1

Light and Colour

- **What is colour?**
- **Why do hot objects become red or white hot?**
- **How does laser light differ from light from the sun?**

Colour may be considered to be the subjective appearance of light as detected by the eye. Broadly speaking, light leaves the generating source, possibly interacts with matter in the course of passage, and then enters the eye. This book is principally about this interaction of light and materials and how it influences our perception of the colour of the light so produced. This perception of colour, the result of an eye–brain combination, is summarised by the term *vision*. Essentially, vision forms a subject area outside of the central area of interest here. However, it should not be totally neglected, and in this introductory chapter, is briefly described.

The chapter is divided into three sections: the first describes the relationship between light and colour (Sections 1.1–1.4); the second concerns light production (Sections 1.5–1.8), while vision and colour perception end the chapter (Sections 1.9–1.13).

1.1 Light and Colour

1.1.1 Light rays

In elementary optics, light can usefully be considered to consist of light *rays*. These can be thought of as extremely fine beams that travel in a series of straight lines from the light source and thence, ultimately, to the eye. The majority of simple optical instruments can be constructed within the framework of this idea. However, the ray concept breaks down when the behaviour of light is critically tested, and the performance of optical instruments, as apart from their construction, cannot be explained in

Colour and the Optical Properties of Materials, Third Edition. Richard J. D. Tilley.

Companion website: www.wiley.com/go/tilley/colour&opticalproperties3e

terms of light rays. Moreover, colour is not conveniently defined in this way. For this, more complex ideas are needed.

1.1.2 Light waves

Early theories concerning the nature of light were divided over whether it was composed of small particles (or *corpuscles*) or whether it was made up of waves. Newton, in his book *Optics* (1704), endorsed the particle concept, which was supported on philosophical grounds by Descartes. On the other hand, Huygens, a contemporary, thought that light was wavelike, a point of view also supported by Hooke. Young provided strong evidence for the wave theory of light by demonstrating the interference of light beams passing through two adjacent slits (1803). Shortly afterwards, Fresnell and Arago explained the polarisation of light in terms of transverse light waves. However, none of these explanations was able to completely refute the particle hypothesis. Nevertheless, the wave *versus* particle theories differed in one fundamental aspect that could be tested. When light enters water it is refracted, that is, the trajectory of the light appears to alter. In terms of corpuscles, this implied a speeding up of the light in water relative to air. The wave theory demanded that the light should move more slowly in water than air. The experiments were complicated by the enormous speed of light, which was known to be about 3×10^8 m s^{-1}, and it was not until April 1850 that Foucault first proved that light moved slower in water than air, and seemingly killed the corpuscular theory then and there. Confirmation of the result by Fizeau a few weeks later removed all doubt.

Over the years, the wave theory became entrenched and reached its peak when Maxwell published his theory of electromagnetic radiation in 1865, which showed that light behaved as an electromagnetic wave, with both electric and magnetic components. [Maxwell's (original) equations that describe electromagnetic waves are fairly indigestible. The compact form now universally known as Maxwell's equations was, in fact, formulated by O. Heaviside, some years later.] Maxwell's theory, which implied the existence of an extensive electromagnetic wave spectrum was confirmed experimentally by Hertz in 1888, whose experiments led to the production and detection of a new long wavelength part of the electromagnetic spectrum, originally described as Hertzian waves, but since called *radio waves*.

It is now known that the spectrum of electromagnetic waves spans an enormous range of wavelengths, from less than 1 pm to greater than 1 km. Historically, the various regions have been given different names, although the boundaries between each region are not sharply defined but grade into one another, with gamma rays at the short wavelength end of the scale and radio waves at the long wavelength end. In these terms, light can thus be defined as that (small) part of the electromagnetic spectrum that is detected by our eyes, called the *visible spectrum*, existing within the approximate range 400–700 nm (Figure 1.1).

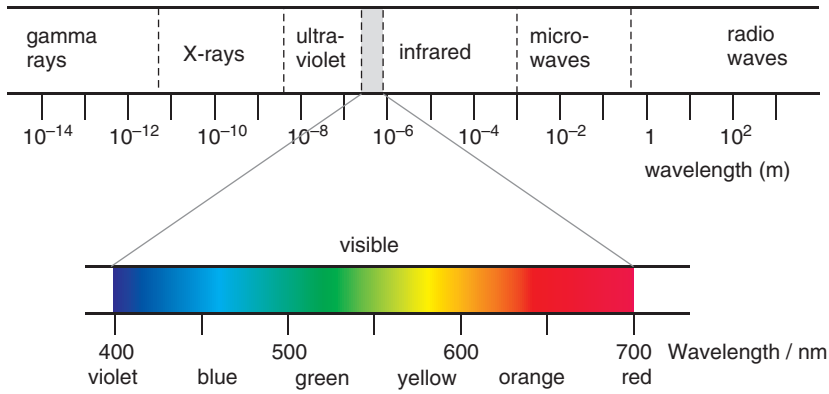


Figure 1.1 ***The electromagnetic spectrum***

The problem for the wave theory was that waves had to exist in something, and the 'something' was hard to pin down. It became called the *luminiferous aether* (often shortened to the *ether*), which had the remarkable properties of pervading all space, being of very small (or even zero) density and having extremely high rigidity. Attempts to measure the velocity of the Earth relative to the luminiferous aether, the so-called aether (ether) drift, by Michelson and Morley, before the end of the nineteenth century, proved negative. Partly in view of this the corpuscular theory of light was revived early in the twentieth century.

1.1.3 Photons

The problem for the electromagnetic wave approach centred on the emission spectrum of a hot object, which can be approximated by the radiation emitted by a *black body*, an idealised object that absorbs and emits all wavelengths perfectly. A reasonable approximation to a source of black-body radiation would be a small pinhole in the wall of a hot furnace. If the irradiance of the radiation issuing from the pinhole is measured as a function of wavelength, a characteristic curve is obtained called a black-body spectrum (Figure 1.2). The shape of the curve is dependent only on the temperature of the body and the maximum in the curve moves to shorter wavelengths as the temperature of the black body increases. The spectrum emitted by the sun is similar to that for a black body at 5700 °C and that from a red hot object is similar to the curve for a black body at 700 °C.

Despite many attempts, the form of the black-body spectrum could not be explained by the classical wave-theory of electromagnetic radiation. The successful theoretical description of this curve by Planck in 1901, which resulted in the *Planck law of black-body radiation,* or *Planck's radiation law*, signalled the start of the quantum theory. The revolutionary concept that Planck employed in the derivation of these equations to successfully reproduce the black-body curve was that the energy absorbed or given out by the atoms and molecules (the 'oscillators' in Planck's time) in the black

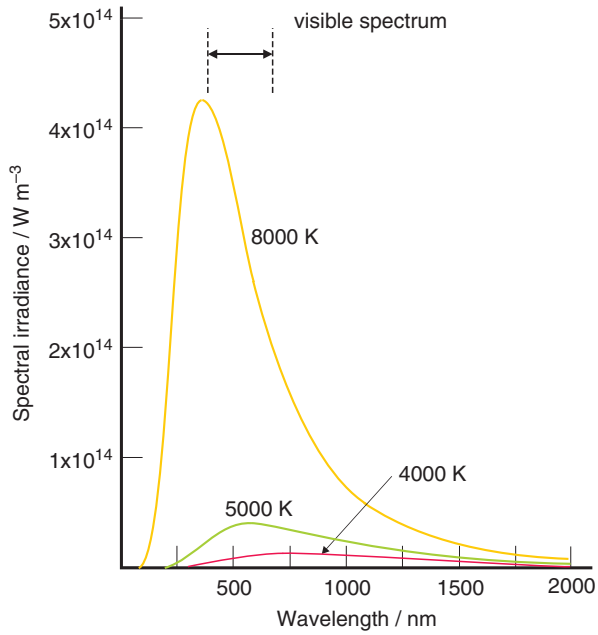


Figure 1.2 ***The radiation emitted from a black body as a function of wavelength***

body could not take any value from a continuous spread of energies, but had to be delivered only in packets or *quanta* q_0, $2q_0$, $3q_0$, and so on. The relationship between the energy of a quantum, E, and the frequency of the radiation, ν, was given by:

$$E = h\nu \tag{1.1}$$

where h is Planck's constant.

The quantisation of radiation proposed by Planck in the derivation of the radiation law was not seized upon instantly. After a lapse of some years, it was exploited by Albert Einstein in his explanation of the photoelectric effect in 1905. Since 1895 it had been observed that when ultraviolet light was used to illuminate the surfaces of certain metals, negative particles, later identified as electrons, were emitted. The details of the experimental results were completely at odds with the wave theory. The electrons, called *photoelectrons*, were only observed if the frequency of the radiation exceeded a certain minimum value, which varied from one material to another. The kinetic energy of the photoelectrons was linearly proportional to the frequency of the illumination. The number of photoelectrons emitted increased as the intensity[1] of the light increased, but their energy remained constant for any particular light source. Very dim illumination still produced small numbers of photoelectrons with the appropriate energy.

Einstein proposed that the quantisation of radiation contained in Planck's formula for black-body absorption and emission of energy could be applied to the radiation itself and not to just the energy exchange. That is to say, light was to be regarded not as a wave but as a hail of bullet-like objects (which are now called *photons*), each of which had an energy $h\nu$ (Eq. (1.1)). The corpuscle theory had re-emerged.

The photoelectric effect was then explained in the following way. Each photon delivered the same amount of energy. If this was sufficiently large, an electron could be ejected from the surface. The energy of each photon is proportional to the frequency of the illumination, so that when the frequency passes a certain threshold, a photoelectron will be ejected, but not before that point was reached. Thereafter, increasing the frequency of the illumination allowed the excess energy to be displayed as an increase in kinetic energy. The kinetic energy of the photoelectrons ejected from a metal under this hail of photons could then be written as:

$$\tfrac{1}{2} m v^2 = E - \phi$$

where ϕ is known as the *work function* of the metal and is simply the energy required to liberate the electron from the metal surface. The intensity of the light indicated the number of photons arriving at the surface so that the *number* of photoelectrons emitted is a function of *intensity*, but the *energy* of these electrons is a function of the *frequency* of the radiation. Einstein thus rescued the wave theory from the dilemma of the luminiferous aether and then seemingly wrecked the self-same theory via his explanation of the photoelectric effect.

1.1.4 Energy levels

The Planck/Einstein idea that light was composed of bullet-like photons means that the emitter must be jumping between two separated energy states, each of which can be assigned a precise energy. To a good approximation, these energy states, usually called *energy levels*, arise from the disposition of the

[1]The imprecise expression *intensity* has largely been replaced in the optical literature by well-defined terms such as irradiance (Appendix A). The term intensity is retained here and elsewhere in this book (in a qualitative way, designating the amount of light) because of the historical context.

electrons that surround atomic cores, modified by factors such as vibration, the presence of impurities and electric and magnetic fields. The energies so derived form a sort of ladder (with variable rung spacing), each separated from the next by an energy gap. At low temperatures, the material is imagined to occupy the lowest energy level, called the *ground state*. The absorption of a photon will then only take place if the photon energy corresponds exactly to the energy interval between ground state, E_0, to an energy level at a higher energy E_1. The relationship between the energy change, ΔE, and the frequency, ν, or the wavelength, λ, of the photon absorbed is:

$$E_1 - E_0 = \Delta E = h\,\nu = h\,c/\lambda$$

where h is the Planck constant and c is the speed of light. In the reverse process, when the material passes directly from E_1 to E_0, it releases a photon identical to that of the absorbed radiation. In both cases, if the frequency associated with the energy change ΔE lies in the visible, colour is perceived.

1.1.5 Waves and particles

This conjunction of the particle and wave descriptions, called wave-particle duality, is evident in the fact that the frequency ν of a photon is itself a wave-like property. In fact, all particles are now believed to exhibit wave-like properties. The momentum p, of a particle (such as an electron, say), in a vacuum, is given by

$$p = (E^2 - m^2c^4)^{1/2}/c$$

where E is the energy, m the particle mass, and c the speed of light. For a photon, $m = 0$, so that:

$$p = E/c$$

The wavelength of the particle is:

$$\lambda = h/p = h\,c/(E^2 - m^2c^4)^{1/2}$$

For a photon, $m = 0$, so that:

$$\lambda = hc/E$$

$$E = hc/\lambda$$

The velocity of a particle is:

$$v = pc^2/E = c\,[1 - (m^2c^4/E^2)]^{1/2}$$

For a photon, $m = 0$, so that:

$$v = c$$

For many purposes, the wave and particle aspects of light can be used interchangeably, as dictated by experiment. At present all experiments show that light and its interaction with matter (i.e. atoms and molecules) is most easily described in terms of photons. The wave aspect of light expresses the fact that the photons do not obey deterministic laws of motion but laws of probability. The waves associated with light photons are one way of describing these probabilities. At its simplest level, the statistical behaviour of a large number of photons is then represented very well by, at the largest scale, light rays, and in more detail by an electromagnetic wave. That is to say, photons are the components of a light beam whilst rays or waves are mathematical descriptions of a beam of light. This idea is

still too simplistic to describe many quantum aspects of light. For example, it has recently been shown that when the slits in Young's classical experiment proving the wave nature of light are replaced by *nanoslits,* the periodic fringes that substantiate the wave nature of the radiation change from λ/2 to less than λ/4 – a phenomenon called *extraordinary Young's interference*. Nevertheless, for the purposes in this book, the division of colour-related phenomena into light wave or photon based will largely be used, except when dealing with the nanoscale, where departures might be expected.

1.1.6 Colour

Colour is a response of the eye–brain combination to light falling on the eye (Sections 1.9–1.12). White is perceived when the light falling on the eye contains all the wavelengths that make up the visible spectrum, roughly in the proportions that these occur in bright sunlight. Colour is perceived when some of these wavelengths are missing.

How this comes about is the subject of the following chapters. Here a brief overview is given. The most commonplace of these are the dyes, paints, and pigments that make up the large majority of colour that surrounds us, including the coats of animals, the feathers of birds, the colours of flowers, and so on. These colours are generated when white light falls on a surface containing these colour-producing materials. Some wavelengths are absorbed, leaving the remainder to be reflected back to the eye. If none return the object is 'black'. If the blues and greens are absorbed the colour is perceived as red or orange.

Colour can arise, though, in many other ways. The most obvious of these is that the light source may not emit all the wavelengths, or not in the correct proportions. Most lasers emit only a narrow wavelength range. Lasers used for levels in construction work, or lasers used in scanning barcodes in supermarkets, emit a narrow range of red light. Displays of all types – TV screens, computer monitors, phone displays – all consist of small elements, each of which emits a single colour. When all emit together the separate colours are interpreted as white by the viewer. Mercury and sodium streetlights, still to be seen, emit a greenish-blue light or a yellow light, respectively.

Apart from this, some transparent objects, when illuminated by white light, are somehow modified by this interaction to show colours. This effect gives rise to rainbows, the 'fire' of diamonds, the colours of soap films and the blue of the sky (Chapters 2–5). Finally, a number of transparent materials can be ordered into multilayers or crystal-like arrays on the same scale as the wavelength of light. These structures give rise to iridescence and other *structural colours*, familiar as the colour of the gemstone opal and in the bright *iridescent colours* of many insects and bird feathers (Chapter 6).

1.2 Light Waves

The nineteenth century saw the apparent triumph of the theory that considered light to be an electromagnetic wave. Any part of the electromagnetic spectrum, including the region allocated to light, is regarded as a wave of wavelength λ with an electrical and magnetic component. The electric field is described as a vector[2] **E** perpendicular to the magnetic field vector, **B**, and they are in phase, so that a peak in the electric-field component coincides with a peak in the magnetic-field component. Moreover, these vectors lie in a plane perpendicular to the direction in which the wave is moving, described by the *velocity vector*, **v**, or the *propagation vector* or *wave vector*, **k**. Thus, **E** and **B** both lie perpendicular to the *direction of propagation*, so that light is regarded as a *transverse electromagnetic* (TEM) wave. The wave is a *progressive wave*, a *travelling wave,* or a *propagating wave*, all terms being used more or less interchangeably. The electric field vector, the magnetic field vector, and the velocity vector can be represented by the three (right-handed) Cartesian axes (Figure 1.3a).

[2]Vectors are given in bold throughout this book.

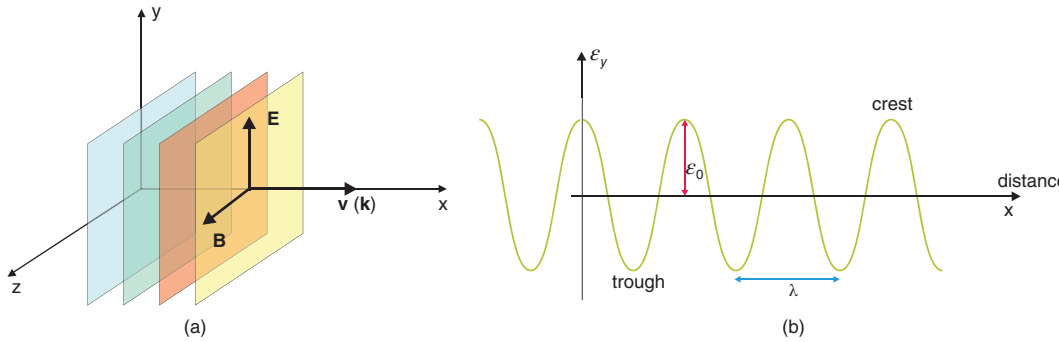


Figure 1.3 ***Light waves: (a) light as a transverse electromagnetic wave. The shaded planes represent the positions of peaks in the electric and magnetic fields; (b) instantaneous representation a light wave travelling along x. The curve shows the magnitude of the electric field vector, $\mathcal{E}_y$, as a function of position***

As far as the topics in this book are concerned, the electric field **E** can usually be considered in isolation. In this case, a one-dimensional continuous electromagnetic wave moving in the $+x$ direction can be conveniently represented by the equation:

$$\mathcal{E}_y = \mathcal{E}_0 \cos [(2\pi/\lambda)(x - vt)] \tag{1.2}$$

or

$$\mathcal{E}_y = \mathcal{E}_0 \cos [k(x - vt)]$$

Here $\mathcal{E}_y$ is the *magnitude* of the electric field vector at position x and time t and v is the speed at which any point on the wave travels. It is properly called the *phase speed* or *phase velocity*.[3] The speed of a *plane* electromagnetic wave in vacuum, denoted by c, is an important physical constant. The term $\mathcal{E}_0$ is the *amplitude* of the wave (the maximum value that the electric field vector takes) and is a constant. The term k $(2\pi/\lambda)$ is called the *wave number* or *propagation constant* and is also the magnitude of the propagation vector. (Other equivalent expressions for the wave equation are given in Appendix A.)

