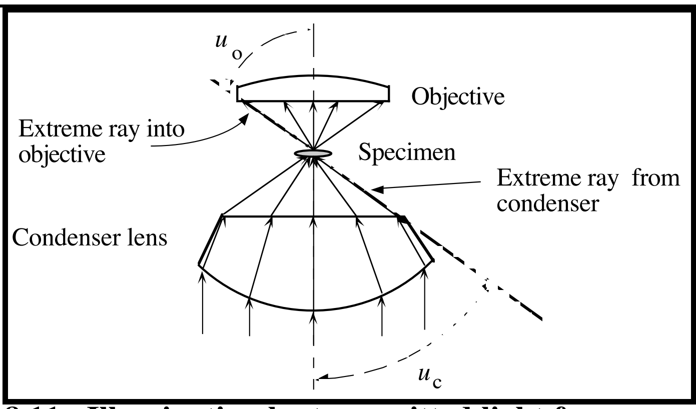


Figure 8.11. Illumination by transmitted light from a condenser

For low power objectives, (i.e. less than 10×) good illumination can also be made with a concave mirror reflecting the light from a convenient source into the microscope.

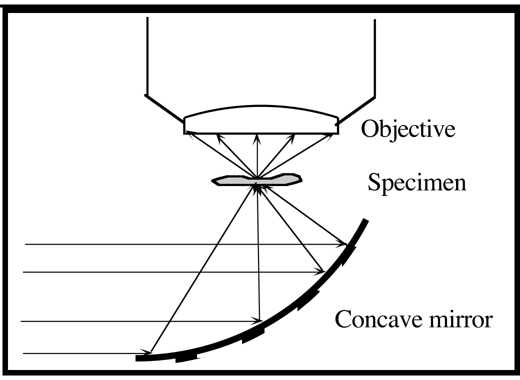


Figure 8.12. Illumination by transmitted light using a mirror

Dark field illumination

Sometimes objects can be more easily seen against a dark background. For example airborne dust becomes visible when viewed against a black cloth, by light scattered sideways out of a strong beam of sunlight. In microscopy this technique is called **dark field illumination** (figure 8.13). The condenser is equipped with a circular opaque stop so that none of the illuminating beam (shaded in the diagram) can enter the objective directly. The image is formed entirely by light scattered from the specimen, and none of the direct illuminating beam can be seen through the microscope.

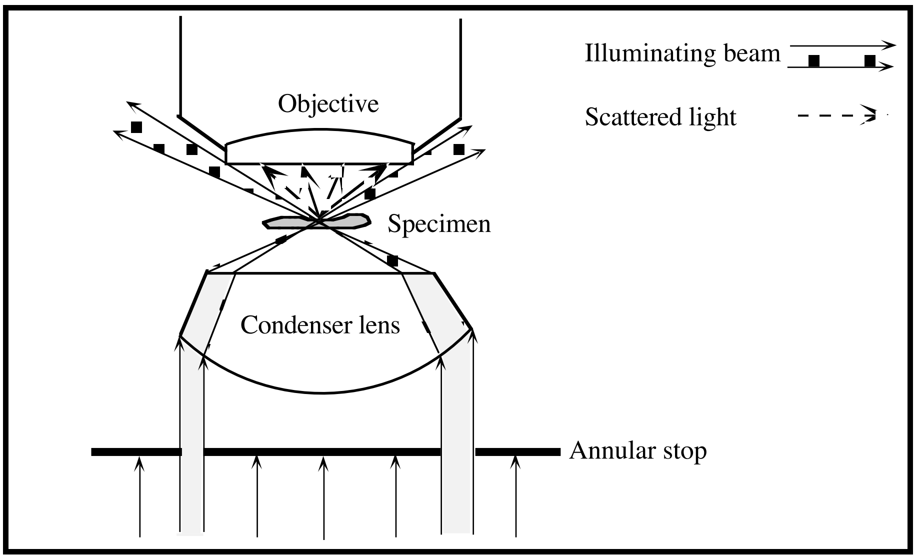


Figure 8.13. Dark field illumination
The illuminating beam misses the objective lens.

