A2 Optics lecture notes

Lance Miller

March 6, 2022

Contents

1 Rays, waves and Fermat **3**

2 Interference **3**
2.1 Waves and coherence . . . 3
2.2 Interference between two coherent point sources . . . 5
2.3 Optical path difference and phase difference . . . 6
2.4 Interference between multiple coherent point sources . . . 7
2.5 The Fraunhofer condition . . . 8

3 Diffraction **8**
3.1 Fraunhofer diffraction from a rectangular aperture . . . 8
3.2 The Fourier Transform . . . 9
3.3 Using the convolution theorem . . . 12
3.3.1 Example 1 . . . 13
3.3.2 Example 2 . . . 13
3.4 The interpretation of diffraction theory . . . 13
3.5 Huygens-Fresnel diffraction . . . 14
3.6 Fresnel-Kirchoff diffraction . . . 15
3.7 From the Fresnel-Kirchoff integral to the Fourier transform . . . 16
3.8 Fresnel or Fraunhofer? . . . 17
3.9 Replacing the distant observer, the importance of Fraunhofer diffraction and the point-spread function . . . 17

4 Diffraction in optical instruments: telescope, microscope and diffraction grating spectrograph **18**
4.1 A simple imaging lens . . . 18
4.2 The astronomical telescope . . . 19
4.3 The angular resolution and resolving power of a telescope . . . 19
4.4 The compound microscope . . . 20
4.5 The linear resolution and resolving power of a microscope with incoherent illumination . . . 20
4.6 Abbé's theory of the resolving power of a microscope for coherent illumination . . . 22
4.7 Aberrations in imaging systems . . . 22
4.8 The diffraction grating as a spectrograph . . . 24
4.9 A practical spectrograph design . . . 27

5 Interference fringes and interferometers **29**
5.1 Interference by division of wavefront or amplitude . . . 29
5.2 Localisation of fringes . . . 29
5.3 Interference fringes from reflections off two surfaces . . . 29
5.3.1 Fringes of equal inclination. . . . 29
5.3.2 Fringes of equal thickness. . . . 31
5.4 The Michelson interferometer . . . 32
5.4.1 The Michelson interference pattern . . . 32
5.4.2 The Michelson interferometer for length measurement . . . 33
5.4.3 The Michelson as a spectrometer . . . 33
5.4.4 The Michelson interferometer as a Fourier Transform Spectrometer . . . 34
5.4.5 Autocorrelation . . . 35

5.4.6 Aether drift experiments . . . 35
5.5 The Fabry-Perot etalon and spectrometer . . . 36
5.5.1 The Fabry-Perot intensity function . . . 36
5.5.2 The fringe pattern . . . 37
5.5.3 Finesse . . . 39
5.5.4 The Fabry-Perot interferometer as a spectrometer: resolution, free spectral range and resolving power . . . 39
5.5.5 The effective number of reflections . . . 39
5.5.6 Use of the Fabry-Perot . . . 40
5.6 Interferometric detection of gravitational waves . . . 40

6 Basics of Lasers **42**
6.1 Interaction of atoms with radiation and the Einstein coefficients . . . 42
6.2 Conditions for laser gain and population inversion . . . 43
6.3 Pumping . . . 44
6.4 Laser design . . . 44
6.4.1 Laser oscillators . . . 44
6.4.2 Optical pumping . . . 44
6.4.3 Collisional pumping . . . 44

7 Coherence **45**
7.1 Temporal coherence . . . 45
7.2 The temporal coherence of lasers . . . 46
7.3 Interference fringe visibility and coherence . . . 47
7.4 Transverse coherence . . . 48
7.5 The astronomical imaging interferometer . . . 50

8 Polarised light **52**
8.1 Introduction to polarised light . . . 52
8.1.1 Linear, circular and elliptical polarised light . . . 52
8.1.2 The Jones vector . . . 54
8.1.3 Unpolarised and natural light . . . 54
8.2 Dichroism . . . 55
8.3 Birefringence . . . 55
8.4 Waveplates and analysis of polarised light . . . 56
8.4.1 Birefringent waveplates . . . 56
8.4.2 Jones matrices . . . 57
8.4.3 The full-wave plate . . . 57
8.4.4 The half-wave plate . . . 57
8.4.5 The quarter-wave plate . . . 58
8.4.6 The circular polariser . . . 58
8.4.7 Achromatic waveplates . . . 58
8.4.8 Analysis of polarised light . . . 58
8.4.9 Study of crystals in polarised light . . . 59
8.5 Induced birefringence . . . 59