Taking t as fixed gives a snapshot of the wave at a single instant (Figure 1.3b). The spatial period of the wave, which is the distance over which the wave subsequently repeats itself, is called the *wavelength* λ. The peaks in the wave are referred to as *crests* and the valleys as *troughs*. The term $[(2\pi/\lambda)(x - vt)]$, the argument of the cosine function, is called the *phase* of the wave, represented by φ. The phase of the wave is usually quoted in radians, in the form $(m\ \pi/n)$, i.e. $3\pi/4$. Clearly the phase of the wave varies along its length and changes by 2π in one wavelength. The phase of a light wave cannot be determined. However, the phase difference between corresponding points on two different waves, say two equivalent crests, can be measured with considerable precision.

Taking x as fixed will show that the magnitude of the electric field vector, $\mathcal{E}_y$, will oscillate up and down between values of $\pm\mathcal{E}_0$. The temporal period of the wave, τ, which is the time over which the wave subsequently repeats itself, is more usually encountered as the reciprocal, $1/\tau$, equal to the temporal frequency ν, which is the number of waves that pass a point per second.

The speed of the wave, v, is related to the frequency ν, by the equation:

$$v = \lambda\nu$$

[3]Strictly speaking, we are discussing wave speed, which is a scalar quantity. Velocity is a vector quantity. However, it makes things simpler to brush over this distinction in the present case.

or for a plane wave in a vacuum by

$$c = \lambda\nu$$

A beam of light is said to be *monochromatic* when it is comprised of a very narrow range of wavelengths and *coherent* when all of the waves that make up the beam are completely *in phase*, that is, the crests and troughs of all the waves are in step and show a constant phase difference. The way in which the electric field vector is constrained describes the *polarisation* of the wave. If the electric field vector remains in one plane, the light is said to be *linearly* (or *plane*) *polarised*.

Normal light, such as that from the sun, say, is not emitted in a continuous stream, but in short bursts lasting about 10^{-8} seconds from a large number of closely spaced point sources. Within each burst from a single point source all of the light waves are in phase and linearly polarised. However, both the phase and polarisation change from burst to burst in a random fashion and are unrelated to those in the preceding burst. Moreover, each emitter acts independently of the others in this respect. This means that the phase and the polarisation of all the various light waves in the beam fluctuate continuously and at random within a fraction of a second. Normal light is thus described as being *incoherent* and *unpolarised*. Because of this, the interaction of daylight with objects can be interpreted (at least as a good approximation) without considering polarisation. Light from lasers is by and large coherent and polarised, and these aspects cannot then usually be ignored.

1.3 Light Waves and Colour

The wavelength range that an eye can distinguish varies from individual to individual. *Perception* of the different wavelengths is called *colour*. In general, it is assumed that the shortest wavelength of light that an average person can detect corresponds to the colour violet, with a wavelength near to 400 nm. Similarly, the longest wavelength of light registered by an average observer corresponds to the colour red, with a wavelength close to 700 nm. Between these two limits, the other wavelengths of the spectrum are associated in the colour sequence from red to orange, green, blue, indigo, and finally to violet (Table 1.1). The divisions between these colours are, of course, artificial, and each colour blends into its neighbours. It is known that the sensitivity of the eyes of animals is different than those of humans. Many insects, for example, can detect wavelengths shorter than humans but do not see so far into the red.

Table 1.1 ***The visible spectrum***

Colour	λ/nm	ν/Hz	ω/rads s^{-1}	Energy/J	Energy/eV	Wavenumber cm^{-1}
Infrared	750	4.00×10^{14}	2.51×10^{15}	2.65×10^{-19}	1.65	13 333
Deep red	700	4.28×10^{14}	2.69×10^{15}	2.84×10^{-19}	1.77	14 286
Orange-red	650	4.61×10^{14}	2.90×10^{15}	3.06×10^{-19}	1.91	15 385
Orange	600	5.00×10^{14}	3.14×10^{15}	3.31×10^{-19}	2.07	16 667
Yellow	580	5.17×10^{14}	3.25×10^{15}	3.43×10^{-19}	2.14	17 241
Yellow-green	550	5.45×10^{14}	3.42×10^{15}	3.61×10^{-19}	2.25	18 182
Green	525	5.71×10^{14}	3.59×10^{15}	3.78×10^{-19}	2.36	19 048
Blue-green	500	6.00×10^{14}	3.77×10^{15}	3.98×10^{-19}	2.48	20 000
Blue	450	6.66×10^{14}	4.19×10^{15}	4.42×10^{-19}	2.75	22 222
Violet	400	7.50×10^{14}	4.71×10^{15}	4.97×10^{-19}	3.10	25 000
Ultraviolet	350	8.57×10^{14}	5.38×10^{15}	5.68×10^{-19}	3.54	28 571

The amount of light that the eye records in any situation, which can *loosely* be called the brightness or intensity of the light, is not the amplitude of the wave but is the *irradiance*, I, which is proportional to the square of the amplitude:

$$I = K\,(\mathcal{E}_0)^2$$

where the value of the constant of proportionality, K, depends on the properties of the medium containing the wave.

Radiation consisting of wavelengths shorter than violet makes up the *ultraviolet* region of the spectrum. Ultraviolet A (UVA) is closest to the violet region, and is taken to have a wavelength range of 400–320 nm. This radiation is largely responsible for suntan. Ultraviolet B (UVB) with an approximate wavelength range of 320–280 nm is more damaging and causes sunburn and may result in skin disease. Ultraviolet *C* (UVC) has the approximate wavelength span of 100–280 nm. Ultraviolet radiation with shorter wavelengths is called the *far ultraviolet* (280–200 nm), *vacuum ultraviolet* (below 200 nm), and *extreme ultraviolet* (10–approximately 120 nm). Radiation with wavelengths longer than red are referred to as *infrared* radiation. Although not visible, the longer wavelengths of infrared radiation, called *thermal infrared*, are detectable as the feeling of warmth on the skin.

1.4 Interference

One of the advantages of the wave description of light is that the interactions between two or more beams are easily explained. If two light waves occupy the same region of space at the same time they add together, or *interfere* to form a product wave. This idea, called the *principle of superposition*, was stated by Young some two centuries ago, in 1802.

Interference can occur between light waves with different relative frequencies, amplitudes, and phases. However, for it to be *observed* the phase difference between the beams must remain constant. That is, the waves must be *coherent*. Many of the difficulties inherent in observing interference effects using normal light stem from the incoherent nature of the wave trains used and efforts must be made to ensure that the incoherence does not destroy any visible interference patterns that may be generated. The standard way of achieving this before the advent of lasers was to pass the light through a pinhole, which approximated the ensuing beam to that from a single point source, and then dividing this into two to give two reasonably coherent wave trains, that could then be made to overlap and show interference effects. This is, in fact, the method originally used by Young to demonstrate that light could be regarded as wave-like. The use of laser light, which approximates to light from a single point source, makes the observation of interference much simpler.

1.4.1 Two waves with the same wavelength

To illustrate the interference of two waves with the same wavelength but differing in phase, suppose, for simplicity, that two identical waves can be represented by:

$$\mathcal{E}_y = \mathcal{E}_0\,\cos\,[2\pi/\lambda\,(x)]$$

and

$$\mathcal{E}_y = \mathcal{E}_0\,\cos\,[2\pi/\lambda(x+\phi)]$$

where a comparison with Eq. (1.2) shows that the time progression is ignored (so that we are dealing with two standing waves). If the two waves are out of step by ϕ, the peak of one is shifted in the $-x$

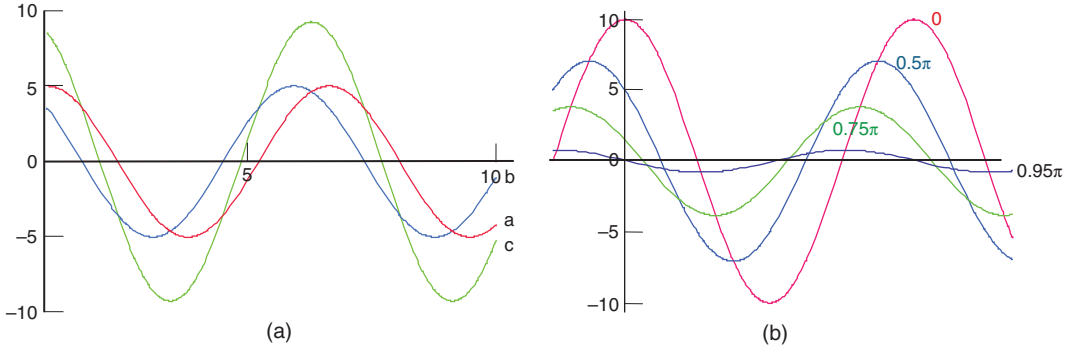


Figure 1.4 ***Interference of light waves: (a) the resultant wave (green) formed by adding two identical waves (red and blue) of amplitude 5 and phase differences of 0.25π; (b) the resultant waves formed by adding two waves of amplitude 5 and phase differences of 0, 0.5π, 0.75π, 0.95π***

direction by a distance ϕ compared to the other and the two waves add to give a new peak at $-\phi/2$ with a peak height of:

$$\mathcal{E}_y = 2\mathcal{E}_0 \cos [2\pi/\lambda(x + \phi/2)]$$

If the two waves are exactly in step, $\phi = 0$ and they will add to produce a resultant wave exactly in step with the two original waves, with twice the amplitude, called *constructive* interference. When there is a phase difference the total wave height drops and the peak of the resultant wave lies between the peaks of the original waves (Figure 1.4a). As the phase difference increases, the resultant amplitude will be less, due to *destructive* interference. If the waves are sufficiently out of step that the crests of one correspond with the troughs of the other, a phase difference of π radians, the resulting amplitude will be zero (Figure 1.4b).

1.4.2 Two waves with different wavelengths

The addition of waves with different wavelengths has a more complex effect. When these waves are added the wavelength of the new wave changes and the wave profile is modified by another wave-like function. For example, again ignoring the time progression, using the simplified equations:

$$\mathcal{E}_1 = \mathcal{E}_0 \cos [k_1 x]$$

and

$$\mathcal{E}_2 = \mathcal{E}_0 \cos [k_2 x]$$

where

$$k_1 = 2\pi/\lambda_1 \text{ and } k_2 = 2\pi/\lambda_2$$

The resultant wave is:

$$\mathcal{E} = 2\mathcal{E}_0 \cos [k_m x] \cos [k_w x]$$

where:

$$k_m = \tfrac{1}{2}(k_1 - k_2) = 2\pi/\lambda_m$$

$$k_w = \tfrac{1}{2}(k_1 + k_2) = 2\pi/\lambda_w$$

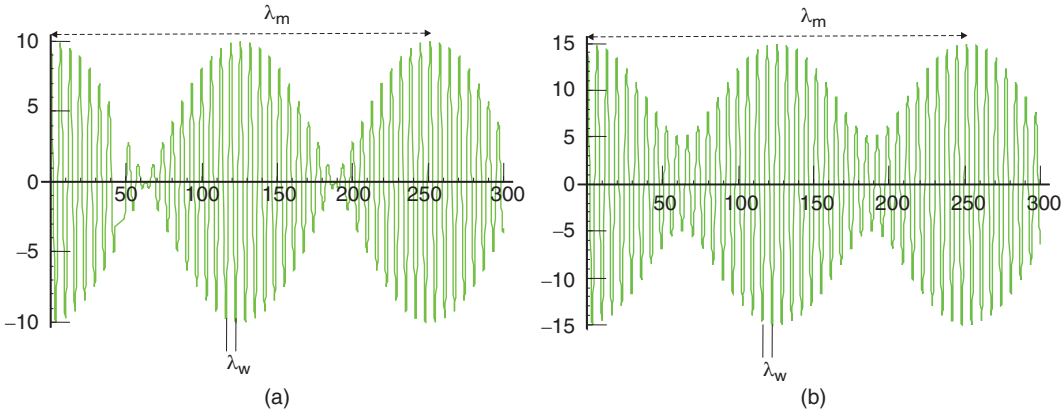


Figure 1.5 ***Beats: (a) the addition of two identical waves of amplitude 5, with*** $k_1 = 1$, ***and*** $k_2 = 0.95$; ***(b) the addition of two waves, one of amplitude 5, with*** $k_1 = 1$ ***and one of amplitude 10 with*** $k_2 = 0.95$

The wavelength of the new wave is λ_w and the wavelength of the modulation is λ_m (Figure 1.5a). The modulation of the wave produces *beats,* which are well known for audible waves and are used for the tuning of musical instruments. The same pattern results if the amplitudes of the two waves differ, but the modulation wave is found to decrease in amplitude (Figure 1.5b).

Note that in both of these cases, the wavelength of the wave has changed. In terms of colour, then, it is possible to say that mixing light of two different wavelengths changes the *colour* of the resultant light wave.

1.4.3 Phase and group velocity

In the preceding discussion, the wave speed was ignored. This omits an important part of the wave nature of light. To correct this, write the wave equations for the two interacting light waves as:

$$\mathcal{E}_1 = \mathcal{E}_0 \cos[(k_1 x - \omega_1 t)]$$
$$\mathcal{E}_2 = \mathcal{E}_0 \cos[(k_2 x - \omega_2 t)]$$

The resultant wave is:

$$\mathcal{E} = 2\mathcal{E}_0 \cos[k_m x - \omega_m t] \cos[k_w x - \omega_w t]$$

where:

$$k_m = \tfrac{1}{2}(k_1 - k_2) = 2\pi/\lambda_m; \omega_m = \tfrac{1}{2}(\omega_1 - \omega_2)$$
$$k_w = \tfrac{1}{2}(k_1 + k_2) = 2\pi/\lambda_w; \omega_w = \tfrac{1}{2}(\omega_1 + \omega_2)$$

The velocity of a plane light wave is given by:

$$v = \lambda\nu = \omega/k$$

(see Appendix A). It is thus clear that the high frequency wave, represented by k_w and ω_w, has a velocity:

$$v_w = \omega_w/k_w$$

This high-frequency wave is called the *carrier wave* or *signal wave*. The velocity is equivalent to the phase velocity described above, and in a vacuum, for a plane wave, this will be equal to the speed of light, c. The modulation wave, however, has a different speed, given by:

$$v_m = \omega_m / k_m = v_g$$

The modulation waves are said to move at the *group velocity*, v_g equal to v_m. The group velocity is often less than the carrier wave frequency, but can be manipulated to be faster or slower, or even zero.

1.4.4 Light pulses

A pulsed light beam, a pulse *train*, consists of a series of short intervals of waves in which the amplitude rises from zero to a maximum and then returns to zero again (Figure 1.6). As with beats, the wave within the pulse is known as the carrier wave or signal wave. A single pulse is called a *wave group* or a *wave packet,* and the shape of the pulse is called the *envelope*. As described above, the carrier wave, in a vacuum, travels at the speed of light, c, while the pulse envelope travels at the group velocity v_g.

Now as beats can be generated by adding together two continuous waves with different amplitudes and frequencies, a pulse train can be generated by adding together a range of waves with different amplitudes and frequencies. Thus, rather counterintuitively, a short pulse of, say, green light, contains a range of different wavelengths and not just that termed green. The details can be obtained rigorously by the methods of Fourier optics. Here the method is illustrated by a specific example.

Suppose we start with a continuous monochromatic carrier wave of wavelength λ_c of 0.5 μm (500 nm) in the blue-green travelling along the x direction, with amplitude of 1 (Figure 1.7a). Ignoring the time dimension for the present, and taking the amplitude $\mathcal{E}_0$ as 1, the wave equation is:

$$\mathcal{E} = \cos\left[(2\pi x/0.5)\right]$$

A single pulse is constructed by multiplying the carrier wave by a function that specifies the envelope that encloses the pulse, here taken as a Gaussian profile with a standard deviation of 0.5 (Figure 1.7b). A plot of the resulting equation:

$$\mathcal{E} = \cos[(2\pi x/0.5)]\ \exp\left[-x^2/2\,(0.5)^2\right]$$

creates a single pulse (Figure 1.7c). In order to generate a set of pulses separated by a distance which is the wavelength of the repetition, λ_g, in space, the wave packet must be repeated at the positions

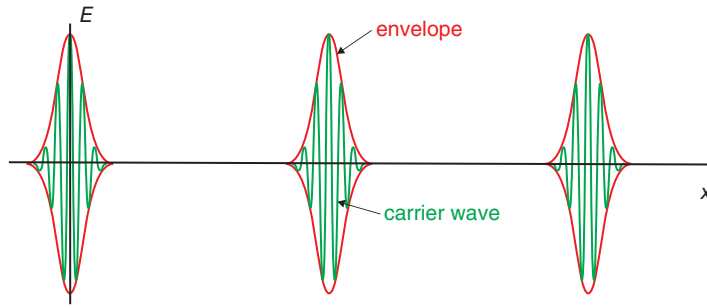


Figure 1.6 ***A pulse train***

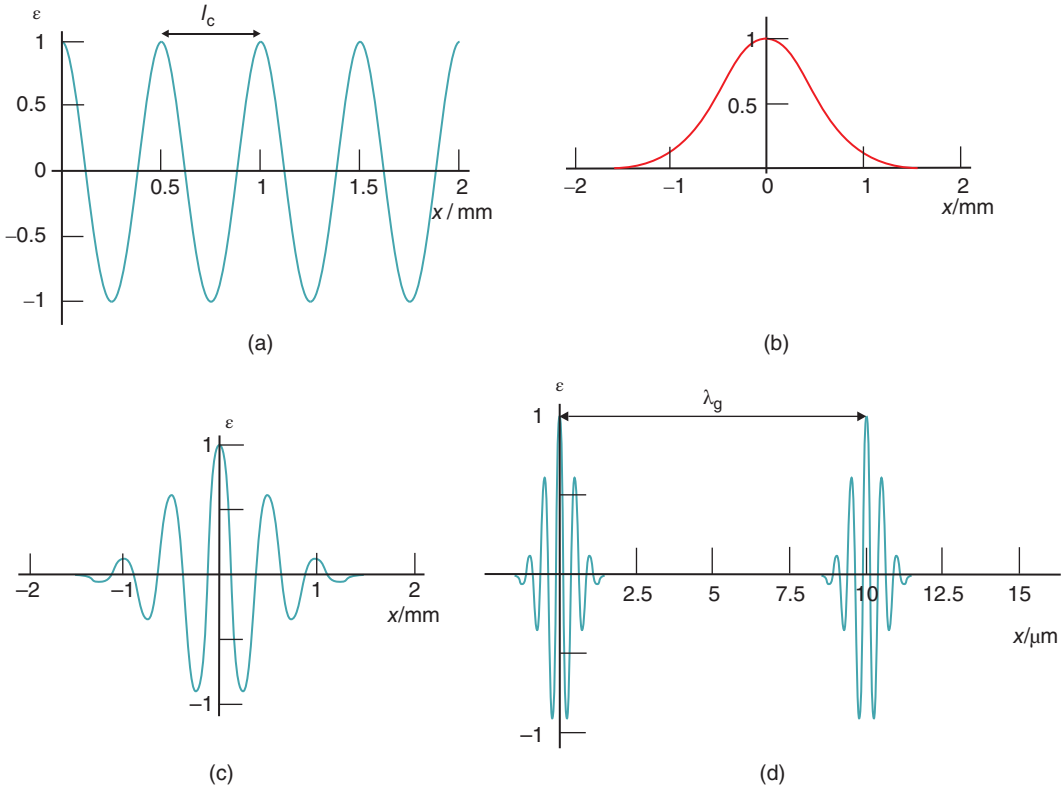


Figure 1.7 ***Pulse formation: (a) Monochromatic carrier wave, λ = 0.5 μm; (b) Gaussian envelope function; (c) pulse formed by multiplying (a) and (b); (d) pulse train with pulse spacing of 10 μm***

specified by a *comb function,* which consists of a set of spikes regularly spaced, exactly like the teeth on a comb. Here we take the wavelength of the comb to be 10 μm (Figure 1.7d).