Illumination by reflected light

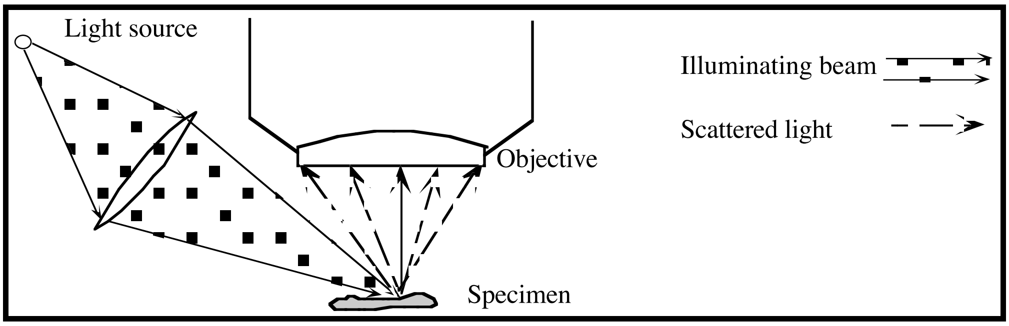


Figure 8.14. Illumination by reflected light

To view specimens in reflected light, illumination from above must be provided. For low power microscopes oblique top lighting may be sufficient (figure 8.14). For higher powers, more complicated systems, such as that illustrated in figure 8.15, are required.

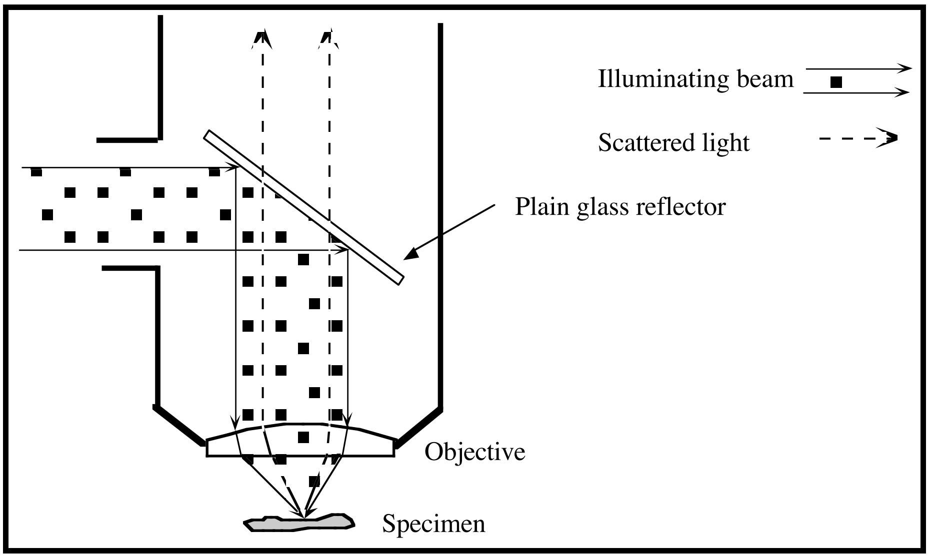


Figure 8.15. Illumination for high magnification with reflected light

8-6 INTERFERENCE MICROSCOPY

Normally when a uniform beam of light passes through a transparent material it emerges with equal intensity across the whole beam. That is true even if the refractive index has different values throughout the material or if the thickness of the specimen varies. However, different parts of the same beam of coherent light which travel through different optical path lengths will have different phases when they emerge from the specimen. Interference is a phenomenon which depends on phase relationships, so we can use interference, in what is called an **interference microscope** to "see" refractive index variations in transparent specimens of uniform thickness. See chapter L4.

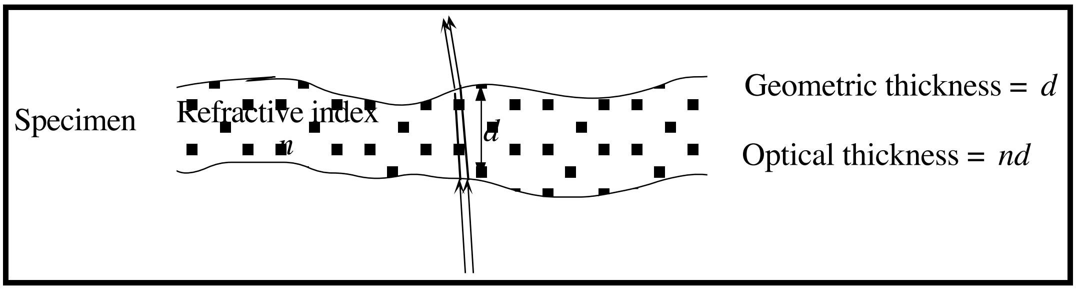


Figure 8.16. Optical path length

In order to get coherence, a single beam of light is split into two parts, one of which goes through the specimen while the other does not. The phase of the light that has passed through a particular part of the specimen will be different from the phase of light that has taken the other path. When the two parts of the beam are recombined interference will occur. Some regions of the specimen will appear bright, others dark.