Appendices **60**

A The reflection of polarised light **60**
A.1 Terminology . . . 60
A.2 **s**-mode reflection and refraction . . . 60
A.3 **p**-mode reflection and refraction . . . 61
A.4 The Brewster angle . . . 61
A.5 Total internal reflection . . . 62
A.6 The phase shift of reflected light . . . 63

1 Rays, waves and Fermat

According to Fermat, light rays follow a path s between two points A and B, such that the time taken is a "stationary value" - usually a local minimum, with respect to variations in the path.

$$\delta \int_s \mathrm{d}t = 0\,, \tag{1}$$

where the integral is along the path s and where δ indicates a variation in the integral with respect to the quickest path. If the speed of light in a medium is v, we may define refractive index $n = c/v$, then $\mathrm{d}t = (n/c)\mathrm{d}s$ and we can define the *optical path length* d_s

$$d_s = \int_s n\,\mathrm{d}s \tag{2}$$

which should also be minimised or be a stationary value.

We can use Fermat's principle to derive all of the equations of geometrical ray optics (e.g. the laws of reflection and refraction, the lens makers equation) in the first-year course.

Exercise: derive Snell's law of refraction using Fermat's principle.

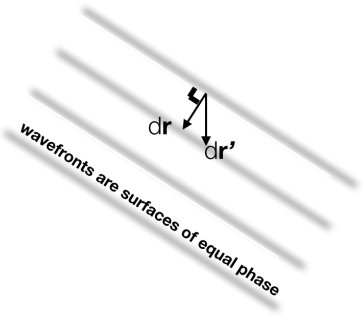


Figure 1: A ray path element d**r**, locally perpendicular to the wavefront, must be the shortest path between wavefront maxima, compared with any other path d**r'**.

We now know that light is a transverse electromagnetic wave. The concept of a ray is still extremely useful: **a ray is a path the defines the direction of propagation of a light wave**. In an isotropic medium, the ray is always locally perpendicular to the *wavefront*, where **the wavefront is the locus of points on a transverse wave that instantaneously all have the same phase (e.g. the set of points that join up the maxima of the wave at some instant of time)**. Fig. 1 illustrates why such a ray must satisfy Fermat's principle (for a more mathematical proof, see Born and Wolf, section 3.3.2).

If a group of rays are describing the same set of wavefronts, they have *normal congruence* - i.e. there is a wavefront surface which is orthogonal to all of the rays. *Malus and Dupin* (in the period 1808-16) proposed that any congruent set of rays will remain congruent even after any series of reflections or refractions (as in Fig. 2). An equivalent route to obtaining the laws of geometrical optics is to note that the wavefront takes the same time along the path A, A', A'' as it does along B, B', B'' (see Hecht, Chapter 4).

2 Interference

2.1 Waves and coherence

In the following we assume that light propagates from a "point source" (an idealised concept) as a spherical wavefront with (complex) instantaneous amplitude

$$u = \frac{u_0}{r}\exp[\mathrm{i}(\omega t - kr)] \tag{3}$$

as a function of time t and at distance r from the source, where u_0 is the normalisation of the wave amplitude, ω is the wave angular frequency and the wavenumber is $k = 2\pi/\lambda$, for light of wavelength λ.

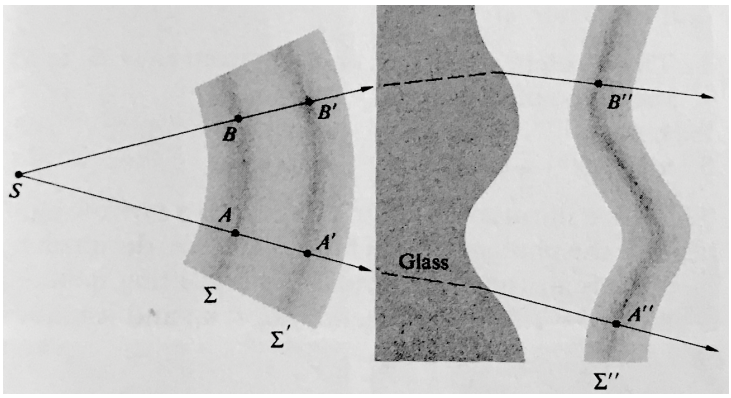


Figure 2: Locations A, B are points on a wavefront at Σ with normal congruence, originating from a source S. The wavefront maintains its congruence at later times, Σ' and even after passing through some complicated refracting medium, Σ''. Figure from Hecht, Chapter 4.