The frequencies and amplitudes of the waves that are needed to reproduce this pulse train are derived by finding the Fourier transform of the composite function consisting of the wave equation, the pulse function, and the comb function. It is found that the pulse is made up of a large number of waves, consisting of harmonics of the carrier wave frequency. These are centred on the wavelength of the carrier wave, λ_c, 0.5 μm, with an increment between each frequency of $1/\lambda_g$, 0.1 μm, so that the wavelengths of the series of harmonics, λ_h, are given by:

$$1/\lambda_h = 1/\lambda_c \pm n/\lambda_g = 2 \pm 0.1n$$

The frequencies and wavelengths that span the visible are given in Table 1.2. It is seen that the pulse now contains wavelengths from *all across* the visible spectrum. The pulse is *multicoloured*.

It is simple to work out what will happen if the pulses are separated by longer gaps, 2 λ_g, 3 λ_g … by repeating the calculation given above. The pulse envelope will always retain the same shape, but the separation of the waves making up the spectrum under the pulse envelope, by $1/\lambda_g$, will change. Thus, a pulse separation 2 λ_g results in a decrease in the separation of the harmonics to $1/(2\,\lambda_g)$; a pulse separation 3 λ_g results in a decrease in the separation of the harmonics to $1/(3\,\lambda_g)$ … and so on. That is, more and more 'colours' can be imagined to occur within the pulse. By extrapolation, it is possible to

Table 1.2 ***Visible harmonics in single pulse of a train***

Relative amplitude	n	$1/\lambda_h$	$\lambda_h/\mu m$	λ_h/nm; (colour)
0.08	−7	1.3	0.769	769 (infrared)
0.16	−6	1.4	0.714	714 (deep red)
0.28	−5	1.5	0.667	667 (red-orange)
0.44	−4	1.6	0.625	625 (orange-red)
0.63	−3	1.7	0.588	588 (yellow)
0.82	−2	1.8	0.556	556 (yellow-green)
0.95	−1	1.9	0.526	526 (green)
1	0	2.0	0.500	500 (fundamental, green-blue)
0.95	1	2.1	0.476	476 (blue-green)
0.82	2	2.2	0.455	455 (blue)
0.63	3	2.3	0.435	435 (indigo)
0.44	4	2.4	0.417	417 (violet)
0.28	5	2.5	0.400	400 (deep violet)

anticipate that a single isolated pulse will contain a *continuum* of wavelengths (or 'colours') from 400 to 700 nm, as well as wavelengths outside the visible which extend into the ultraviolet and infrared.

What happens if the pulse is shortened or lengthened? The pulse width is characterised by the standard deviation, σ, of the envelope function. The mathematics shows that as the standard deviation of the envelope function gets broader the pulse envelope itself gets narrower. The spatial frequencies lying within the pulse envelope remain the same spacing, $1/\lambda_g$, but as the pulse narrows the central frequency corresponding to λ_c becomes more dominant. By extrapolation, it is possible to anticipate that a single continuous beam (an infinitely broad pulse) will now be composed of a single colour corresponding to λ_c.

1.4.5 Superluminal and subluminal light

The fixed *phase velocity* of a *plane light wave* in a vacuum, c, is one of the fundamental constants of physics. In a transparent solid, the speed of a planar light wave is less than c due to the interaction of the light with the electrons around the atoms that makes up the solid. The extent of the slowing in a material is be represented by its *absolute refractive index*, given by:

$$n = c/v$$

where c is the velocity of light in a vacuum and v is the phase velocity of light in the medium (Chapter 2). The reports of superluminal light (fast light) and subluminal light (slow light) do not refer to this aspect of things but to the speed of light *pulses* not of continuous monochromatic plane waves.

A single light pulse is made up of an infinite number of superimposed monochromatic waves. In order to understand fast and slow light it then becomes necessary to define the speed of the light pulse (or a group of such pulses). This is not as easy as all that. The peaks and troughs of the component waves will still move at the phase velocity defined above. However, the pulse shape depends on the interference of all of the harmonics. In a vacuum this is not a problem. The phase velocity of each component is the same, c, and the pulse travels through space as a well-defined shape. However, because of normal dispersion (Chapter 2), a pulse will gradually spread out and change shape as it crosses a

normal transparent solid such as a slab of glass, because each of the monochromatic components of the pulse will move at its own individual phase velocity. For short propagation distances, it is reasonable to ignore pulse distortion, so that a pulse velocity can be written v_g, which is related to a pulse refractive index n_g:

$$n_g = c/v_g$$

To give an example, the apparent speeding up of a light pulse can be achieved by directing a pulse of light at a cell containing Cs atoms. These atoms are initially put into an excited state via laser excitation. As the first photons in the leading edge of the pulse reach the cell, they trigger stimulated emission from the excited Cs atoms. The pulse is now regenerated somewhat ahead of the main peak of the following pulse. As the pulse moves into the cell, photons start to become absorbed by the now ground state Cs atoms. This means that a new pulse is recreated somewhere in front of the original pulse. Thus, when the new pulse emerges from the cell, it is earlier than an equivalent that does not pass through the cell, forming superluminal light.

Similar light pulse/atomic interactions, starting with atoms in the ground state, can lead to the pulse appearing to slow, termed subluminal light.

1.5 Light Sources

Until the end of the nineteenth century, artificial illumination was via incandescence – firelight, candles, oil lamps, or gas light. Incandescence is the emission of light by a hot body. The sun provides a commonplace example and is regarded as producing (more or less) white light. The light emanating from the upper part of a candle flame or firelight also arises from incandescence. In this case, small particles of carbon are heated to high temperatures in the flame and emit light that is perceived as red-yellow in colour.

Gas lighting is slightly different. The increasing availability of coal gas in towns and cities led to the development of gas lamps. In these, a gas jet was ignited and used to heat a mantle, which, when hot, emitted a greenish-white light. The gas mantle was a small cotton or other fabric bag, the fibres of which were impregnated with lanthanoid oxides. Initial operation burnt off the cotton matrix, leaving a delicate tracery of oxides that became incandescent under the action of the interior gas jet. (Camping gas lamps still operate in a similar way.)

At the end of this period, new light sources began to be invented in parallel with the generation and availability of electricity. In 1897, Nernst invented the ‘glower’. This lamp consisted of a bar of electrically conducting ceramic made from a mixture of lanthanoid oxides similar to those in gas mantles. The ceramic bar became incandescent under the action of an electric current. Although Nernst glowers were widely used and were more efficient than the competing incandescent carbon filament electric light bulbs developed by Edison, they fell into disuse after the successful introduction of tungsten filament lamps following the invention of the Coolidge process for the production of ductile tungsten wires that allowed for the fabrication of lamp filaments. Throughout the twentieth century, tungsten filament lamps dominated the lighting market.

Although incandescence was the most widespread source of artificial light during this epoch, other types of lighting were well known. These use various forms of luminescence for light production (Section 1.7). Luminescence is a term that signifies the production of light from a cool body, generally caused by transitions between the various energy levels available to the component atoms or molecules. Neon signs and various forms of luminescence were used in specialist light-generating ways, such as, in the case of neon signs, for advertising. These latter mechanisms relied directly on atomic transitions in a way that was obscured in the complex incandescence reactions. Instead of generating a continuous white-light spectrum, these new light sources tended to give out coloured light.

Moreover, the colour produced could be tuned to a greater or lesser extent by changing the atomic makeup of the light-emitting phase.

Recently, many more light emitters have been synthesised, apparent in the many forms of display now commonplace, including TV screens, computer screens, and the displays on many hand-held devices such as mobile phones. These mainly use some form of luminescence, including light-emitting diodes (LEDs), OLEDs (organic light-emitting devices), and related technologies. The ways in which these generate light are explained in the following chapters, but a point of importance is that the light emitted from these sources is incoherent and usually unpolarised, as with the other light sources described.

Towards the middle of the twentieth century, advances in communication technologies reinforced the utility of using radiation directly to carry signals. This necessitated the use of coherent radiation. Initially, the push came from radio, as radio waves are normally emitted as a coherent wave train, not as incoherent waves. The wavelength of the waves used for carrying signals continually decreased from long waves, via medium waves, to short waves. At the same time, the engineering skills required to encode greater and greater information on these carrier waves increased to an amazing extent, making television and stereo broadcasting a norm. Unfortunately, the production of coherent radiation seemed to be stuck somewhere in the microwave region. The idea of using lower wavelengths, especially optical wavelengths, was, though, enormously attractive, and a great deal of effort was invested into breaking into this wavelength range. Success came in the 1960s, with the invention of the laser.

Lasers are a completely new sort of lamp compared with those already described. The word *laser* is an acronym for the expression *Light Amplification by Stimulated Emission of Radiation*. The first laser to be made was the ruby laser, and the first laser light emitted was on May 15, 1960. Since then a vast number of lasers have been produced, including solid state lasers, gas lasers, semiconductor diode lasers, and dye lasers – so much so that lasers have become ubiquitous in modern life, being used as pointers, at check-outs in supermarkets, in surveying and measurement, in micromachining, microsurgery, and so on.

1.6 Incandescence

1.6.1 Incandescence and black-body radiation

Incandescence is the emission of light by a hot body. When light from an incandescent object is spread out according to wavelength by a prism, the result is a continuous fan of colours following the sequence listed in Table 1.1 and called a *continuous spectrum*. The radiation emitted is both incoherent and unpolarised.

Incandescence comes about in the following way. At absolute zero, all atoms and molecules making up the solid are in the lowest possible energy state. As the temperature increases they absorb energy and are promoted to higher energies and, at the same time, atoms and molecules which have already absorbed energy loose some and they fall back to lower energies. (The energy levels involved in this process will be described in more detail in later chapters.) The radiation emitted in this way effectively extends over a continuous range of energies. For a solid somewhere above room temperature all the wavelengths of the emitted energy lie in the infrared, and although the radiation is invisible it may be detectable as a sensation of warmth. At a temperature of about 700 °C the shortest wavelengths emitted creep into the red end of the visible spectrum. The colour of the emitter is seen as red and the object is said to become *red hot*. At higher temperatures, the wavelengths of the radiation given out extend increasingly into the visible region and the colour observed changes from red to orange and thence to yellow, as in the example of a candle flame, mentioned above. When the temperature of

the emitting object reaches about 2500 °C all visible wavelengths are present and the body is said to be *white hot*. The sun provides a perfect example, and the 'colour' white is applied to light that is a combination of wavelengths that spans the visible spectrum with a similar composition as that of the radiation from the sun.

These qualitative colour changes can be understood in terms of the radiation emitted by a black body (Section 1.1.3). The spectrum emitted by the sun is similar to that for a black body at 5700 °C and that from a red hot object is similar to the curve for a black body at 700 °C. The black-body emission curve derived by Planck can be written in a number of forms. The *spectral radiance*, L, within a black body at equilibrium at temperature T in the frequency range ν to $\nu + \delta\nu$ or the wavelength range λ to $\lambda + \delta\lambda$ is:

$$L_\nu = 2h\nu^3/c^2\,[\exp(h\nu/k_B T) - 1]$$

units $W\,m^{-2}\,sr^{-1}\,Hz^{-1}$

$$L_\lambda = 2hc^2/\lambda^5\,[\exp(hc/\lambda k_B T) - 1]$$

units $W\,m^{-3}\,sr^{-1}$

In these equations, h is Planck's constant, c is the speed of light, k_B is Boltzmann's constant and T the temperature of the body (K). This equation is often seen in the form describing the *spectral irradiance* (if the energy falls upon a surface), or the *spectral exitance* (if the energy is observed after leaving a black body via a pinhole not large enough to disturb the thermal equilibrium within), I_ν in the frequency range ν to $\nu + \delta\nu$ or I_λ, in the wavelength range λ to $\lambda + \delta\lambda$ for a black body at a temperature T:

$$I_\nu = 2\pi h\nu^3/c^2\,[\exp(h\nu/k_B T) - 1]$$

units $W\,m^{-2}\,Hz^{-1}$

$$I_\lambda = 2\pi hc^2/\lambda^5\,[\exp(hc/\lambda k_B T) - 1]$$

units $W\,m^{-3}$or as the corresponding spectral energy density, u_ν in the frequency range ν to $\nu + \delta\nu$ or u_λ, in the range λ to $\lambda + \delta\lambda$ for a black body at a temperature T:

$$u_\nu = 8\pi h\nu^3/c^3\,[\exp(h\nu/k_B T) - 1]$$

units $J\,m^{-3}\,Hz^{-1}$

$$u_\lambda = 8\pi hc/\lambda^5\,[\exp(hc/\lambda k_B T) - 1]$$

units $J\,m^{-4}$.

In the mid-twentieth century, it was realised that the cosmos was filled with background electromagnetic radiation. The peak of the radiation lies in the microwave part of the electromagnetic spectrum. Naturally, it is invisible to optical instruments, and was first mapped using radio telescopes and latterly by satellites. The spectrum of this radiation fits that of a blackbody, and indeed this radiation – called the cosmic microwave background radiation, is possibly the most accurately measured black-body radiation curve available. It is interpreted as lending strong support to the 'Big Bang' theory of the origin of the universe.

1.6.2 The colour of incandescent objects

From the point of view of the colour of incandescent objects, one of the most important attributes of the emission curve is the variation in the position of the maximum as the temperature of the black

Table 1.3 ***Colour temperature of incandescent sources***

Light source	Colour temperature/K
Mean noon sunlight	5400
Electronic flash	~ 7000
Blue flash bulb	~ 6000
Tungsten filament photographic lamps	~ 3400
Tungsten filament light bulb, 100 W	2850
Standard candle	1930

body increases. This was derived before the Planck radiation law, and represents the final success of classical electromagnetic theory. The relationship, known as the *Wien displacement law*, is:

$$\lambda_{max}\, T = \text{constant}$$

where T is the temperature of the body (K) and the constant has a value of 0.002 898 m K. It can be derived from the black-body radiation law by differentiating with respect to λ and setting the result equal to zero. The colour of an incandescent object is thus controlled by the maximum of the emission curve. The second factor of importance is the spread of the spectrum. A cool body will be perceived as initially showing a colour when the peak of the curve is close enough to the visible range that some radiant energy creeps into the low energy (red) end of the spectrum. As the temperature of the incandescent object increases the peak moves to higher energies, following the displacement law, and the spread moves further across the visible spectrum, resulting in the colour sequence of dull red, red hot to white hot, to blue-white.

The colour of an incandescent object with an emission spectrum that resembles that of a black body closely is described by its *colour temperature* (Table 1.3). Most solids behave in this way over some range of temperature and wavelength and stars are a close approximation over the whole of the wavelength range. If the match between the emission spectrum and the theoretical black-body curve is only approximate, the term used is *correlated colour temperature*. This expression is used for light sources that are not incandescent.

The most important incandescent object for us is the sun, which is the ultimate source of energy on Earth. The solar spectrum has a form quite similar to a black-body curve corresponding to a solar temperature of about 5780 °C (about 6000 K), which has a maximum near 480 nm. The form of the spectrum when it reaches the surface of the Earth is a function of a number of variables, including the elevation of the sun, the amount of scattering material in the atmosphere, and so on. Light is perceived as white if it has a make-up like that of the solar spectrum from an overhead sun on a clear day. Stars that are cooler than the sun appear redder whilst those that are hotter are perceived as blue-white. The *effective temperature* of a star is the temperature calculated as if it were a black body radiating with the same energy over the same wavelength ranges (Table 1.4). The effective temperature is generally a good approximation to the surface temperature of a star. The hottest visible stars are the Orion type, with blue-white colour and an effective temperature of approximately 25 000 K whilst the reddest naked-eye star is μ Cephei, the Garnet Star, with a temperature of approximately 2600 K.

1.7 Luminescence

In contrast to light emission by a hot body, incandescence, the emission of light by bodies at relatively low temperatures, ‘cold light’, is generally called *luminescence*. It is more scientifically defined as

Table 1.4 ***Effective star temperatures***

Star colour and example	Effective temperature/K
Blue-white, Bellatrix	25 000
White, Sirius	11 000
Yellow-white, Sirius – Solar	7 500
Solar, the Sun	6 000
Orange-yellow, Arcturus	4 200
Orange, Antares	3 000
Deep orange-red, μ Cephei	2 600

the spontaneous emission of radiation from an electronically or vibrationally excited species not in thermal equilibrium with its environment. Fluorescent lighting and white light emitting LED lamps, common forms of indoor illumination, fall into the category of luminescence, as do lasers, described in the following section. However, the term encompasses a considerable range of phenomena, including light emitted by many organic and inorganic materials following exposure to light (termed photoluminescence), electric fields (electroluminescence), chemical reactions (chemiluminescence), as well by living organisms such as fireflies, jellyfish, and fungi (bioluminescence). Luminescence has been recorded from ancient times and occurs in many folklore fables and stories.

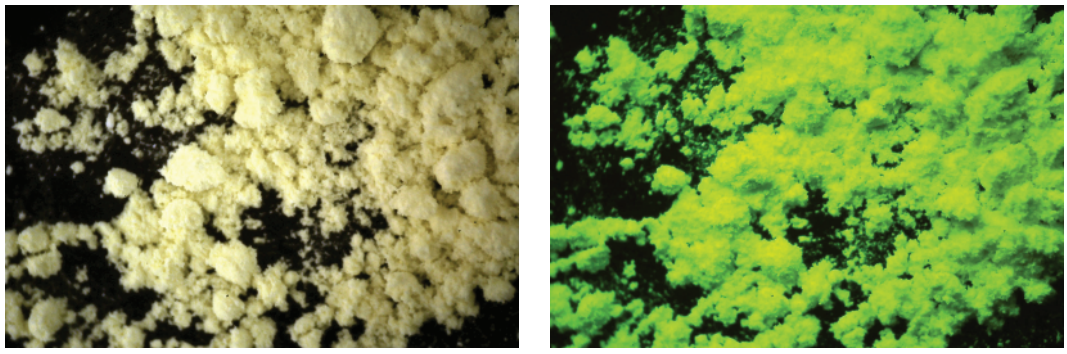
Solids that give rise to luminescence are called *phosphors* or latterly, *luminescent materials*. The term *phosphor* is derived from the element phosphorus. This element was first isolated in about 1674 or 1675 by the alchemist Brandt, who discovered that the material shone with a pale greenish light in the dark, and gave it the name phosphorus, which is from the Greek, *phos* (light) and *phero* (I carry). The name for the glow from phosphorus, *phosphorescence*, was taken over and quite widely applied to many other forms of 'cold light' that came from a variety of organic and inorganic bodies, including decaying organisms.