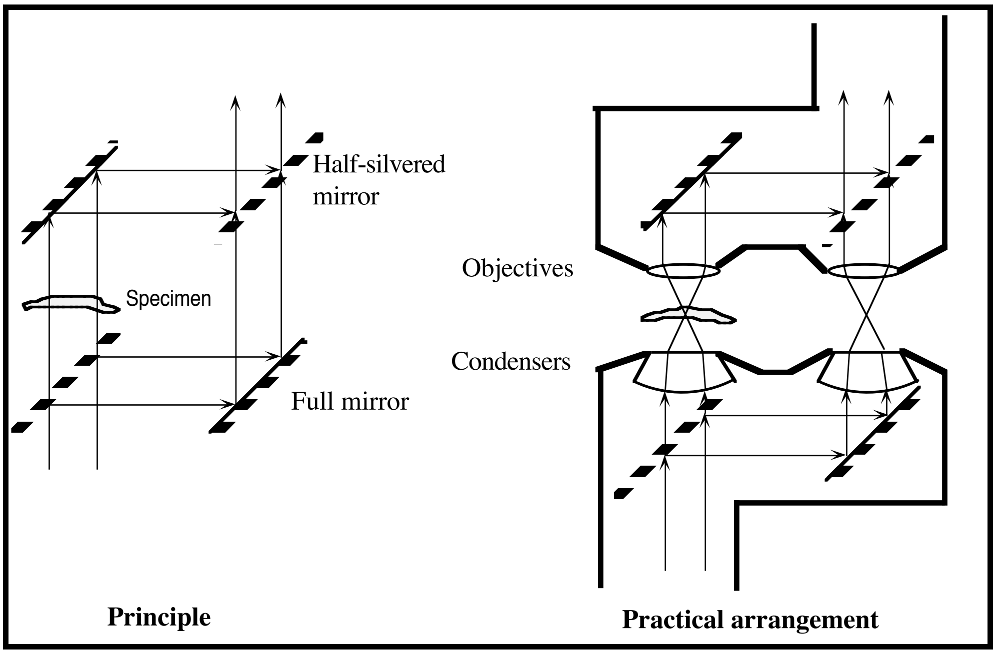


Figure 8.17. Interference microscopy

QUESTIONS

Q8.2 What is the angular magnification of a small magnifier, focal length 5 cm, used with relaxed vision?

Q8.3 A magnifier with a focal length of 25 mm is held close to the eye to examine a small object. As shown in §8-2, if the object is at the focal point of the lens, the image will be at infinity.

(a) Calculate the angular magnification in this case.

Now if the object is just inside the focal length, you will still get an image on the far side of the lens but it will be closer.

(b) Use the lens formula to calculate where the object must be for the image to be at 0.25 m.

(c) Calculate the ratio, $\dfrac{\text{image height}}{\text{object height}}$, and hence the angular magnification in this situation.

Q8.4 (a) A microscope is made up of an objective with a focal length of 16 mm, numerical aperture 0.25 and a 10× eyepiece. The optical tube length has the standard value, 160 mm. Calculate the angular magnification.

(b) If the eyepiece restricts the diameter of the intermediate image to 15 mm, how big is the field of view of this microscope?

Q8.5 Suppose that the objective in the previous question has a resolving power given by

$$R = \frac{0.6\ \lambda}{\text{N.A.}} \quad .$$

What is the finest detail that we could observe on the specimen? What angle does this detail subtend at the eye when it is viewed through the microscope? Does the resolution of the eye or the resolution of the objective determine the finest detail observable with this instrument?

Discussion questions

Q8.6 What is the angle subtended by a TV screen for comfortable viewing? Given that the TV image is made up of 625 lines, how does the resolution of the eye compare with the angular separation of the lines?

Q8.8 Why is the magnification of a simple magnifier defined in terms of angles rather than the actual sizes of image and object?

Q8.9 Spectacles are not used to magnify objects. What are they used for? Discuss.

Q8.10 Photographers alter the apertures or f-numbers of their camera lenses. Why? What is the relation between aperture setting and exposure time?

Q8.11 The real image formed on the retina of the eye is inverted. Why don't we see things upside down?

INDEX