If the source is very distant, then local to the observer with some vector position coordinate $\mathbf{x}$, the wave may be approximated as a plane wave propagating as:

$$u = u_0 \exp[\mathrm{i}(\omega t - \mathbf{k}.\mathbf{x})]\,, \tag{4}$$

where now we write the wavenumber as a vector $\mathbf{k}$ whose direction indicates the direction of propagation of the plane wave with respect to the coordinate system used to define $\mathbf{x}$. An optical system may be designed so that plane waves are output by the optics, if required (e.g. the *collimator* of a spectrograph).

For much of what we write below, we consider position vectors $\mathbf{x}$ in the direction of wave travel that are parallel to the wavevector $\mathbf{k}$, in which case we can just write the dot product $\mathbf{k}.\mathbf{x} = kx$, thus

$$u = u_0 \exp[\mathrm{i}(\omega t - kx)]\,. \tag{5}$$

We may also choose a different convention and write the complex exponent the other way round,

$$u = u_0 \exp[\mathrm{i}(kx - \omega t)]\,. \tag{6}$$

The choice of sign convention does not affect the observed intensity (equation 8). We may introduce a phase shift ϕ by allowing u_0 to be a complex constant:

$$u_0 = |u_0| \exp[\mathrm{i}\phi]\,. \tag{7}$$

Thus, u is the instantaneous complex amplitude of the wave, with peak amplitude u_0. Writing it as a complex wave in this way allows more compact maths and allows us to choose the wave to have any starting phase that we wish - usually, for any particular realisation of the wave, we would choose just one component, e.g. the real part. In an electromagnetic wave, the amplitude is a vector, e.g $\mathbf{E}$, but for much of these lectures we treat the wave amplitude as being a scalar, effectively considering light to have only a single polarisation. We will revisit this towards the end of the course (notes Section 8).

The *irradiance* or *intensity* of light that we measure is

$$I \propto \langle u u^\star \rangle\,, \tag{8}$$

where the angle brackets indicate averaging over some interval of time much longer than the time period of the wave, and where the average is normally taken of the real part of the time dependence. I is a real quantity that is a measure of the power flowing per unit area of the wave[1].

[1] Usually we want I to have units of energy per unit time per unit area, and the units of wave amplitude u depend on what type of wave we are dealing with. We now know that light is an electromagnetic wave. If the wave amplitude is the amplitude of the electric field vector $\mathbf{E}$, then the instantaneous energy crossing unit area per unit time is given by the Poynting vector,

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}\,. \tag{9}$$

For a plane wave in a vacuum, consider the real part of a complex wave, say $E = E_0 \cos(\omega t - kx)$, where E_0 is the electric field peak amplitude, and in our notation $E_0 \equiv u_0$. The magnitudes of these vectors are related by $B = \mu_0 H = \frac{E}{c}$, also $c = (\epsilon_0 \mu_0)^{-1/2}$ and the time-averaged intensity then is

$$\begin{aligned} I &= \langle S \rangle \\ &= \epsilon_0 c \langle |E|^2 \rangle \\ &= \frac{1}{2} \epsilon_0 c |E_0|^2\,, \end{aligned} \tag{10}$$

which is the total energy density in the E and H fields times the wave speed, the speed of light.

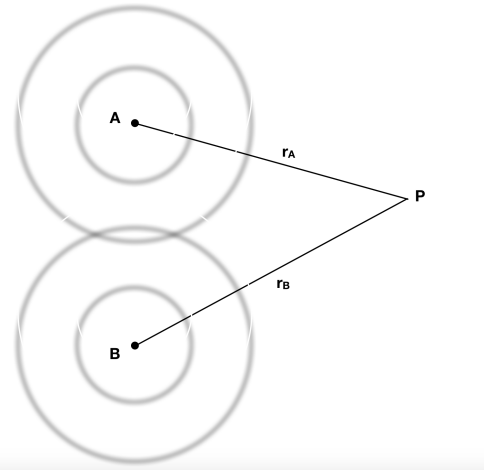


Figure 3: Two idealised point sources at A and B emit light waves isotropically. We are interested in the amplitude and intensity measured at point P.