The modern understanding of photoluminescence can be thought to start with Ritter, in 1801, who, investigating solids that would glow after being illuminated with daylight, discovered that the effect was greatest when the sample was placed in the dark region beyond the violet end of the spectrum. He therefore postulated that an invisible form of *light,* termed *ultraviolet,* existed and was instrumental in producing the effect.

In 1852, Stokes published the results of an extensive investigation in which he invented the term *fluorescence* for the light that he observed emanating from crystals of the mineral fluorspar on illumination with ultraviolet light (Figure 1.8). This fluorescent light emission ceased when the exciting light was turned off, unlike the more long-lasting phosphorescence that seemed to decay slowly, if at all, in the dark. Stokes' law, proposed at this time, stated that the fluorescent light emitted has a longer wavelength (lower energy) than the exciting radiation. The wavelength difference is known as the *Stokes shift*. A.-E. Becquerel, in France, also studying the emission of light by solids after illumination, was of the opinion that fluorescence was simply a short duration form of phosphorescence – in which he was essentially correct.

Roughly coincident with the fluorescence and phosphorescence studies of Stokes and Becquerel, researches by Faraday, Giessler, Crookes, and others showed that gases and some solids produced light when bombarded with cathode rays, that is, energetic electrons (Section 7.6). To distinguish this light from fluorescence and phosphorescence, the effect was called *cathodoluminescence*.

Yet another form of luminescence had an enormous impact on twentieth century science. H. Becquerel, the son of A-E. Becquerel, was studying phosphorescence when he discovered, in 1896, that uranium salts emitted a new sort of radiation, uranium rays. Pierre and Marie Curie



(a) (b)

Figure 1.8 *(a) a pale yellow phosphor based on zinc sulphide, ZnS, in normal daylight; (b) the same material irradiated with ultraviolet light with a wavelength range of 350–380 nm*

followed this up and discovered radioactivity and the highly radioactive element radium. The radioactivity is intense and this causes many materials to emit light, now called *radioluminescence*.[4] Following this discovery, radium was widely used in luminous paints for watch and other instrument dials that could be seen in the dark, as well as in a number of patent medicines. This has now ceased due to the harmful effects of the radiation emitted.

These preceding studies pointed the way towards a general understanding of luminescence. It is now clear that luminescent materials are able to gain energy from an energetic, often 'invisible' source (ultraviolet light, electric fields, X-rays, energetic particles from radioactive decay, and so on[5]) and re-emit some of this energy in the form of light. For this reason, luminescence is now subdivided into a number of categories depending on the nature of the exciting source (Table 1.5). However, this does not imply that a single process operates in all cases and there are no general mechanisms for luminescence. Thus, the mechanism of light emission by a firefly is quite different from light emission by a fluorescent lamp, although both may be termed luminescence. In fact, the phosphorescence of phosphorus occurs when the solid is exposed to moist conditions, which causes mild oxidation. The light so generated is, therefore, properly described as chemiluminescence – light production as the result of an ongoing chemical reaction.

1.8 Laser Light

1.8.1 Emission and absorption of radiation

When a photon of energy $h\nu$ is *absorbed* by an atom or molecule, it passes from the normally occupied ground state, energy E_0 to an upper or excited state, energy E_1 (Figure 1.9a). If the atom is in the excited state and makes a transition to the ground state, an identical photon will be emitted. In this description, the actual emission mechanism is ignored. In 1917, Einstein suggested that there should be *two* possible types of emission process:

[4]Marie Curie described the phenomenon thus: 'One of our joys was to go into our workroom at night; we then perceived on all sides the feebly luminous silhouettes of the bottles or capsules containing our products. It was really a lovely sight and one always new to us. The glowing tubes looked like faint, fairy lights'.

[5]Materials that emit light in response to the impact of high energy particles or X-rays are often termed *scintillators* rather than phosphors.

Table 1.5 ***Luminescence***

Type	Definition	Source of energy	Section
Bioluminescence	Luminescence in a living organism	Gibbs energy of chemical reactions	9.14
Cathodoluminescence	Luminescence due to electron bombardment (cathode 'rays')	Electron kinetic energy	9.6, 9.7
Chemiluminescence	Luminescence during a chemical reaction	Gibbs energy of chemical reactions	9.14
Electroluminescence	Luminescence resulting from the application of an electric field (can be equivalent to cathodoluminescence)	Electrical potential energy	LEDs 10.7 OLEDs 9.17, 9.14
Electrochemiluminescence (Electrogenerated chemiluminescence)	Luminescence via the chemical reaction of one or two molecules excited by the application of an electric potential	Electrical potential energy	9.14
Photoluminescence	Luminescence after irradiation by visible or ultraviolet light	Ultraviolet and visible photon energies	9.1–9.6
Fluorescence	Rapid decay of excited state		
Phosphorescence	Slow decay if excited state		
Radioluminescence	Luminescence as a result of radioactivity	energetic particles and γ rays	9.13
Sonoluminescence	Luminescence as a result of sound waves	Acoustic vibrations, pressure waves	9.15
Thermoluminescence	Luminescence following an increase of temperature	Thermal energy	9.12
Mechanoluminescence	Luminescence following deformation or friction	Strain energy, thermal energy, chemical bond energy	9.15
Triboluminescence	Luminescence following grinding, crushing	Chemical bond energy	9.15

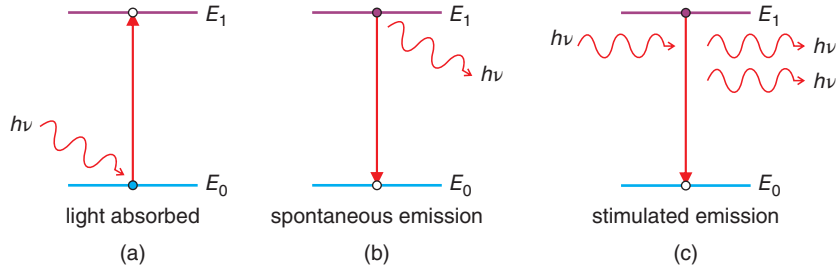


Figure 1.9 ***Light absorption and emission: (a) a photon excites an atom from the ground state*** E_0 ***to an excited state*** E_1***; (b) spontaneous emission produces photons at random; (c) stimulated emission produces photons in phase***

1. *Spontaneous emission*. An atom in an excited state can randomly change to the ground state (Figure 1.9b).
2. *Stimulated emission*. A photon having an energy equal to the energy difference between the two levels, that is, $(E_1 - E_0)$, can interact with the atom in the excited state, causing it to fall to the lower state and emit a photon at the same time (Figure 1.9c)

In spontaneous emission, the light photons all have the same frequency, because the energy difference between the excited and ground state is constant at $(E_1 - E_0)$ but they possess random phases and polarisation, making the light *incoherent*. In stimulated emission, the photon produced not only has the same frequency but also the same phase and polarisation as the one that caused the emission, making the light *coherent*. It is these two important features of stimulated emission on which the special properties of laser light depend.

1.8.2 Energy-level populations

Under conditions of *thermal* equilibrium, the relative populations of a series of energy levels will be given by the Boltzmann law, which, for two energy levels can be written as:

$$(N_1/N_0) = \exp\left[-(E_1 - E_0)/k_B T\right]$$

where k_B is Boltzmann's constant, T is the absolute temperature, E_1 and E_0 are the energies of the excited state and the ground state, respectively, and N_1 and N_0 are the numbers of atoms (the *populations*) in each of these energy levels. For ordinary atoms, in a gas, liquid, or solid at ordinary temperatures, the fraction N_1/N_0 will be negligible for energy levels that are sufficiently separated to give rise to visible light. Atoms can be assumed to be in the ground state as far as visible light emission is concerned.

Under these conditions, when a photon of the appropriate energy interacts with an atom in the ground state, it will be absorbed and shortly afterwards re-released by spontaneous emission (Figure 1.10a). This will be repeated at each atom in the ground state. There will be no amplification and we may well see a net absorption of energy. To obtain laser amplification, one needs to ensure that stimulated emission is the dominant process occurring. This means that there are more atoms in the excited state of energy E_1 than in the ground state, E_0. In this instance, a photon interacting with an excited atom can cause energy to be released by stimulated emission and two photons emerge. If most atoms are in the excited state, amplification can occur (Figure 1.10b). The situation in which more atoms are in the excited state than the ground state is called a *population inversion*.

From the Boltzmann equation, it is obvious that an increase in temperature cannot achieve this object. Even an infinite temperature will only result in equal numbers of atoms in E_0 and E_1. To obtain

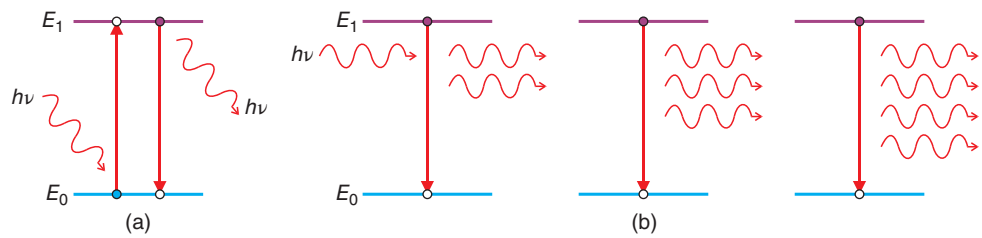


Figure 1.10 ***Amplification: (a) absorption of a photon and the subsequent spontaneous re-emission will not lead to amplification; (b) when most atoms are in the excited state stimulated emission can lead to amplification***

a population inversion, therefore, a *nonequilibrium state* must be achieved. The crux of laser action is how to create such a nonequilibrium situation in a material and then exploit it to produce the desired amplification. Chapters 7 and 10 provide examples of practical ways in which this is achieved.

1.8.3 Rates of absorption and emission

In the previous section, it was implicitly implied that the rate of spontaneous emission was fast. This aspect must be looked at in more detail to obtain a better understanding of laser action. When equilibrium between absorption and emission holds, the rate of depopulation of an upper level ($-\mathrm{d}N_1/\mathrm{d}t$) by spontaneous emission is given by a first-order rate law:

$$-\mathrm{d}N_1/\mathrm{d}t = A_{10}\,N_1$$

where the negative sign denotes that the number N_1 of atoms in the upper state E_1 (per cubic metre, say) is decreasing with time. The rate is proportional to the number of atoms in the state, N_1. The rate constant, denoted here as A_{10}, is called the *Einstein coefficient for spontaneous emission*, and the suffix 10 means that we are considering a transition from the excited state, E_1 to the ground state E_0. The number of downward transitions due to spontaneous emission per second will be given by:

$$A_{10}N_1$$

Similar rate laws can be written for the cases of stimulated emission and for absorption, but in this case the rates are proportional to the numbers of atoms in the relevant state and, in addition, the number of photons present. The reactions can be taken to be first order with respect to both of these quantities.

The rate at which atoms in state E_0 are excited to state E_1 is then given by

$$-\mathrm{d}N_0/\mathrm{d}t = B_{01}\,\rho(\nu_{01})\,N_0$$

where N_0 is the number of atoms in state E_0 (per cubic metre, say), $\rho(\nu_{01})$ is the radiation density responsible for absorption, which is the number of quanta per cubic metre incident per second at the correct excitation frequency ν_{01}, and B_{01} is the Einstein coefficient for absorption of radiation. Similarly, the rate of depopulation of state E_1 by stimulated emission is given by

$$-\mathrm{d}N_1/\mathrm{d}t = B_{10}\,\rho(\nu_{10})\,N_1$$

where N_1 is the number of atoms in state E_1 (per cubic metre), $\rho(\nu_{10})$ is the radiation density responsible for depopulation, which is the number of quanta per cubic metre incident per second at the correct frequency ν_{10}, and B_{10} is the Einstein coefficient for stimulated emission of radiation. Now the correct frequency for excitation will be the same as that for depopulation, so that $\nu_{10} = \nu_{01}$, which we can simply write as ν, and the radiation density will be the same in each case, so that we can write:

$$\rho(\nu_{10}) = \rho(\nu_{01}) = \rho(\nu)$$

The number of stimulated downward transitions per second will be given by

$$N_1\,B_{10}\,\rho(\nu)$$

while the total number of upward transitions in the same time will be given by

$$N_0 B_{01} \rho(\nu)$$

At equilibrium, the total number of transitions in each direction must be equal, hence:

$$N_0 B_{01} \rho(\nu) = N_1 A_{10} + N_1 B_{10} \rho(\nu)$$

so

$$\rho(\nu) = N_1 A_{10}/(N_0 B_{01} - N_1 B_{10})$$

In addition, at equilibrium the Boltzmann distribution applies, thus:

$$N_1/N_0 = \exp(-h\nu/k_B T)$$

and by making this substitution:

$$\rho(\nu) = A_{10}/[\exp(h\nu/k_B T) B_{01} - B_{10}]$$

This expression represents the radiation density at frequency ν. At thermal equilibrium, this should be identical to Planck's black-body equation:

$$\rho(\nu) = 8\pi h\nu^3/c^3 [\exp(h\nu/k_B T) - 1]$$

which leads to the conclusion that:

$$B_{01} = B_{10} = B$$

and:

$$A_{10}/B = 8\pi h\nu^3/c^3$$

The *ratio* of the rate of spontaneous emission to stimulated emission under conditions of thermal equilibrium is given by

$$R = A_{10}/\rho(\nu) B = \exp(h\nu/k_B T) - 1$$

This is an extremely interesting result. At 300 K, at visible wavelengths, R is much greater than 1. This shows that for light, stimulated emission will be negligible compared to spontaneous emission and reinforces the idea that it will be impossible to make a laser under equilibrium conditions. On the other hand, if the wavelength increases beyond the infrared into the microwave and radio wave regions of the electromagnetic spectrum, R becomes *much less* than 1 and *all* emission will be stimulated. Hence, radio waves and microwaves arise almost entirely from stimulated emission and are always coherent. This is one of the main reasons that communications in the early part of the twentieth century was able to use radio waves.

Perhaps because of this equation and the towering reputation of Einstein, it seems that for the first part of the twentieth century it was felt that lasers were not feasible. In the middle of the century, scientists started to explore stimulated emission at microwave frequencies, developing the *maser*. This soon led to the first lasers (often called *optical masers*), the ruby laser, and then the He-Ne gas laser, produced in 1960. Once the way to overcome the production of laser light was understood, laser development became prolific.

1.8.4 Cavity modes

Supposing that a population inversion is obtained between energy levels that would give rise to visible light, it is still necessary to design the equipment so that amplification of the signal takes place. One of the most important of these design features is the shape of the laser-emitting phase. This is referred to as the *cavity* that the laser medium (or gain medium) occupies. Suppose that this is simply a crystal rod. Once the population of active centres is in a state of population inversion, it is in an unstable condition, and after a short time, some spontaneous emission will occur from E_1. Naturally, these photons will rapidly leave the crystal rod, and although in so doing a few other atoms might lose energy via stimulated emission, no amplification will occur. It is necessary to prevent these photons from leaving the crystal in order to increase the chances of stimulated emission occurring. The simplest way to achieve this is to coat the ends of the crystal rod with a highly reflecting mirror. In this case, the photons are reflected to and fro, causing stimulated emission from the other populated E_1 levels. Once started, the stimulated emission rapidly depopulates these levels in an avalanche. In order to allow some light to emerge, one of the mirrors is not perfect and allows a small amount of light to pass. There will then be a burst light emerging from the cavity that is not only coherent but also shows amplification.

This simplest cavity geometry is simply cylindrical with one end fitted with a completely reflecting mirror and the other with an almost perfect mirror, appropriate to the wavelength of the light generated by the stimulated emission. There are several consequences of this, which are easiest to explain if the light trapped in the cavity is regarded as a wave. It is clear that initially all photons will be emitted in a random direction, but only those that are emitted more or less parallel to the long axis of the cavity will bounce to and fro and so cause the stimulated emission avalanche. In terms of wave optics, the photons form a series of standing waves in the cavity, which is described as *resonance*. The standing waves form only if there is a node at each reflecting surface. The allowed waves are called *longitudinal cavity modes*, and are given by the condition that a complete number of half wavelengths must fit into the length of the cavity, i.e.:

$$m_c = l/(\lambda/2) = 2\,l/\lambda$$

where m_c is an integer, l is the cavity length, and λ the wavelength of the mode. The frequency of a mode is given by

$$\nu_m = m_c\, v/2l$$

where v is the velocity of the light waves in the cavity, given by $\nu_m\,\lambda$, and ν_m is the frequency of the mode m_c. The separation of the modes is given by

$$\nu_m - \nu_{m-1} = \Delta\nu = v/2l$$

The velocity of light in the cavity is given by

$$v = c/n$$

where c is the velocity of light in a vacuum and n is the refractive index of the cavity medium so that

$$\nu_m - \nu_{m-1} = \Delta\nu = c/2ln$$

How does this work out in practice? The emission from the upper to the lower energy level has been written as a single energy with a negligible width. In the case of real materials, atoms and molecules are in continuous motion, vibration in solids, translation in gases, and the sharp idealised energy levels

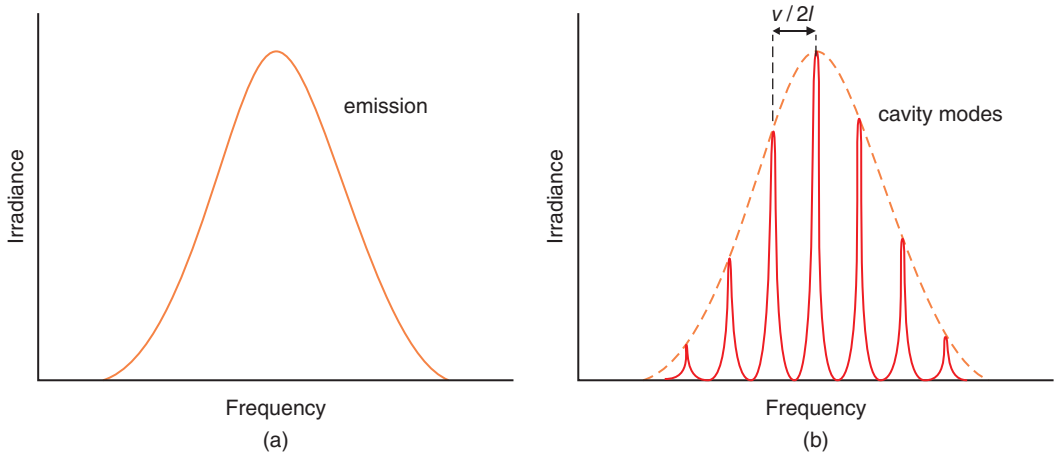


Figure 1.11 *(a) The emission from an excited state* E_1 *to the ground state* E_0 *is not sharp, but consists of a range of frequencies dependent on temperature and other factors; (b) In a laser cavity, only certain frequencies, the cavity modes, are allowed to propagate*

in Figures 1.9 and 1.10 give rise to a spread of energies (or of frequencies or wavelengths) called the *transition bandwidth* (Figure 1.11a). Only that part of this output that fulfils the longitudinal mode criterion will then be allowed to grow. The output from the cavity will then be composed of a set of modes (Figure 1.11b). These modes will depend on the shape of the initial emission pulse and the overall power of the excitation process. A change in temperature will change the length of the laser cavity. This will cause a change in mode spacing and in the mode wavelengths. For precise work, carefully stabilised lasers must be used.