Now let us consider the wave amplitude and intensity that we measure if we are at a point P which is distance r_A and r_B from two sources of light, denoted A and B, as in Fig. 3. At any instant of time,

$$u_P = u_A + u_B \,, \tag{11}$$

where u_A, u_B are the instantaneous amplitudes at P due to the light from A and B respectively,

$$\begin{aligned} u_A &= u_{A,0} \exp[\mathrm{i}(\omega_A t - k_A r_A)] \,, \\ u_B &= u_{B,0} \exp[\mathrm{i}(\omega_B t - k_B r_B)] \,. \end{aligned} \tag{12}$$

The resulting intensity, averaged over time, is

$$\begin{aligned} I &\propto \langle (u_A + u_B)(u_A^\star + u_B^\star) \rangle \\ &\propto \langle u_A u_A^\star \rangle + \langle u_B u_B^\star \rangle + \langle u_A u_B^\star + u_B u_A^\star \rangle \end{aligned} \tag{13}$$

If $k_A \neq k_B$ and/or $\omega_A \neq \omega_B$, the last term averages to zero, when averaged over a large number of oscillations of the wave, and so we have

$$I = I_A + I_B \,. \tag{14}$$

The sources in this case are said to be *incoherent*. The total intensity is the sum of the intensities of the sources when viewed independently.

However, if $k_A = k_B$ and $\omega_A = \omega_B$ the sources are said to be *coherent* and the last term in equation 13 causes *interference* effects to be seen: the total intensity is no longer the sum of the individual intensities. We expect the total intensity to depend on the path difference between the sources: energy is conserved but is redistributed into bright, spatially-dependent *fringes*, which we will look at in detail later.

This idealised view of coherence implies we are dealing with purely monochromatic light, which cannot be produced by any real physical light source, so in Section 7, we will look in more detail at coherence in more realistic systems.

2.2 Interference between two coherent point sources

From the above we may write

$$u_P = \frac{u_{0,A}}{r_A} \exp[\mathrm{i}(\omega t - k r_A)] + \frac{u_{0,B}}{r_B} \exp[\mathrm{i}(\omega t - k r_B)] \,. \tag{15}$$

Let us consider the case where the point P is very distant from A and B, such that $r_A, r_B \gg x$, where x is the distance between A and B. Then the rays to P may be considered parallel, as in Fig. 4. Then, $r_B \simeq r + (x \sin\theta)/2$ and $r_A \simeq r - (x \sin\theta)/2$, and if the sources have equal emitted amplitude, $u_{0,A} = u_{0,B}$,

$$\begin{aligned} u_P &\simeq \frac{u_0}{r} \exp[\mathrm{i}(\omega t - kr)] \left(\exp[\mathrm{i}kx \sin\theta/2] + \exp[-\mathrm{i}kx \sin\theta/2]\right) \\ &\simeq 2\frac{u_0}{r} \exp[\mathrm{i}(\omega t - kr)] \cos\left[\frac{\delta}{2}\right] \,, \end{aligned} \tag{16}$$

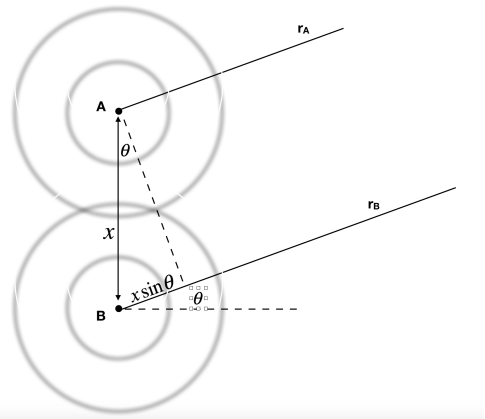


Figure 4: Two idealised point sources at A and B emit light waves isotropically, viewed at a very distant location.

where $\delta \equiv kx\sin\theta$ is the phase difference between the waves from the two sources. Thus the intensity is

$$\begin{aligned} I &\propto \frac{1}{2}uu^\star \\ &\simeq \frac{2|u_0|^2}{r^2}\cos^2\left[\frac{\delta}{2}\right] . \end{aligned} \tag{17}$$

There is a factor $1/2$ in the intensity normalisation from taking the real part and time averaging. Note that maxima occur when $\delta = 2n\pi$ or $x\sin\theta = n\lambda$, for integer n, i.e. when the path difference between the sources is an integer number of wavelengths.

For future use, note also the convenient relation

$$\cos^2\frac{\delta}{2} = \frac{1}{2}\left(1+\cos\delta\right) . \tag{18}$$

If, instead, the sources were incoherent, we would simply add their (time-averaged) intensities, and we would obtain

$$I \propto \frac{|u_0|^2}{r^2} \tag{19}$$

The peak intensity in the coherent sources case is twice the intensity of the incoherent case, but the average intensity over the intensity pattern is the same, so energy is conserved, but is concentrated in the peaks of the interference pattern.

2.3 Optical path difference and phase difference

The quantity δ above is an example of the phase difference between two coherent waves. It is the variation in δ that determines how the interference pattern looks and this is generally the most important quantity to consider.

Recall that we also defined the optical path length along a ray as $d_s = \int_s n\,\mathrm{d}s$ (equation 2), for refractive index n, so the phase difference δ is related to the *difference* in optical path lengths, Δd_s between two rays, $\delta \equiv k_0 \Delta d_s$, where k_0 is the wavenumber in vacuum. The optical path length definition already includes the refractive index n and hence the dependence of the wavenumber k on n: $k = k_0 n$. In a medium with high refractive index, the wave travels more slowly than in vacuum and so wavelength is shortened with respect to the vacuum wavelength, and the wavenumber $k \equiv 2\pi/\lambda$, is larger than in vacuum.

The first question of your problem set is a guided exercise in working out the optical path difference between rays reflected from two plane parallel surfaces: we will need this calculation later in the course.

2.4 Interference between multiple coherent point sources

The above treatment may be easily extended to multiple point sources. Consider a row of N point sources, each separated from its neighbour by distance x. The far-field amplitude is

$$\begin{aligned} u &\propto \sum_{m=0}^{N-1} \exp[im\delta] \\ &= 1 + \mathrm{e}^{\mathrm{i}\delta} + (\mathrm{e}^{\mathrm{i}\delta})^2 + \ldots (\mathrm{e}^{\mathrm{i}\delta})^{N-1}\,, \end{aligned} \tag{20}$$

for convenience dropping the dependence on distance and where we define $\delta \equiv kx\sin\theta$. This is a standard geometric series: consider the quantity:

$$\begin{aligned} u(\mathrm{e}^{\mathrm{i}\delta} - 1) &= u + (\mathrm{e}^{\mathrm{i}\delta})^N - 1 - u \\ &= \mathrm{e}^{\mathrm{i}N\delta} - 1\,, \end{aligned} \tag{21}$$

so

$$\begin{aligned} u &= \frac{\mathrm{e}^{\mathrm{i}N\delta} - 1}{\mathrm{e}^{\mathrm{i}\delta} - 1} \\ &= \frac{\mathrm{e}^{\mathrm{i}N\delta/2}[\mathrm{e}^{\mathrm{i}N\delta/2} - \mathrm{e}^{-\mathrm{i}N\delta/2}]}{\mathrm{e}^{\mathrm{i}\delta/2}[\mathrm{e}^{\mathrm{i}\delta/2} - \mathrm{e}^{-\mathrm{i}\delta/2}]} \\ &= \mathrm{e}^{\mathrm{i}(N-1)\delta/2}\frac{\sin(N\delta/2)}{\sin(\delta/2)}\,, \end{aligned} \tag{22}$$

and so the intensity is

$$I(\delta) \propto \langle uu^\star \rangle = \frac{I_0}{N^2}\left[\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right]^2\,, \tag{23}$$

where the normalisation term $1/N^2$ is included[2] so that we can define $I = I_0$ for $\delta = 0$. This function is shown in Fig. 5 for $N = 8$ - in reality N is likely to be 1000s. The function has maxima as the denominator