By extension, it is apparent that, in general, there will be transverse modes as well as longitudinal modes, in the laser emission. These must be taken into account when the optics of a laser beam are considered. Laser cavity design is therefore of considerable importance in practice.

1.8.5 Coherence length and bandwidth

The emission from an excited state E_1 to the ground state E_0 is not sharp, as has been assumed up until now, but consists of a range of frequencies dependent on temperature and other factors. A spectral line thus has a width, called the linewidth, quoted in nm, or the bandwidth, quoted in Hz (Figure 1.11a). This spread is made up from two factors:

1. The duration of the transition, approximately 10^{-8} to 10^{-9} seconds, giving rise to the *natural linewidth*.
2. The temperature of the medium containing the atoms, as thermal motion, vibrations, and interactions with other atoms considerably broadens the natural linewidth. The spectra from atoms in a low-pressure gas are considerably sharper than those from atoms in a high-pressure gas, which themselves are much sharper than the spectra from atoms or ions in solids.

The extent of the line width or bandwidth of a spectral line is often taken to be the line width at a value of irradiance equal to ½ of that of the maximum irradiance, $\frac{1}{2} I_{max}$.

The relationship between the line width and the bandwidth is derived in the following way. Suppose a line has a linewidth $\Delta\lambda$ centred upon a wavelength λ_m. This is converted to a bandwidth, $\Delta\nu$ centred on a frequency ν_m by the relationship:

$$\nu = c/\lambda$$

where c is the velocity of light and we presume that the light is moving thorough air or a vacuum. It is found that:

$$\Delta\nu/\Delta\lambda = \nu_m/\lambda_m$$

The concept of coherence described in Section 1.2 needs to be amplified when discussing laser light. Imagine a beam of light emerging from a laser pointer. This light can be considered, to a reasonable approximation, as coming from a single point source. (It is analogous to normal incoherent light that has passed through a pinhole, but of much higher irradiance, of course.) As the wavefront spreads away from the pointer, the beam can be thought of a *spatially coherent* if corresponding points on the wavefront always show the same relative phase. That is to say that there is a fixed-phase relationship between two points at different positions across the wavefront. Similarly, as the wavefronts pass by a fixed point, if the wave always shows a fixed phase then the wave is said to be *temporally coherent*. The electric field arising from a laser beam can show perfect spatial and temporal coherence, a high degree of spatial coherence and poor temporal coherence, or a high degree of temporal coherence and poor spatial coherence.

The coherence length of a light beam is the length of a wave (with a fixed phase), measured in wavelengths, before there is an abrupt change of phase. The coherence length, Δs, is simply the number of complete waves in the stretch multiplied by the wavelength:

$$\Delta s = N\,\lambda$$

The coherence time of laser light, Δt_c, is the time for the coherent stretch of wave to pass a fixed point. In air, this is given by the coherence length divided by the velocity of light in air, which is conveniently taken as c:

$$\Delta t_c = \Delta s/c$$

The bandwidth of a spectral line is related to the coherence length and the coherence time by;

$$\Delta\nu \approx 1/\Delta t_c$$

To give some illustrative figures, for a high-pressure lamp such as a mercury vapour lamp, the linewidth is about 1 nm, and the coherence length is about 0.5 mm; for a low-pressure lamp such as a sodium lamp, the linewidth drops to 1×10^{-2} nm and the coherence length is as much as 3 cm or so; for an ordinary He–Ne laser, the linewidth is approximately 1×10^{-3} nm and the coherence length is approximately 30 cm; for a high stability He–Ne laser, the linewidth is as narrow as 1×10^{-6} nm, with a coherence length of approximately 300 m.

In order to observe interference effects, the light waves must be both spatially and temporally coherent. When an ordinary He–Ne laser is used as the light source, strong interference will be observed if the light paths taken by the divided laser beam differ by less than 30 cm. This is easily obtained in practice, as the spatial coherence of the beam is high and the temporal coherence extends over some 30 cm. It is for this reason that laser light is normally just called coherent light. The coherence length of laser light is important when high-visibility holograms are required as the object beam and reference beam path length difference must be within the coherence length. In addition, for the most demanding work, the spatial coherence might need to be improved, which is achieved by inserting a filter, which is effectively a pinhole, to exclude waves that lack high spatial coherence.

1.8.6 Supercontinuum light

The light from lasers is typically made up of only a short range of wavelengths and is, for many purposes, considered to be monochromatic. However, laser light that spans a significant wavelength range, say over the visible, is of considerable interest for many purposes such as in optical frequency combs. Light from a laser that has a narrow spectral bandwidth which is converted into light that has a

broad spectral bandwidth, but with the coherence remaining high, is known as *supercontinuum light*. An infrared laser beam converted into a beam of white light spanning the whole of the visible is an example.

Supercontinuum light can be produced by propagating an optical pulse of laser light with a high irradiance through a nonlinear medium (Section 4.9). In fact, provided the irradiance is high enough, all transparent solids will function in this way due to the Kerr effect. Thus, passing an intense pulse of infrared light simply through a slab of glass will result in the production of a pulse of white light provided that the glass is not too thick. Unfortunately, this requires very high irradiance light, and the pulse spreading tends to be rather small.

In 2000, the use of a photonic crystal fibre (Sections 2.9 and 6.10) with a small solid core was found to produce broad spectrum supercontinuum light from pulses of relatively low irradiance, due to the strongly nonlinear nature of the fibre core. Note that supercontinuum light generated in this way consists of a broad pulse made up of vast numbers of sharp frequencies (a frequency comb) separated by the repartition rate of the laser pump. The physical processes that lead to the development of this light are complex and depend on the pulse duration, power, and the wavelength of the waves that make up the pulse. The topics needed to take this explanation further forward are refractive index, nonlinear optical materials, and photonic crystal optical fibres. These are covered in later chapters.

1.9 Vision

Vision in humans and other animals involves a complex set of reactions that take place in two principal types of photoreceptor cells located in the retina of the eye, *rods* and *cones*. (There is also a third type of photoreceptor in the eye called intrinsically photosensitive retinal ganglion cells, which respond to light rather slowly, and may be important in the registration of the onset of dawn and dusk.) There are about 10^8 rods and 4×10^6 cones in an eye. In humans, the rod cells, of about 0.002 mm diameter, are about four times as sensitive as cones and are responsible for vision at low light intensities. Although they detect light all across the visible, the peak sensitivity is at approximately 500 nm. The rod cells are not sensitive to colour and give rise to a monochrome image. Moreover, they saturate in high light levels, making them unresponsive under bright conditions.

The cone cells, approximately 0.006 mm diameter, are sensitive to bright light and form the daylight colour detection system. They exist in three varieties, with peak sensitivities in three different regions of the visible: L cones, most sensitive to red, λ(peak) ~560 nm, M cones, most sensitive to green, λ(peak) ~530 nm, and S cones, most sensitive to blue λ(peak) ~420 nm (Figure 1.12a). Human vision is thus termed *trichromatic*. The human eye is optimally sensitive to green light and is noticeably less sensitive to red and especially blue light (Figure 1.12b).

There is considerable variation across the human population in the sensitivity ranges of the cone cells, giving rise to a variation in colour vision. The sensitivity of the eye to colour depends not only on the amount of light but also on which area of the retina is being stimulated. The most sensitive region, called the *fovea*, is almost directly behind the lens of the eye, and predominantly contains cone cells. The maximum sensitivity of a normal eye to bright white light focused on the fovea, which is the sum of the contributions of the cone and rod cells, the photopic spectral luminous efficiency function, is for a wavelength close to 555 nm (Figure 1.12c). Colour blindness results from a fault or deficiency in one or more varieties of the cone cells or in the way in which these cells communicate with the brain.[6]

[6]The existence of colour blindness itself was first recorded as such by John Dalton, who realised that his own perception of colours was different than the majority of his friends (but the same as his brother's), and for many years the condition was known as Daltonism. A more recent study of his careful observations suggests that he was unable to distinguish the colour red. It is of interest to learn that Dalton himself felt that he possessed some visual advantages over his friends because of the nature of the abnormal sensitivity of his eyesight. He did not find that he was at a disadvantage at all.

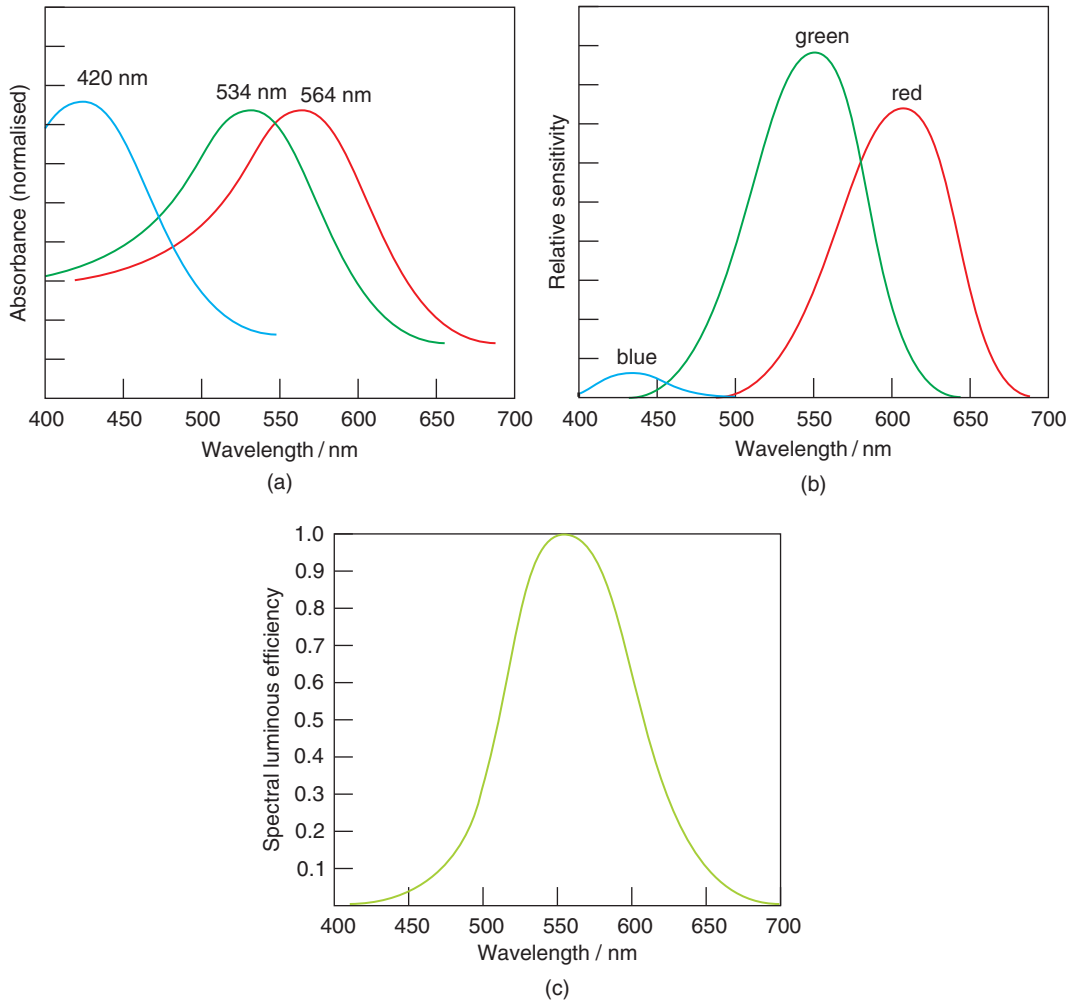


Figure 1.12 ***The sensitivity of the eye to light, schematic: (a) sensitivity of the cone cells in a normal eye to light as a function of the wavelength; (b) visual sensitivity of a normal eye to red, green, and blue light; (c) overall visual sensitivity of a normal eye to light***

Trichromaticity, considered normal in humans, is fairly common among old-world primates, but most mammals can only detect two colours and are referred to as dichromats. Many insects, including butterflies and beetles and some birds, have a trichromatic colour detection system that is able to see ultraviolet, blue, and green, with detection peaks of sensitivity at approximately 350, 450, and 550 nm, but are unable to record red. There are also numerous tetrachromic animals, including most birds, lizards, turtles, and crocodiles that have four different cone cell receptors. These can detect ultraviolet light with a peak wavelength at approximately 370 nm in addition to three colours with absorption peaks at approximately 445, 508, and 565 nm, and so have a very increased richness in vision compared to humans. Some species of butterfly are believed to be able to detect up to eight wavelength ranges, but the record for vision seems to be held by certain species of mantis shrimps, which have 12 different photoreceptors and can detect wavelengths from 300 nm in the ultraviolet to 720 nm deep red.

The physiological response of the eye–brain combination arises when light waves fall upon the light sensitive *retina*, which makes up the inner surface of the eye. The light that is poorly absorbed by the rod cells, red and blue/violet, gives rise to the red-purple colour of the membrane. In 1876, Boll reported that this red-purple pigment bleached in the presence of light to a colourless form. The change was found to be reversible, and in the dark the purple colour was regenerated. This important photochromic reaction is the source of vision. The compound involved became known as 'visual purple' and is now called *rhodopsin*.

When light photons impinge on both rod and cone cells, they are absorbed by stacks of photoreceptor molecules, which are bleached in the process. This sends a nerve impulse to the brain. The system is remarkably sensitive and there is considerable evidence to suggest that in the rod cells just one photon is enough to stimulate the nerve. The light absorbing pigments consist of a protein, an *opsin*, bound to a light-absorbing molecule, *retinal*. The receptor in the rod cells is the rhodopsin mentioned above, while those in the cone cells are called *cone opsins*. The opsin part of the receptor, consisting of 364 amino acid residues in humans, is arranged in the form of seven helices, which penetrate the cell wall and enclose the retinal, which is bound to the amino acid lysine 296 (Figure 1.13). The opsin proteins differ from one cone cell to another and from rhodopsin in the rods, and it is these differences that confer the differing sensitivities to the receptors. However, the differences are rather small. For example, the amino acid sequences in the green (M) and red (L) cone receptors in humans differ in only three of the amino acid residues in 364.

In humans, retinal is derived from the compound β-carotene, an orange pigment found in carrots. This is transformed into vitamin A in the liver, which then forms retinal. The visual pigments in animals also consist of retinal plus opsin. In fact, there are two forms of vitamin A: A_1, which gives $retinal_1$ (11-*cis*-retinal, the aldehyde of vitamin A_1), and A_2 which gives $retinal_2$ (3-dehydro-retinal). $Retinal_1$ is used by all mammals and birds and will just be referred to as retinal in what follows.

The framework of the processes triggering vision is well established. It is described here with respect to rod cells, which have been studied in most detail. The chromophore of rhodopsin is the *cis*- form of the molecule *retinal*, 11-*cis*-retinal (Figure 1.14a). This *cis*-retinal molecule is bound to the opsin via the amino acid lysine, to form rhodopsin (Figure 1.14b). The *cis*-retinal by itself is not coloured and has an absorption maximum between 370 and 380 nm. However, when joined to the opsin the absorption maximum moves to about 500 nm. Molecules that can cause the deepening of the colour of a chromophore are called *bathochromes,* and the resultant movement of the absorption maximum is referred to as a *bathochromic shift*. The bathochromic shift comes about because of the particular conformation of the *cis*-retinal molecule in conjunction with the protein. The bonding and slight differences in the various forms of the opsin molecules produces different bathochromic shifts and hence makes the cones sensitive to the different wavelengths of red, green, and blue light.

The molecular mechanism leading to the nerve impulse hinges on the fact that retinal can exist in two isomeric forms, the *cis*- form already described and a *trans*- form, called *all-trans*-retinal. Under the

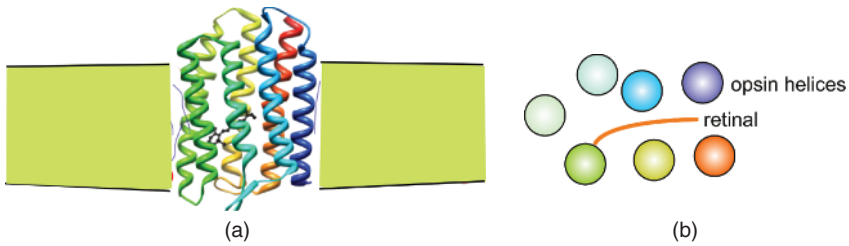


Figure 1.13 ***(a) The schematic structure of an opsin protein in the cell wall of a photoreceptor; (b) The opsin protein molecule is in the form of seven helices arranged to enclose a retinal molecule***

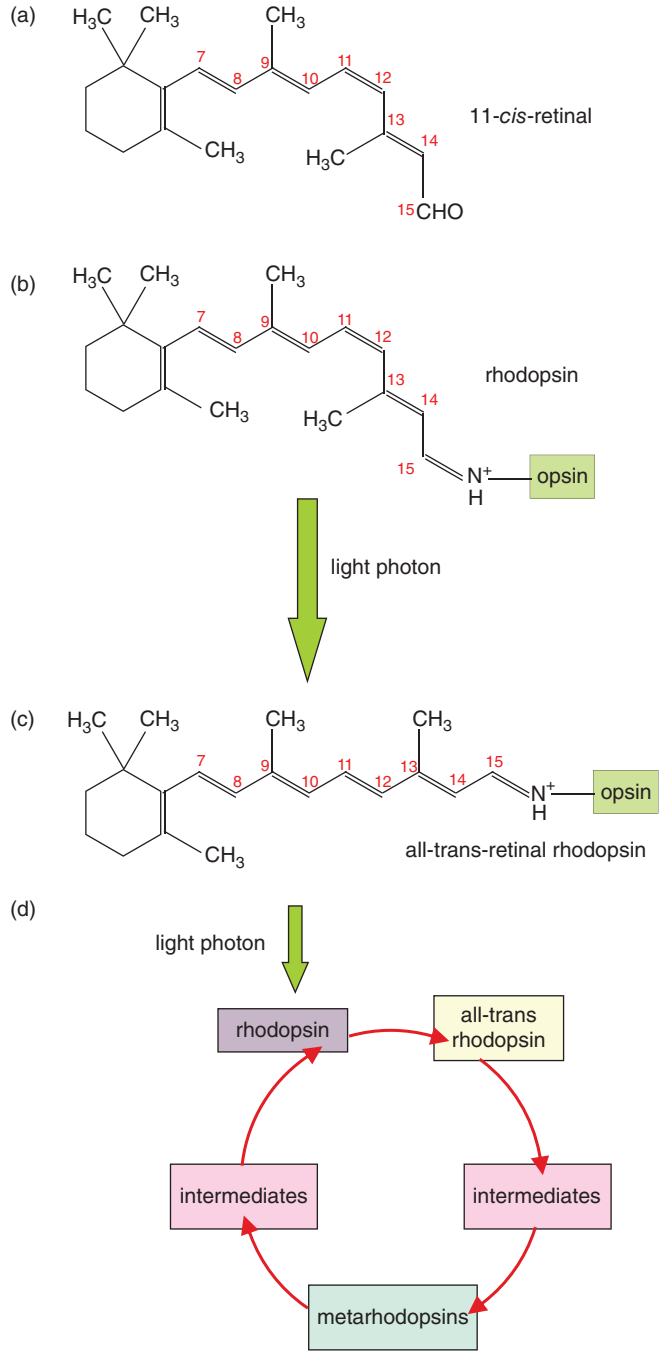


Figure 1.14 ***The structures of (a) 11-cis-retinal: (b) rhodopsin; (c) all-trans-retinal rhodopsin, produced by the action of light on (b); (d) cycle of chemical changes producing vision***

influence of a photon the *cis*-retinal molecule changes to *all-trans*-retinal rhodopsin, (Figure 1.14c). Absorption of light by rhodopsin drives the molecule through several intermediates to the bleached state, which can consist of a number of different molecules, metarhodopsin I, metarhodopsin II, and so on, depending on the conditions experienced. Thereafter, the reaction reverses, again passing through a number of intermediates, so that the *trans*-retinal readopts the *cis*- conformation and reforms rhodopsin (Figure 1.14d). Another photon can trigger the cycle again. Each cycle takes only a fraction of a second and can repeat indefinitely in normal light conditions so as to send a stream of nerve impulses to the brain. These impulses end when the light is extinguished and all molecules revert to rhodopsin.

It is worth commenting on the enormous complexity of vision. The description of the cycle occurring in rod cells and presumed to occur in the cone cells described above is only true at moderate light intensities. At lower light intensities, the *trans*-retinal molecule in rod cells is released completely from the opsin. Two processes then operate, depending on the weakness of the light signal. At the 'higher' of these lower intensities, the *trans*-retinal is transformed back to the *cis*-conformation by the action of enzymes in the eye itself, whereupon the molecule is reattached to the opsin. At the lowest light intensities, the *trans*-molecules actually leave the eye completely, enter the bloodstream, and are reprocessed to the *cis*-form in the liver, an occurrence that contributes to the length of time it takes to become fully 'dark adapted'.

Rhodopsin has another role to play in the broader picture of life. In 1971, it was found that some purple halobacteria, bacteria which inhabit very salty environments, are coloured purple by a version of rhodopsin called bacteriorhodopsin. This consists of 247 amino acid residues, arranged in seven helices, with the photoactive retinal attached to lysine 216. It is, however, not used for vision, but as an analogue to chlorophyll in plants. Absorption of light by chlorophyll initiates a chain of electron transfer reactions that eventually provide the energy for plant growth. In the purple halobacteria, the bacteriorhodopsin converts photons of green sunlight into energy for the metabolism of the bacterium. In essence it appears that the *cis*- / *trans*- change acts as a proton pump that is activated by a photon, and the resulting electrochemical potential created transports ions across the cell membrane and so initiates the energy-building steps.

Other members of this family of light-activated molecular pumps have since been identified. In the bacterium *Natronomonas pharaonis,* which is found in high-salt-concentration lakes in Egypt and Kenya, the active rhodopsin, NpHR halorhodopsin, transports chloride ions (Cl^-) following stimulation by a photon of light. The response is found for photons of wavelength between 450 and 600 nm, with a peak in the yellow at 589 nm.

In 2002, another family of proteins that acted as light activated ion pumps were discovered – the channelrhodopsins. The single-celled alga, *Chlamydomonas reinhardtii* contains two such proteins, channelrhodopsin-1 (ChR1) and channelrhodopsin-2 (ChR2), which act as single-protein membrane channels activated by blue light. The ion pump ChR2, for example, absorbs photons with wavelengths from about 410–530 nm, with a peak at 470 nm, thus absorbing in the blue region of the spectrum. It serves to pump sodium ions (Na^+) across the cell membrane ions. In 2008, a third light-activated channel rhodopsin (VChR1) was isolated in the multicellular alga *Volvox carteri*, commonly found in lakes, ponds, and water-filled ditches. This protein responds to photons of wavelengths 450–600 nm, with a maximum response in the green (535 nm) and yellow (589 nm).

An attribute of light that is not very important for day to day vision in humans is polarisation. Whilst the human eye is unable to detect polarisation direction,[7] many insects have this ability and are called *polarotactic* insects. The capability arises because the molecule responsible for photoreception in

[7]In fact, this is not altogether correct. The visual phenomenon called Heidinger's brushes, a faint small hourglass shape centred in the field of view, is caused by the detection of polarised light in the retina of the eye. It is not observed by everyone.

the eyes of all animals, rhodopsin, is a dipolar molecule with an optical axis. These molecules absorb polarised light energy maximally when the direction of polarisation is parallel to the optical axis of the molecule. In insects' eyes these molecules are aligned in a fixed direction, making them polarisation sensitive. In humans, the molecules are free to rotate, so that the orientation of the optical axis is random and polarisation perception is lost.

The polarisation of sky light is detected in the eyes of many migratory butterflies (*Lepidoptera*) such as the painted lady (*Cynthea cardui*) and the monarch (*Danaus plexippus*) and is used in their internal 'compass'. The African scarab beetle *Scarabaeus zambesianus*, is able to detect the polarisation of moonlight and uses this to navigate.

1.10 Colour Perception

Recognition of colour is a function not only of the physical make-up of the light falling on the eye and physiological factors but also of psychological biases. The 'colour' of an object in this sense is changed by factors such as surface roughness or texture. Indeed, the perception of surface colour in itself is complex. In addition, subsurface scattering, which returns some incident light on a body to the exterior, is of importance in the appearance of many objects, such as skin, cosmetics, and paint. Because of this interplay, it is possible to distinguish a hard red plastic surface from a red velvet surface, even though in terms of physics, the colours of both may be identical, originating in the same dye or pigment. It is clear that when describing the appearance of an object in colour terms, it is necessary to consider specular (mirror-like) reflection, diffuse (non-mirror-like) reflection and subsurface scattering, as well as the make-up of the light that is reflected or scattered. Moreover, human eyes vary in colour-interpreting ability. It appears that an average person can distinguish more than a million different colours. All of these aspects are implied when the colour of an object or a light source is mentioned in a colloquial way. Because of this, colour is difficult to quantify.

Despite the complexity inherent in the concept of colour and its perception, it has been found that all colours can be precisely specified by three parameters. Colours can then conveniently be represented by points in a three-dimensional coordinate system. There are many diagrammatic ways of representing the three attributes, and these are called *colour spaces*. The way in which the coordinates of any colour in the colour space are derived is called a *colour model*. There are many colour models, of which only three will be described briefly here.

One widely used colour model takes as initial parameters the three attributes *hue*, *saturation,* and *brightness* to give the *HSB model*. These characteristics are generally described as follows:

1. *Hue*, which corresponds to the wavelength or frequency of the radiation. The hue is given a colour name such as red or yellow.
2. *Saturation* or *chroma*, which corresponds to the amount of white light mixed in with the hue and allows pale, 'washed out' colours to be described.
3. *Brightness, lightness, luminance*, or *value*. These terms describe the amount of light, that is, the number of photons reaching the eye.

This model is also given the acronyms *HVC* (hue, value, chroma), *HSL* (hue, saturation, luminance), *HIS* (hue, intensity, saturation) and *HCL* (hue, chroma, luminance). One way of building a colour space in terms of this colour model is to arrange the hue around the periphery of a disc with the degree of saturation of the colour represented by the distance from the centre of the disc along the radius. Brightness is defined by an axis perpendicular to the centre of the disc (Figure 1.15a). This arrangement has been quantified in constructions such as the *Munsell colour cylinder* or *Munsell colour solid* (Figure 1.15b).

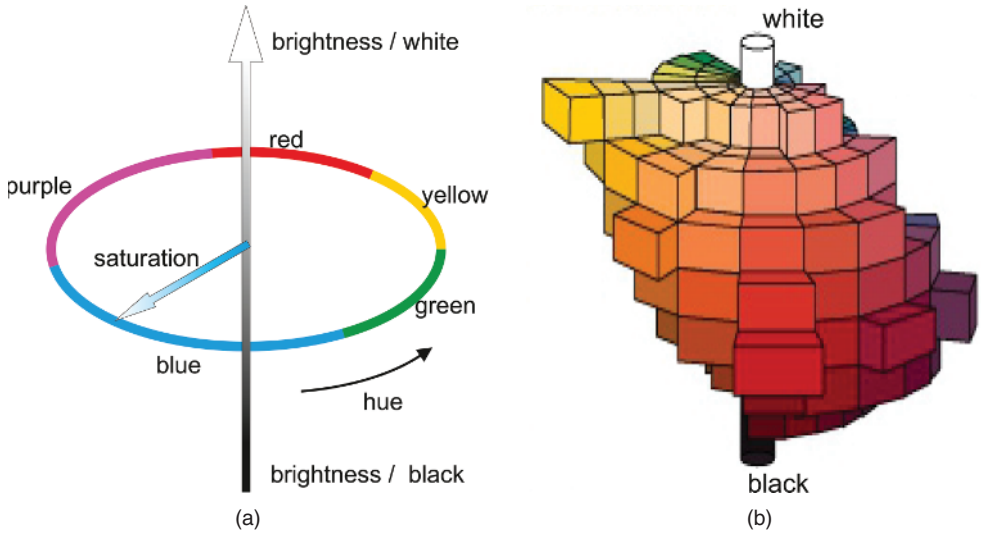


Figure 1.15 ***The representation of colours in the HSB colour model: (a) hue is given by a point on the circumference of a planar disc, saturation by the distance along the radius from the centre of the disc and lightness by the vertical axis; (b) the solid representation of the colours on the Munsell colour cylinder***

1.11 Additive Coloration

Additive colour mixing occurs when two or more beams of differently coloured light combine (i.e. overlap on a perfectly white surface, or arrive at the eye simultaneously). Colours on displays are produced by additive coloration as the screen is composed of small dots of three different light emitting substances each of which shines with one of three colours when activated. Additive coloration is also used in the painting technique known as pointillism. In this method of painting, the image is built up by placing small dots of relatively saturated colour onto the canvas, making sure that they do not overlap. When viewed from a distance of a few metres such pictures appear bright and dynamic.

The colour patterns on the wings of many butterflies and moths are produced in a similar way. The wings are tiled with a fine mosaic of scales, each of which reflects only one colour. The colour perceived by the eye is an additive colour arising from the numerous closely spaced scales. The range of colours that can be produced by rather a few basic pigments is remarkable. For example, some perceived purples arise from mixtures of black, white, and red scales, while some greens arise from mixtures of yellow and black scales.

It has been found that the majority of additive colours can be produced by mixing just three *additive primary colours*, red, green, and blue. (Strictly speaking, any fairly monochromatic intervals near to these colours will suffice.) Moreover, mixing equal quantities of these three primary colour lights will produce white light. There are a number of ways of quantifying the amounts of each primary colour light mixed, which are represented by the values r of the *red* component, g of the *green* component, and b of the *blue* component, thus:

$$\text{colour} = r + g + b$$

Use of these additive three primaries is called the *RGB colour model*.

A simple colour space can be constructed by using Cartesian axes to represent the amount of the three primary colours, red, green, and blue, while the diagonal represents the transformation from

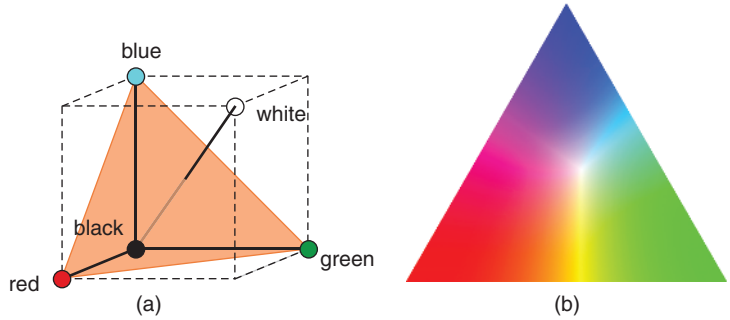


Figure 1.16 ***(a) RGB colours represented by Cartesian axes, with black to white along the body diagonal; (b) a section of (a) taken normal to the body diagonal passing through red, green, and blue corners of the cube forms a colour triangle***

black to white (Figure 1.16a). Sections through this colour space allow representation of colours by a planar figure. Such representations are called *chromaticity diagrams*. A simple example is given by taking the triangular sheet running diagonally through the cube normal to the black–white diagonal which cuts the corners of the cube that represent pure red, green, and blue. This produces a *colour triangle* (Figure 1.16b). Other colours can be specified by coordinates in the plane of the colour triangle. The location given by the coordinates corresponds to the amounts r, g, and b making up the colour. The coordinates that specify the centre of the colour triangle represent the case when the three primary colours are mixed in equal amounts. This will be a shade of grey, but is usually represented by the colour white. The range of available colours that can be obtained by mixing lights corresponding to the three vertices is the *gamut* of colours available. These diagrams represent hue and saturation but not lightness (i.e. the grey tone), which must still be added as a third axis perpendicular to the chromaticity diagram if this information has to be displayed.

The study of light mixing has been quantified by the Commission Internationale de l'Eclairage, (CIE), which has, on a number of occasions, refined the rather simple colour triangle concept so as to allow colour perceptions to be more accurately characterised. A colour is specified by a pair of x- and y-coordinates, which are derived from the r, g and b values noted above by the application of a standardised set of equations. In this representation, the triangular shape has been distorted into an outline something like a parabola, depending on the way in which the x- and y-axes are plotted. A commonly encountered form of the CIE chromaticity diagram is that first proposed in 1931 (Figure 1.18a). The spectral colours, running from red to violet, are arranged around the outer edge of the shape and colours not seen in the spectrum, the purples, and browns, are found to lie between the red and violet ends of the curve. Points within the area of the diagram represent colours formed by the additive mixing of light values. The colours are *fully saturated* along the outer edge of the curve and become less and less saturated as the centre of the diagram is approached. Standard daylight white is represented by a point close to the coordinates $x = y = 0.33$, shown as W in Figure 1.17b.

If a straight line is drawn through the point W and extended to the boundaries of the curve, the pair of colours reached, when mixed, will give white light. For example (see Figure 1.17b), a line connects the colours red, of wavelength 700 nm, and cyan (blue-green), of wavelength 492 nm, and passes through the point W. The proportions of the end colours red and cyan-green light needed to produce white light is given by the *lever rule*, (Figure 1.17c):

$$\text{amount of red light} = r/(r + c)$$

$$\text{amount of cyan light} = c/(r + c)$$

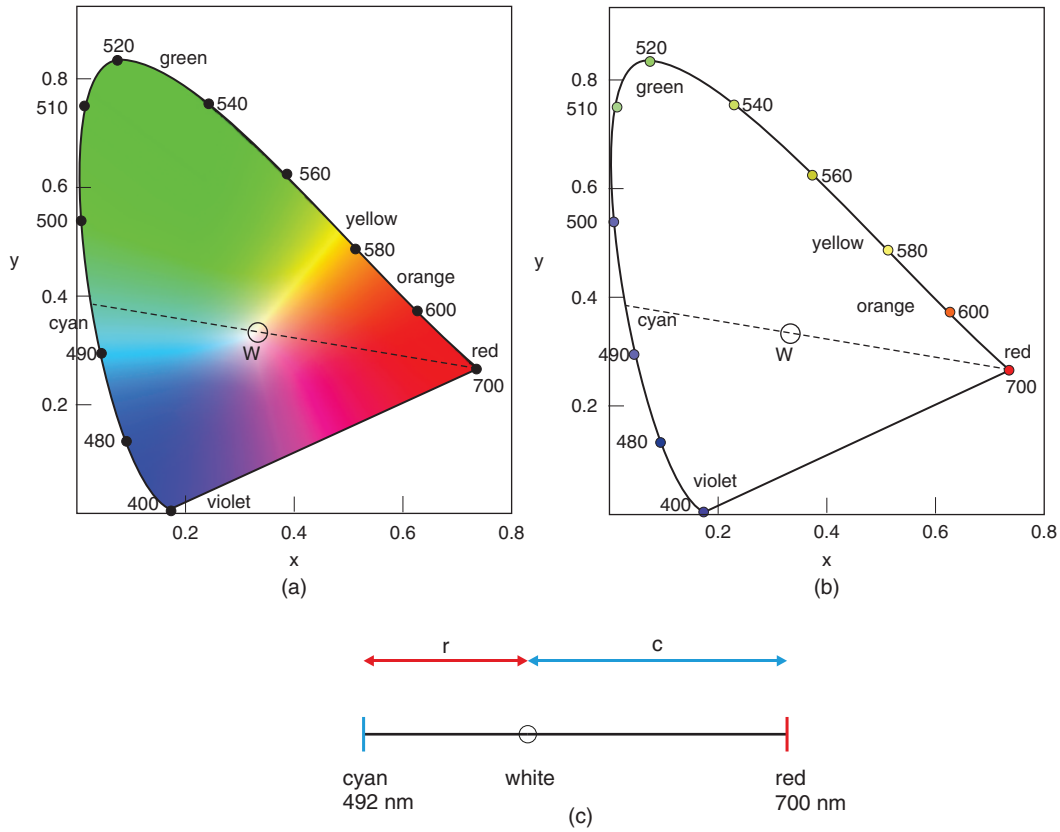


Figure 1.17 ***The CIE 1931 chromaticity diagram: (a) The figures marked around the outer edge of the curve denote the wavelength of the colour; (b) A straight line through white (W) links two complementary colours on the periphery; (c) The lever rule gives the proportions of complementary colours which are needed to create white light***

Measurement shows that mixing red of wavelength 700 nm and cyan light of wavelength 492 nm in the proportions 39% red to 61% blue will produce white light. The colours at the ends of a line through the point W are called a *complementary pair* of colours. If one of these colours is subtracted from white light, the colour remaining is called the *complementary colour* to the first.

As with the colour triangle, all planar chromaticity diagrams represent hue and saturation but not the exact value of lightness, which must still be added as a third axis perpendicular to the chromaticity diagram if this information has to be displayed. In general terms, therefore, the white region on the chromaticity diagram should be represented by grey, with white and black being extremes on the vertical axis perpendicular to the plane of the figure.

The accurate rendition of additive colouration is of prime importance in displays. Additive colouration and the interconvertion between various colour models are most easily explored using a computer which has photography or drawing editing software installed. On most of these packages, seven or eight or so different colour models are available, including RGB and at least one CIE model. The coordinates of any colour are given and comparisons between several systems are rapidly made.