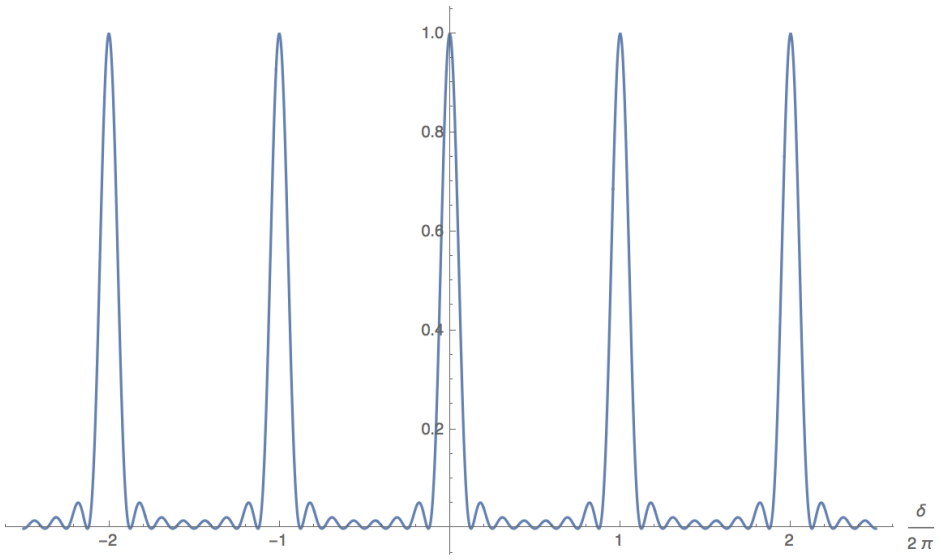


Figure 5: The ideal diffraction grating response of equation 23 for $N = 8$, plotted as a function of $\delta/2\pi = x\sin\theta/\lambda$.

$\sin(\delta/2) \to 0$ and then $I \to I_0$, at $\delta = 2n\pi$ for integer n, or $x\sin\theta = n\lambda$ - this is the *grating equation*, where the value of n is the *order* of the diffraction maximum. There are much smaller secondary maxima when $Nx\sin\theta = (p + \frac{1}{2})\lambda$ for integer p, and zeros (minima) of the intensity occur when $Nx\sin\theta = q\lambda$ for integer $q \neq Nn$ (note that the product Nd is the total width of the grating). We will look at more realistic diffraction gratings in section 4.8.

[2] This allows a convenient definition of I_0, the $1/N^2$ is not a derived factor. As $\delta \to 0$, $\frac{1}{N^2}\left[\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right]^2 \to 1$.

2.5 The Fraunhofer condition

In the above treatments, by placing the observer P towards infinite distance, the rays to the observer became parallel, and the relative phases of the contributions to the amplitude at P became *a linear function of the coordinate x lying in the plane of the aperture*. This is the *Fraunhofer condition* and we term interference or diffraction observations that satisfy this condition as being *Fraunhofer interference* or *Fraunhofer diffraction*.

The importance of the Fraunhofer condition lies in the correspondence with the far-field limit, and this is the condition in which we operate a wide range of optical instruments. For example, the Fraunhofer diffraction of an astronomical telescope is responsible for the image that we see if we observe a point-like source such as a distant star - in that case we observe the telescope's *point spread function*, which is the Airy profile diffraction pattern for an ideal telescope (see below), and which provides the ultimate limit on the angular resolution attainable with that telescope[3]. See sections 3.9 & 4.3.

3 Diffraction

Diffraction may be viewed as the extension of interference phenomena to the case of finite apertures and obstructions. Credit for the first written description of diffraction is given to Grimaldi in 1664, when he observed the diffraction pattern in the shadow of a rod. The phenomenon was the subject of much debate until the work of Young, Fresnel and Kirchoff in the 19th century established it as arising from the wave nature of light. Note that Huygens' wavelet theory did not attempt to explain diffraction - Huygens thought that light could only travel in straight lines.

3.1 Fraunhofer diffraction from a rectangular aperture

Consider the case of a plane wave passing through a 2D plane aperture and the resulting intensity being viewed at point P a large distance away. Let us divide up the aperture into small 2D elements $\mathrm{d}x\mathrm{d}y$, and in the limit as $\mathrm{d}x$,$\mathrm{d}y \to 0$ we shall adopt the Huygens scheme and assume that each part of the wavefront at the aperture may be replaced by a coherent point source which is isotropically radiating. At the observer, we define a coordinate system in a plane perpendicular to the optical axis, as shown in Fig. 6. The distance r between a point on the aperture at coordinate x, y and the observer at P, with

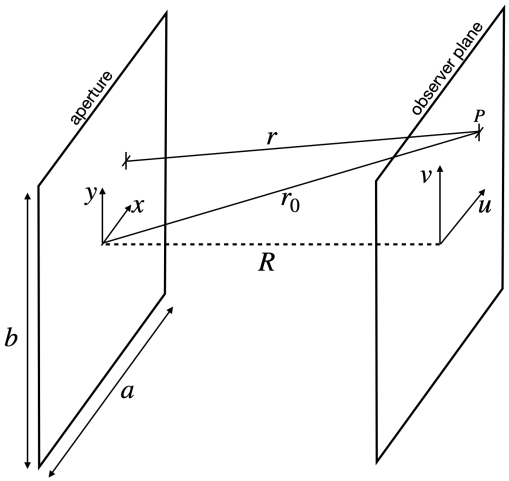


Figure 6: Geometry of diffraction by a finite aperture, of dimension $a \times b$, as seen by an observer at P.

coordinate u, v, is $r = [R^2 + (u-x)^2 + (v-y)^2]^{1/2}$ and the distance r_0 between the centre of the aperture and P is $r_0 = [R^2 + u^2 + v^2]^{1/2}$, where R is the distance between the aperture and the observer plane.

[3] In practice, large telescopes on the ground do not attain this resolution limit, because the turbulent refracting medium that is the Earth's atmosphere causes the ideal plane wavefront from a star to become deformed, making the actual resolution worse than the Fraunhofer diffraction limit. Telescopes in space, such as the *Hubble Space Telescope*, its successor the *James Webb Space Telescope* or the *Euclid* mission come close to the Fraunhofer diffraction limit.

Then,

$$\begin{aligned} r &= [R^2 + u^2 + v^2 + x^2 + y^2 - 2(ux + vy)]^{1/2} \\ &= [r_0^2 + x^2 + y^2 - 2(ux + vy)]^{1/2} \\ &\simeq r_0 \left[1 + \frac{x^2 + y^2}{2r_0^2} - \frac{ux + vy}{r_0^2}\right] \\ &\simeq r_0 - \frac{ux + vy}{r_0} \end{aligned} \tag{24}$$

when $r_0^2 \gg \{ux, vy\} \gg \{x^2, y^2\}$. Let us write $u' \equiv u/r_0$, $v' \equiv v/r_0$, these being the two "direction cosines" of the angles at P between the vector r_0 and the u, v coordinates in the observer plane.

Then the far-field diffraction pattern at point P due to the element $\mathrm{d}x\mathrm{d}y$ is

$$\begin{aligned} \mathrm{d}u_P &\propto \frac{1}{r} \exp[\mathrm{i}kr] \\ &\propto \exp[\mathrm{i}kr_0] \exp[-\mathrm{i}k(u'x + v'y)] \,, \end{aligned} \tag{25}$$

where we have neglected the $1/r$ dependence, since for $r_0^2 \gg \{x^2, y^2\}$, $1/r \simeq 1/r_0$ and hence this term may be assumed invariant with x, y, in this "distant observer" approximation. The expression satisfies the Fraunhofer condition, because the exponent depends only linearly on the coordinates x, y in the aperture plane. Integrating over the aperture, whose dimensions are a, b,

$$\begin{aligned} u_P &\propto \exp[\mathrm{i}kr_0] \int_{y=-b/2}^{y=b/2} \int_{x=-a/2}^{x=a/2} \exp[-\mathrm{i}k(u'x + v'y)] \mathrm{d}x\mathrm{d}y \\ &\propto \exp[\mathrm{i}kr_0] \int_{y=-b/2}^{y=b/2} \exp[-\mathrm{i}kv'y] \mathrm{d}y \int_{x=-a/2}^{x=a/2} \exp[-\mathrm{i}ku'x] \mathrm{d}x \\ &\propto \exp[\mathrm{i}kr_0] \left[\frac{\exp[-\mathrm{i}kv'y]}{-\mathrm{i}kv'}\right]_{y=-b/2}^{y=b/2} \left[\frac{\exp[-\mathrm{i}ku'x]}{-\mathrm{i}ku'}\right]_{x=-a/2}^{x=a/2} \\ &\propto \exp[\mathrm{i}kr_0] \frac{2\sin(kv'b/2)}{kv'} \frac{2\sin(ku'a/2)}{ku'} \\ &\propto \exp[\mathrm{i}kr_0] \, a\, b \, \mathrm{sinc}(kv'b/2) \, \mathrm{sinc}(ku'a/2) \,. \end{aligned} \tag{26}$$

The observed intensity then is

$$I(u', v') \propto |u_P|^2 \propto a^2 \, b^2 \, \mathrm{sinc}^2(kv'b/2) \mathrm{sinc}^2(ku'a/2) \,. \tag{27}$$

Notice that the benign constant phase term $\exp[\mathrm{i}kr_0]$ disappears in the expression for the intensity ($|\mathrm{e}^{\mathrm{i}\theta}| = 1$).