The confusion that colour blindness can cause is easily understood in terms of a chromaticity diagram. For example, Dalton had a lack of red receptors (see footnote 6). The CIE 1931 chromaticity

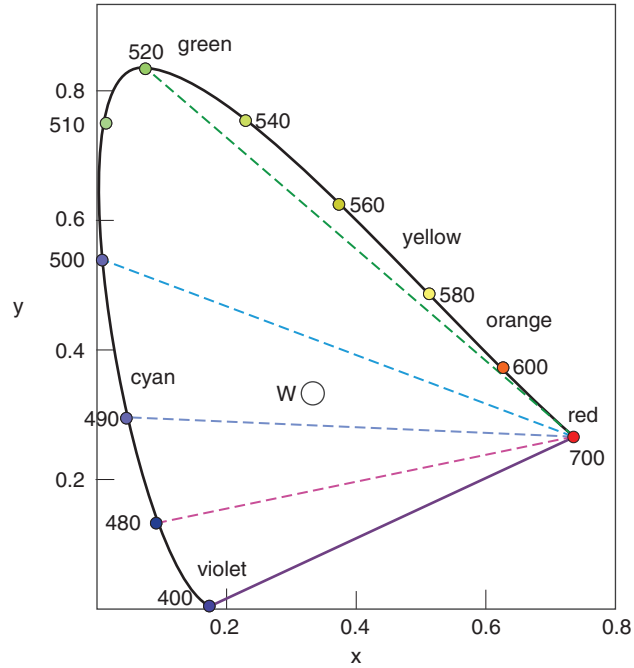


Figure 1.18 ***The dashed lines represent the loci of colour confusion for a person with red defective vision plotted on the CIE 1931 chromaticity diagram***

diagram can be used to illustrate the difficulties this causes. Any colour formed by mixing red with another colour, C, around the periphery of the curve will not be differentiated from any other colour along the line joining red to C. These lines show the loci of *colour confusion* (Figure 1.18). Other types of colour blindness will lead to other loci of colour confusion.

1.12 Subtractive Coloration

Subtractive coloration occurs when some wavelengths of light are *removed* from the incident beam of (generally white) light by absorption, scattering, or reflection. The colours perceived by the eye in this process are said to be due to *subtractive colour mixing*. A common way to produce subtractive colours is by selective absorption. This method has been used for centuries to produce beautiful 'stained glass' images. When daylight illuminates a stained glass window, some wavelengths are absorbed by ingredients in the glass. An observer inside the building sees only the residual colours which are not absorbed and pass through the material (Figure 1.19). Colour filters act in the same way and absorb some wavelengths strongly and transmit the remainder. Figure 1.20a shows the fraction of light transmitted as a function of wavelength for a commercial glass colour filter. The filter absorbs red light strongly and transmits violet and blue-green light, (Figure 1.20b). If the filter is held up to the light it will look blue-green. When it is viewed in reflected light is appears dark, as red light is absorbed and blue-green passes through the glass. For the same reason stained glass windows in medieval churches look impressive when viewed inside the building, with light transmitted through the glass, yet often look dull when viewed from outside the building in reflected light.

By analogy with additive coloration, one would expect to be able to combine three *subtractive primary colours* to produce the whole range of subtractive colours. These subtractive primary colours

Figure 1.19 ***Mediaeval stained-glass window in Gloucester Cathedral, viewed from inside the building. Source: Reproduced with permission, Gloucester Cathedral, Gloucester, UK: www.gloucestercathedral .org.uk.***

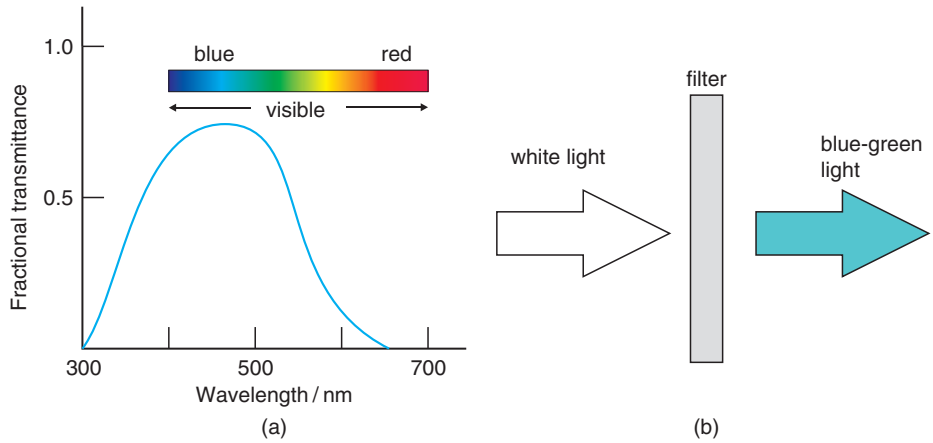


Figure 1.20 ***(a) The fractional transmittance of a commercial blue colour filter. About 3/4 of the blue light incident on the filter is transmitted but most red light is absorbed; (b) When the filter is viewed in transmitted white light it will appear blue-green***

are *cyan*, which is red absorbing and transmits/reflects blue and green, *magenta*, which absorbs green and transmits/reflects blue and red and *yellow*, which is blue absorbing and which transmits/reflects green and red. If the three subtractive primaries are mixed in equal amounts we obtain *black*, as one primary will absorb red, one will absorb green and one absorb blue, thus removing the whole of the visible spectrum. Colour construction using these three subtractive primary colours is described as employing the *CMY model*, where the letters simply represent the initial letters of the colourants.

If the wavelength range of light absorbed is rather small, the colour remaining is called the *complementary* colour to that absorbed (Table 1.6). It is seen that the additive and subtractive primary colours are complementary colours.

Table 1.6 ***Complementary colours***

Wavelength/nm	Colour absorbed	Complementary colour
400–435	Violet	Yellow-green
435–480	Blue[a]	Yellow[b]
480–490	Blue-green	Orange
490–500	Green-blue	Red
500–560	Green[a]	Magenta[b]
560–580	Yellow-green	Violet
580–595	Yellow	Blue
595–605	Orange	Blue-green
605–700	Red[a]	Cyan[b]

[a]Additive primary colours.
[b]Subtractive primary colours.

Colour printers use cyan, yellow, and magenta dyes to produce the coloured images. These dyes are deposited upon white paper, and absorb the appropriate subtractive primary colour. White light reflected from the dyes is depleted in these colours and yields the appropriate toned image by subtractive coloration. Although the overlap of cyan, yellow, and magenta produces black, this tone is often not dark enough for many representations. Printers therefore often add black to the trio. This system of colour production is known as the *CYMK model* of colour formation, where the letter K stands for the black component. Although these four colours are satisfactory for many colour printing applications, more hues, intermediate between the CYMK set, are used to obtain more accurate colour rendition, in, for example, high-quality art reproductions.

1.13 The Interaction of Light with a Material: Appearance

The processes of reflection, scattering, and absorption give rise to the world of colour around us. The blue sky on a clear day is due to scattering of light by the atmosphere. The appearance of water bodies is often due to reflection of the sky, blue in fine weather, grey or dark during storms. The 'white horses' on waves are coloured by scattering of light by the many small droplets of water present that make up the foam, exactly analogous to opal glass. Plants, animals, soil, and rocks are predominantly coloured by absorption.

Broadly speaking, if the energy of the scattered wave/photon is the same as that of the incident wave/photon the scattering is called *elastic* scattering, and otherwise *inelastic* scattering. Elastic scattering from a surface is normally called *reflection*, and elastic scattering into a transparent solid is called *refraction*. Elastic scattering from ordered collections of small scattering centres is called *diffraction*. For historical reasons, the term scattering itself, especially elastic scattering, is usually reserved for the interaction of light with randomly distributed small particles at relatively low density.

1.13.1 Reflection

The appearance of an object will rest on reflection from the surfaces present, and, for a transparent solid, by refraction into the material. The reflectivity of the surface will depend on surface roughness and texture. If the surface is smooth, the reflection is said to be *specular* or mirror-like. A fine beam is reflected as a beam such that the angle of incidence is equal to the angle of reflection (Figure 1.21a). If the surface is rough or made up of many small particles bonded together, the reflection is partly

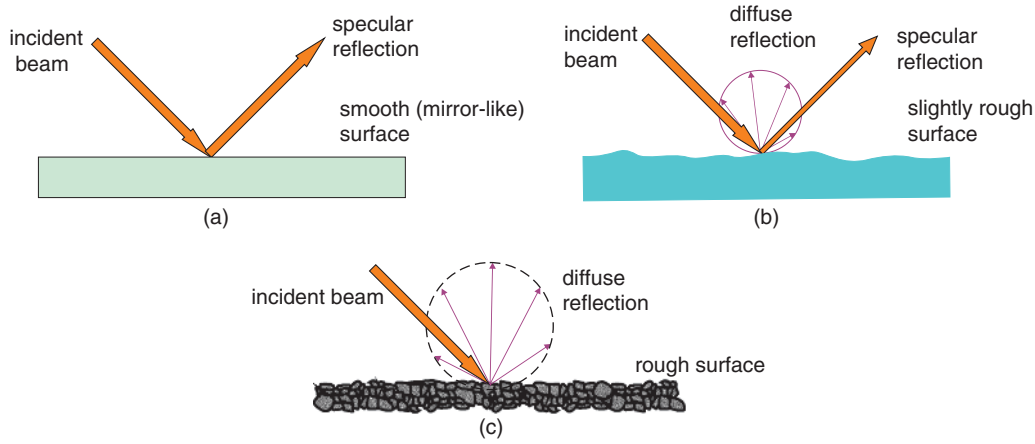


Figure 1.21 ***(a) Reflection of light from (a) a smooth surface; (b) an uneven surface; (c) a rough surface made up of particles***

made up of a specular component as well as a *diffuse* component (Figure 1.21b). The diffuse reflection component increases with surface roughness at the expense of the specular component, so that a finely ground powder shows only diffuse reflection (Figure 1.21c). The *gloss* of a surface is a measure of the relative amounts of diffuse to specular reflection. Glossy surfaces have a large specular component.

1.13.2 Diffuse reflectance

The absorption spectrum of a material gives important information about structure and available energy levels. This information can be extracted from the transmitted light from a semi-transparent solid as the absorption peaks will be absent from the recorded spectrum. However, for powders or films that are essentially opaque, the same information can be retrieved by measuring the diffuse reflectance spectrum of the material. This method is also involved when subtractive coloration is to be quantified.

The diffuse reflectance of an opaque film to a monochromatic light source which contains both scattering and absorption centres is given by the Kubelka-Munk (sometimes Schuster-Kubelka-Munk) function:

$$\mathrm{f}(R) = (1 - R)^2/(2\,R) = k/s$$

R = total reflectance
k = absorption coefficient
s = scattering coefficient

The Kubelka-Munk function applies when the particles that comprise the sample are much smaller than the thickness of the layer under scrutiny, the absorbing and scattering centres are uniformly distributed, and the layer examined is thick enough that no light escapes through the back of the sample, i.e. in effect the sample is infinitely thick for optical purposes. The Kubelka-Munk function was originally devised to account for the reflectance of paints but is now widely used for paper and finely powdered samples such as catalysts or pigments. The formula is quadratic and can be expanded as:

$$R_\lambda = 1 + k/s - \sqrt{[(k/s)^2 + (2\mathrm{k/s})]}$$

or

$$R_\lambda \approx 1 - \sqrt{(2k/s)}$$

R_λ = reflectivity at wavelength λ

For a mixture of pigments, the scattering and absorption coefficients can be written in terms of the respective concentrations as:

$$k = c_1 k_1 + c_2 k_2 \ldots$$

$$s = c_1 s_1 + c_2 s_2 \ldots$$

1.13.3 Elastic scattering

Elastic scattering occurs when light interacts with randomly distributed small particles at relatively low density in a gas, liquid, or solid, or when light meets with internal boundaries within transparent solids. For many ideally transparent solids such internal surfaces are a major cause of loss of transparency. Glass, the best known of transparent solids, is, in effect, in the liquid state, and no internal boundaries occur. However, it is relatively simple to cause a glass to crystallise, and the resulting tiny crystallites act as scattering centres. The resulting scattering, which can contain diffuse and specular components, renders the material nontransparent. The resulting material transmits a certain amount of the incident light and scatters the remainder away from the incident beam (Figure 1.22). Such materials are termed *translucent*. (Scattering is described in Chapter 5.)

Translucency is a desirable property of fine porcelain, which consists of crystallites of mullite ($\sim Al_6Si_2O_{13}$) dispersed throughout a glassy matrix. An important property of early Chinese porcelain was this translucency, which European manufacturers found impossible to reproduce until the secrets of the Chinese process was discovered and relayed to Europe, in an early case of industrial espionage.

Opaque glasses, such as *opal* glasses, are deliberately made with large numbers of scattering centres present. The resultant scattering renders the material white because the scattering affects all wavelengths of the incident light equally. Note that the light reflected from the upper surface as well as that emerging from a translucent material will contain a diffuse and specular component (Figure 1.23).

Most plastics as fabricated are noncrystalline and have no internal boundaries, rendering them transparent. If these contain impurities, inhomogeneities or polymer crystallites light is scattered and they take on a slightly milky appearance.

Nonglassy solids are mainly composed of polycrystalline aggregates or ‘grains’. The grain boundaries between each crystallite scatter light, and any impurity phases that exist in the matrix or the grain

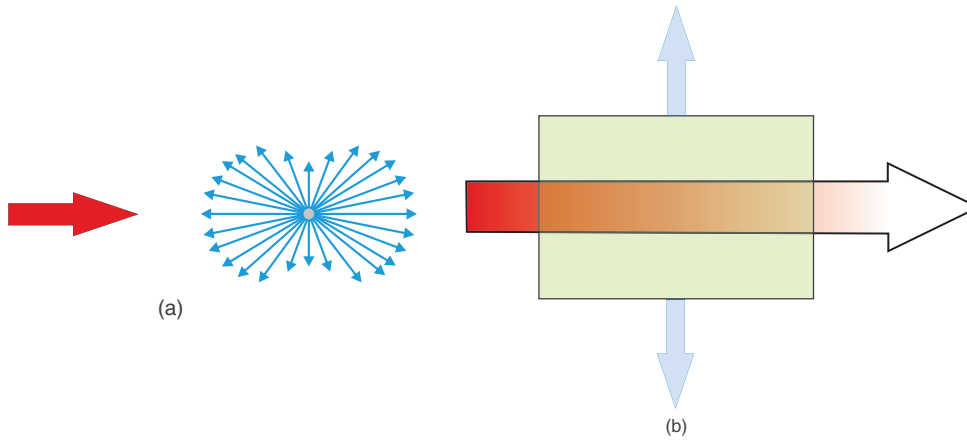


Figure 1.22 ***Scattering (schematic; (a) Scattering from a single particle; (b) the passage of light through a translucent material containing many scattering centres***

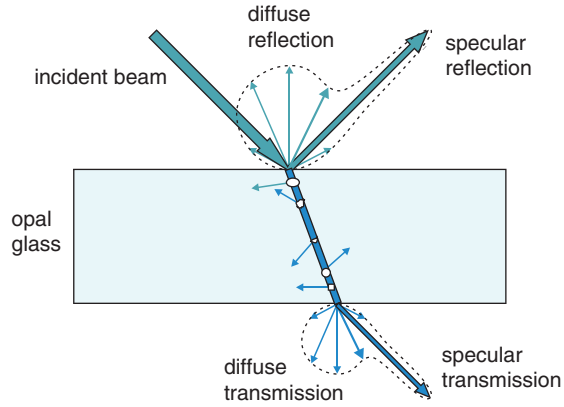


Figure 1.23 ***Scattering and reflection by a semi-transparent opal glass (schematic)***

boundary regions enhance this effect, so that polycrystalline solids are invariably opaque. However, it is of considerable benefit to make such materials transparent. This can be achieved by careful processing that achieves a high density, so that internal pores and bubbles of gas are eliminated, and a solid composed of small, evenly sized crystallites with no impurity grain boundary phases present is produced. In this way, transparent refractory ceramics such as alumina (Al_2O_3), aluminium oxynitride ($\sim Al_{23}O_{27}N_5$), and SiAlONs (phases occurring in the $SiO_2 - Al_2O_3 - Si_4N_3 - AlN$ system) have been produced. These and similar materials have uses as lamp housings and windows that need to be stable in air to temperatures of 2000 °C or more. In addition, these are hard and durable ceramics, and are favoured for applications such as specialist optical windows and domes where resistance to abrasion and erosion are important selection criteria.

As well as diffuse reflection, subsurface scattering is of considerable importance in modifying the appearance of a surface. This is particularly so when the surface is composed of layers with different optical properties, such as skin. Controlling these forms of scattering and reflection are of great importance to the cosmetics industry and imitating them is vital to both artists and allied personnel involved in the representation of skin tones in computer generated images.

1.13.4 Inelastic scattering

Inelastic scattering arises when some energy, usually a small amount, is transferred from the incident photons to another component of the material. The major forms of inelastic scattering are Raman, Brillouin, and Compton scattering. Raman scattering is widely used as a spectroscopic tool in the investigation of molecular structure. In this, a photon interacts with the vibration or rotation of a molecule, either giving up or gaining some energy. The energy difference between the scattered photon and the incident photon then gives details about the relevant energy level spacing in the molecule. Brillouin scattering involves the interaction of photons with varying wave-like deformations of the solid structure, typically vibrational deformations characterised by phonons. Again, there is an exchange of energy that can be analysed to give the phonon spectrum of the solid. Compton scattering occurs when a high-energy photon, typically an X-ray or gamma ray, is scattered by electrons in the solid, losing energy in the process while simultaneously imparting momentum to the electron.

1.13.5 Absorption

Absorption is generally the term reserved for use when almost all of the incident radiation is taken up by the material as opposed to inelastic scattering when the changes are rather small. In the absence

of absorption centres (or scattering, of course), a pure material will be transparent. The absorption properties of a material vary over the electromagnetic spectrum and no material is transparent over all wavelength ranges. Silicon, for example appears metallic over the visible spectrum but is transparent to infra-red wavelengths. Absorption is the root cause of subtractive coloration, as described above. During absorption, the photon energy excites the electrons on the component atoms or molecules that constitute the absorption centres into higher-energy levels. Often, the absorbed energy is manifested as a rise in temperature of the body, but on occasion some of this energy might be re-emitted from the absorption centres as light of lower energy, giving rise to fluorescence or luminescence.

1.13.6 Attenuation

As a beam of light passes through a transparent or semi-transparent material, it gradually loses intensity, a process generally called *attenuation* (formerly *extinction*). Attenuation is due to the interaction of light with the constituents present in two basic ways – scattering or absorption, described above. When attenuation takes place in a homogeneous solid, the amount of light transmitted by a semi-transparent plate of thickness x is given by:

$$I_x = I_o \exp(-\alpha_e x) \tag{1.3}$$

where I_x is the irradiance leaving the plate,[8] I_o is the incident irradiance, and α_e is the (Napierian) *linear attenuation coefficient* (formerly *extinction coefficient*). Equation (1.3) is known as *Lambert's law* or *Beer's law*, although it was first clearly set out by Bouguer and should, by rights, be called Bouguer's law. The *attenuation length* is defined as $1/\alpha_e$. The amount of light removed from the beam is thus:

$$I_r = I_o - I_x = I_o - I_o \exp(-\alpha_e x) = I_o [1 - \exp(-\alpha_e x)]$$

If the attenuation of the beam is solely due to absorption, the attenuation coefficient is replaced by the (*Napierian*) *linear absorption coefficient*, α_a. Similarly, if the attenuation is solely due to scattering the attenuation coefficient is replaced by the (Napierian) linear *scattering* coefficient, α_s. For nonhomogeneous solids, these coefficients may vary with direction. Note that the degree of attenuation will vary significantly across the spectrum and the attenuation coefficient is not a constant.

It is sometimes convenient, as when discussing the absorption of X-rays, to define a *mass absorption coefficient* μ, which describes the decrease in transmitted irradiance through a homogeneous material of density ρ and thickness x:

$$I_x = I_o \exp(-\mu \rho x)$$

in this case,

$$\mu = \alpha_e / \rho$$

Attenuation is often associated with the presence of chemical or physical centres, which may be atoms, molecules or larger particles, distributed throughout the bulk of a material. In the case of the mass absorption coefficient described above these are the totalities of the atoms that make up the material itself. In this case, if the atoms in the material are supposed to absorb radiation independently of each other, the mass absorption coefficient of the phase is simply related to the weight

[8] The symbol I is used for irradiance instead of E to avoid confusion with the use of E for energy throughout this book. See also Appendix A.

fraction of each atom species present. Thus, the mass absorption coefficient of a material M with a formula $A_xB_yC_z$ is:

$$\mu_M = (\text{weight fraction A})\,(\mu_A) + (\text{weight fraction B})\,(\mu_B) + (\text{weight fraction C})\,(\mu_C)$$

The weight fraction of each species is given by:

$$\begin{aligned}\text{weight fraction A} &= \text{mass of A present/total mass}\\ &= x(m_A)/[x(m_A) + y(m_B) + z(m_C)]\end{aligned}$$

and so on, where m_A is the molar mass of species A, m_B is the molar mass of species B, and m_C is the molar mass of species C.

More often, attenuation is associated with a dilute concentration of centres distributed throughout the bulk phase. In this case, the degree of attenuation is often taken to be a function of the concentration of these centres. This is taken into account by writing the Beer–Lambert–Bouguer law as:

$$\log (I_x/I_o) = -\varepsilon\, c\, x$$

where I_x is the irradiance after passage through a length of sample x, I_o is the incident irradiance and c is the *molar concentration* of the active centres. The quantity ε is called the *molar (decadic) attenuation coefficient*. The attenuation coefficient has units of area and can therefore be regarded as an attenuation *cross-section*. In a pure liquid the Beer-Lambert-Bouguer law is often written in the form:

$$\log (I_x/I_o) = -a\, x$$

where $a = \varepsilon\, c$ is the *molar (decadic) attenuation (or absorption) coefficient*. The attenuation will be due to molecular or atomic processes taking place in the pure medium.

The dimensionless product $A = \varepsilon\, c\, x$ is called the *absorbance* (sometimes the *optical density*) and the ratio I_x / I_o is the *transmittance* or *transmissivity*, T. Thus, we can write:

$$\log\, T = -A$$

The Beer-Lambert-Bouguer law finds use in the measurement of concentrations. For example, the clarity or otherwise of polluted air is often measured by comparing the irradiance of light at a certain time with the irradiance on a fine day.

The relationship between these quantities can be expressed thus:

$$\begin{aligned}\text{Incident irradiance } (I_o) = {}&\text{Amount reflected } (I_r) + \text{Amount scattered } (I_s) + \text{Amount absorbed } (I_a)\\ &+ \text{Amount transmitted } (I_t)\end{aligned}$$

$$I_o = I_r + I_s + I_a + I_t$$

or

$$1 = R + S + A + T$$

where R is the *fraction* of light reflected, S the *fraction* of light scattered, A the *fraction* of light absorbed, and T the *fraction* of light transmitted and the quantities measured are the appropriate

irradiance values. In good-quality optical materials the amount of light scattered and absorbed is small and it is often adequate to write:

$$I_o = I_r + I_t$$
$$1 = R + T$$

The units used for attenuation vary, depending on the application. These variations are set out in Appendix A.

1.13.7 Structural colour, iridescence, and electron excitation colour

Broadly speaking, when a nonluminous object is described as coloured, it means that, when illuminated with white light, the light leaving the object and entering the eye is lacking in some wavelengths (or energy bands) that were present initially. In general, there are two main ways that interaction can occur. In one, the light interacts with the constituents of a material but does not exchange energy with the electrons encountered. Colour generated by this relatively passive interaction is termed structural colour. Alternatively, the incident white light can excite electrons in the material from the normally occupied energetic ground state to an excited state, so that light leaving the object is depleted in one or more frequencies and is thus coloured. We will term this electron excitation colour to differentiate it from structural colour.

Structural colours are generated by refraction and dispersion, interference caused by reflection, scattering, and diffraction. Dispersion (Chapter 2) is responsible for the rainbow, chromatic aberration of lenses, and the 'fire' of diamonds. In these and related colour-producing interactions, each wavelength present in the incident white light is slowed down as it traverses a transparent solid. The slowing is wavelength dependent, so that the initial white beam splits into a fan of colours. Reflection from a transparent thin film or a stack of such films similarly gives rise to many well-known colours (Chapter 3). These are produced by interference that effectively removes some wavelengths from the incident radiation. The best known of these colours is that seen in soap films, but many of the shining metallic colours seen in nature, typically displayed in feathers, butterflies, beetles, dragonflies, and a host of other animals, fall into this category. These shining metallic colours often change hue as the viewing angle is changed, and are termed *iridescent* (Figure 1.24). The polarisation of light (Chapter 4) may also be changed by reflection and the colours of a number of insects are influenced by polarisation changes, although these are not detected by human eyes. Scattering of light (Chapter 5), which depends on wavelength, also gives rise to structural colours, most notably the blue of the sky or the blue cast of far-off mountains. Finally, diffraction (Chapter 6), in which the incident white light interacts with planar structures (gratings) or three-dimensional stackings (photonic crystals), which have a similar spatial frequency to the components of the incident beam, gives rise to many colours observed in nature. The iridescent colour of precious opal, as well as the colours of some beetles, butterflies, and other animals arise in this way, as do the colours of many holograms.

Colour that arises by the excitation of electrons is best treated by considering the light beam to consist of photons covering a range of energies matching that of the visible spectrum. Photons falling onto a surface containing atoms or molecules with appropriate energy-level spacings will interact with the electrons in the ground state, and those with the precise energy required for electron promotion are absorbed. The light that leaves is depleted in these energies (colours), and the object is seen as coloured.

Transition metal cations (Chapter 7) generally have suitable energy level spacings and are generally regarded as being coloured. For instance, Cu^{2+} cations endow compounds, including rocks and minerals, with a blue colour, whilst rubies and emeralds owe their colour to Cr^{3+} cations. Laser light from ruby, titanium-sapphire, and Nd lasers, invariably of a single colour, arises when the excited electrons fall back to the ground state. Molecular materials are coloured in an analogous way to transition

Figure 1.24 ***Damselfly exhibiting metallic blue structural colours surrounded by green colour due to electronic excitation of chlorophyll molecules in leaves***

metals, but here the orbitals involved spread over the whole or large parts of the molecule rather than being restricted to a single cation (Chapter 8). Although a number of important inorganic pigments such as ultramarine are coloured in this way, the coloured materials that spring readily to mind are plant colours. The green of leaves arises from absorption of red and blue wavelengths present in white light by chlorophyll molecules (Figure 1.24). Many flowers and fruit are coloured in a similar way, with flavones producing yellow colours, carotenes, orange and anthocyanins, reds, and blues. The vast majority of inks and dyes are based on molecular colouring constituents. Although the energy level schemes are rather different, semiconductors are coloured in a similar fashion (Chapter 10).

Further Reading

The following seven books contain a vast amount of material of relevance to the whole of this book.

Hecht, E. (2002). *Optics*, 4e. San Francisco: Addison-Wesley.

Nassau, K. (2001). *The Physics and Chemistry of Color*, 2e. New York: Wiley-Interscience, Ch. 1 and 2 and Appendix A.

Bohren, C.F. (2001). *Clouds in a Glass of Beer*. New York: Dover originally published by Wiley New York, (1987).

Bohren, C.F. (2006). *What Light Through Yonder Window Breaks?* NY: Dover originally published by Wiley, New York, (1991).

Saleh, B.E.E. and Teich, M.C. (2007). *Fundamentals of Photonics*, 2e. Hoboken, NJ: Wiley.

Lynch, D.K. and Livingston, W. (1995). *Color in Light and Nature*. Cambridge: Cambridge University Press.

Bohren, C.F. and Clothiaux, E.E. (2006). *Fundamentals of Atmospheric Radiation*. Weinheim: Wiley-VCH.

Watson, B. (2016). *Light: A Radiant History from Creation to the Quantum Age*. London: Bloomsbury.

Colour, from the point of view of artist's pigments, is the subject of:

Findlay, V. (2009). *Colour; Travels through the Paintbox*. London: Folio.

The original description of the electromagnetic wave theory is:
Maxwell, J.C. (1865). *Philos. Trans. R. Soc. London, Ser. B* 155: 499.

The interaction with light and nanostructures, (subwavelength optics), giving rise to non-classical effects, such as the extraordinary Young's interference are described in the reviews:
Luo, X. (2018). *Adv. Opt. Mater.* 6: 1701201.
Luo, X. (2019). *Adv. Mater.* 31: 1804680.

Light described in terms of quantum electrodynamics is explained lucidly and non-mathematically in:
Feynman, R.P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton, NJ: Princeton University Press.

The fascinating history of the theories of light is given by:
Cantor, G.N. (1983). *Optics after Newton*. Manchester, UK: Manchester University Press.

Fast and slow light are described by:
Sambles, J.R. (2015). *Science* 347: 828.
Giovanni, D. *et al*. (2015). *Science* 347: 857–860.
Wang, L.J. *et al*. (2000). *Nature* 406: 277.
Mognai, D. *et al*. (2000). *Phys. Rev. Lett.* 84: 4830.
Liu, J. *et al*. (2017). *Adv. Mater.* 29: 1605349.

A historical perspective of luminescence is:
Harvey, E.N. (1957). *A History of Luminescence from The Earliest Times Until 1900*. Philadelphia, PA: American Philosophical Society.

For more information on supercontinuum light see the reviews:
Dudley, J.M., Genty, G., and Coen, S. (2006). *Rev. Mod. Phys.* 78: 1135–1184.
Dudley, J.M. and Taylor, J.R. (2009). *Nat. Photonics* 3: 85–90.

The (coincidental) relationship between the sensitivity of the human eye and the solar spectrum is discussed by:
Stoffer, B.H. and Lynch, D.K. (1999). *Am. J. Phys.* 67: 946–958.

For discussions of the molecular basis of vision see:
Hunt, D.M., Carvalho, L.S., Cowing, J.A., and Davies, W.L. (2009). *Philos. Trans. R. Soc. London, Ser. B* 364: 2941–2945.

The third type of photoreceptor in the eye of humans, called intrinsically photosensitive retinal ganglion cells – ipRGCs, is described by:
Lok, C. (2011). *Nature* 469: 284–285.

For a discussion of bacteriorhodopsin see:
Whitford, J. (2005). *Proteins, Structure and Function*, 114–119. Chichester: Wiley.

The crystal structure of the important intermediate metarhodopsin II is given by:
Choe, H.-W. *et al*. (2011). *Nature* 471: 651–655.

The evolution of primate colour vision is detailed by:
Jacobs, G.H. and Nathans, J. (2009). *Sci. Am.* 300 (4): 40–47.

The complexity of the retina and much information about vision in specialist circumstances is to be found by consulting:

Temple, S., Hart, N.S., Marshall, N.J., and Collin, S. (2010). *Proc. R. Soc. London, Ser. B* 277: 2607–2615.

Many aspects of vision and the interpretation of visual images, including optical illusions, are detailed in the series of articles:

Russ, J.C. (2004). *Seeing the Scientific Image*, *Proc. Roy. Microscopical Soc.*, 39: Part 1, 97–114; Part 2, p179–193; Part 3, p267–281.

The Kubelka-Munk function was originally devised to account for the reflectance of paints, see (1931). *Zeit. f. Tekn. Physik* 12: 593.

There are a number of demonstrations of relevance to this chapter, including diffuse versus specular reflection available at: http://demonstrations.wolfram.com/index.html

Problems and Exercises

Quick Quiz

1. Light from the sun is:
- **(a)** polarised and coherent
- **(b)** unpolarised and incoherent
- **(c)** polarised and incoherent

2. The long wavelength end of the visible spectrum is bounded by:
- **(a)** the infrared region
- **(b)** the ultraviolet region
- **(c)** the microwave region

3. The energy of green light is:
- **(a)** greater than that of red light
- **(b)** less than that of red light
- **(c)** the same as that of red light

4. An incandescent object is classified in terms of its:
- **(a)** colour spectrum
- **(b)** colour temperature
- **(c)** colour correlation

5. Light from a blue sky:
- **(a)** is made up of only blue light
- **(b)** lacks red light
- **(c)** has light of all colours present

6. In the HSB colour model, the initial H is equivalent to:
- **(a)** the wavelength of the light
- **(b)** the amount of whiteness in the light
- **(c)** the irradiance of the light

7. Colour printing uses:
- **(a)** additive colouration
- **(b)** RGB colouration
- **(c)** subtractive colouration

8. The number of colours that can be produced by mixing three primary colours is called:
- **(a)** the range
- **(b)** the gamut
- **(c)** the scope

9. Mixing the three subtractive primary colours gives:
- **(a)** white
- **(b)** black
- **(c)** neutral

10. A transparent glass containing small crystallites will have an attenuation coefficient:
- **(a)** the same as the pure glass
- **(b)** less than the pure glass
- **(c)** greater than the pure glass

Calculations and Questions

1.1 Calculate the energy (J and eV) and wavelength of light photons with frequencies of (a) 7×10^{14} Hz, (b) 6×10^{14} Hz, (c) 5×10^{14} Hz, (d) 4×10^{14} Hz. Describe the colour associated with each of each of these frequencies.

1.2 The colour of ruby crystals is due to absorption and emission of light of wavelengths 556 and 407 nm together with emission of light at 694.3 nm. What are the energies of the corresponding energy levels associated with these transitions above the ground state?

1.3 (a) Taking the sun to be a black body with a temperature of 6000 K and a peak wavelength of 483 nm, estimate the value of the constant in the Wien displacement law. (b) Antares is a bright-red-coloured star in the constellation Scorpius, so named because of its colour similarity to the planet Mars (Ares in Greek). If the temperature of Antares is estimated at 3500 K, explain the red colour of this star. (c) A carbon arc gives a brilliant white light that was used, historically, in projectors. If the peak wavelength is at 580 nm, in the yellow region of the spectrum, estimate the arc temperature.

1.4 The planet Mars has a striking resemblance to the star Antares, and the two can be easily confused (see Q1.3). The surface temperature of Mars varies between 21 °C and −6 °C at the equator and is nearer to −70 °C at the poles. Explain the colour of this planet.

1.5 Spots of oil paint, red, green, and blue, are arranged in a nonoverlapping array. What would the colour be when viewed from a few metres away? The same three colours are mixed together on the artist's palette. What would the colour be?

1.6 What is the energy level separation, ΔE, that gives rise to following electromagnetic radiation types: (a) X-rays, $\lambda = 10^{-10}$ m; (b) ultraviolet, $\lambda = 5 \times 10^{-7}$ m; (c) infrared, $\lambda = 5 \times 10^{-4}$ m; (d) microwaves, $\lambda = 5 \times 10^{-1}$ m; (e) radio waves, $\lambda = 100$ m?

1.7 (a) The pair of complementary colours orange ($\lambda = 600$ nm) and cyan ($\lambda = 490$ nm) have x and y values on the CIE chromaticity diagram (Figure 1.17) of (orange: $x = 0.672$, $y = 0.373$), (cyan: $x = 0.454$, $y = 0.295$). What amounts of light of these two colours needs to be mixed to obtain white light? (b) The pair of complementary colours violet ($\lambda = 400$ nm) and yellow ($\lambda = 570$ nm) have x and y values on the CIE chromaticity diagram (Figure 1.17) of (violet: $x = 0.173$, $y = 0.005$), (yellow: $x = 0.444$, $y = 0.555$). What amounts of light of these two colours needs to be mixed to obtain white light?

1.8 The irradiance of a light beam traversing a solution placed into a cell of 10 cm path length drops to 80.3% of the incident intensity. Calculate the linear absorption coefficient of the solution.

1.9 A dyestuff is quoted as having an attenuation coefficient of 3×10^4 dm^3 mol^{-1} cm^{-1}. What path length in a 10^{-3} M solution would cause the light irradiance to fall to 20% of the incident irradiance?

1.10 A cell of length 1 cm contains a solution of a double-stranded segment of a DNA molecule. The attenuation coefficient of the segment is estimated to be 1.54×10^{-6} l mol^{-1} m^{-1}. A beam of ultraviolet radiation of $\lambda = 260$ nm is diminished by 3.5% on traversing the cell. Determine the concentration of the molecules present.

1.11 The mass absorption coefficient for X-rays from a molybdenum target, with a wavelength of 0.0709 nm are: Fe, 3.81 m^2 kg^{-1}; O, 0.145 m^2 kg^{-1}. (a) Determine the mass absorption coefficient of iron oxide Fe_2O_3. (b) Estimate the thickness of oxide that will reduce the irradiance of an X-ray beam to 1/1000 of the incident value. Molar masses, Fe, 55.85 g mol^{-1}; O, 16 g mol^{-1}; densities: Fe, 7873 kg m^{-3}; Fe_2O_3, 5240 kg m^{-3}.

1.12 Calculate the spectral irradiance, I_λ at wavelengths of 350, 500, and 650 nm for a black body at 6000 K.

1.13 Construct the wave profiles shown in Figure 1.4a,b and Figure 1.5a,b.

1.14 Derive the equation $\Delta\nu / \Delta\lambda \approx \nu_m / \lambda_m$ (Section 1.8.5).

1.15 Determine the mean frequency, bandwidth, coherence time, and coherence length for the following spectral lines.

- **(a)** from a high-pressure mercury lamp, $\lambda_m = 546.1$ nm, $\Delta\lambda = 1$ nm
- **(b)** from a low-pressure Kr lamp, $\lambda_m = 605.6$ nm, $\Delta\lambda = 1.2 \times 10^{-3}$ nm
- **(c)** from a low-pressure Cd lamp, $\lambda_m = 643.85$ nm, $\Delta\lambda = 6.5 \times 10^{-4}$ nm
- **(d)** from a stabilised He–Ne laser, $\lambda_m = 632.8$ nm, $\Delta\lambda = 1 \times 10^{-6}$ nm