For this square aperture, the diffraction pattern is the product of two functions, one in u', the other in v'. It is often convenient to consider this as a 1D problem, if $b \gg a$: i.e. the aperture becomes a tall, narrow slit. In that case, the diffraction pattern becomes confined to the horizontal, u', plane, because for large b the first sinc expression falls off rapidly to zero with v'. Because of the separation of variables, we could just consider the 1D case to get the u' component alone.

3.2 The Fourier Transform

Review the mathematics of the Fourier transform from your *Mathematical Methods* course. You will also need the convolution theorem from that part of that course.

A wavepacket of shape $T(t)$ as a function of time t has a Fourier transform

$$\widetilde{T}(\omega) = \int_{-\infty}^{\infty} T(t) \exp[-\mathrm{i}\omega t] \mathrm{d}t \,, \tag{28}$$

where the angular frequency ω is related to the wave frequency ν by $\omega = 2\pi\nu$. Thus the Fourier transform tells us what range of frequencies exist in a pulse of light (say). The Fourier transform has an inverse

$$T(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widetilde{T}(\omega) \exp[\mathrm{i}\omega t] \mathrm{d}\omega \,. \tag{29}$$

The factor $1/2\pi$ is sometimes instead placed in front of the forward transform, or sometimes its square-root is placed in front of both integrals: the latter is the convention adopted in the Mathematical Methods

course. In general, for the results we obtain, it does not matter if we consider the forward or the inverse Fourier transform.

The Fourier transform does not have to involve the physical quantities of time and frequency, any mathematical variables that are related by an integral of this form are said to be *Fourier transform pairs*. Some common Fourier transforms are reproduced in Fig. 7 from the textbook by Bracewell. These are well worth familiarising yourself with, because they arise in many branches of physics, but especially in optics. In this figure, the Fourier transform has been defined as

$$\tilde{f}(s) = \int_{-\infty}^{\infty} f(x) \exp[-\mathrm{i}2\pi s x] \mathrm{d}x \,. \tag{30}$$

One of the most important is that the Fourier transform of a tophat is a sinc function. Let us define the *aperture function* for the 2D aperture in section 3.1 as

$$T(x,y) = \begin{cases} 1 & -a/2 < x < a/2 \ \& \ -b/2 < y < b/2 \\ 0 & \text{otherwise,} \end{cases} \tag{31}$$

which is a "top-hat" function in 2D. This describes the aperture of section 3.1 as it has value of unity where light passes through the aperture, and zero everywhere else. Then (neglecting the benign phase term outside the integral), the diffraction pattern in section 3.1 is

$$\begin{aligned} u_P(u',v') &\propto \int_{y=-b/2}^{y=b/2} \exp[-\mathrm{i}kv'y] \mathrm{d}y \int_{x=-a/2}^{x=a/2} \exp[-\mathrm{i}ku'x] \mathrm{d}x \\ &\propto \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T(k'_x, k'_y) \exp[-\mathrm{i}(k'_y v' + k'_x u')] \mathrm{d}k'_x \mathrm{d}k'_y \end{aligned} \tag{32}$$

which is just a 2D Fourier transform, if we identify the Fourier transform variables as $k'_x \equiv kx$ and $k'_y \equiv ky$, remembering that $k \equiv 2\pi/\lambda$ is the wavenumber of the light: i.e. to calculate diffraction patterns using the Fourier transform, we first define the aperture function $T(x,y)$, scale the aperture coordinates by the wavenumber k, and then take the Fourier transform.

In the above, we have looked at a special case of the rectangular aperture, but we can see that in general **the Fraunhofer diffraction pattern of an aperture is (proportional to) the Fourier transform of the aperture function, when scaled by the wavenumber**, valid providing the Fraunhofer condition applies. The resulting pattern depends on the direction cosines u', v' and hence we should think of the diffraction pattern as being an angular pattern. In general, the result of the Fourier transform will be a complex-valued function of the direction cosines: the diffraction pattern has an "amplitude" and a "phase". Remember to take the modulus squared of the result in order to obtain an expression for the light intensity.

We have not yet considered the constant of proportionality, we have not dealt very well with how we go from the case of a set of discrete oscillators to an element $\mathrm{d}x\mathrm{d}y$ of a continuous aperture: we will come to that later.

Note two important, related, general principles about the Fourier transform:

1. A broad function in one domain transforms to a narrow function in the other, and vice versa - in quantum mechanics, this is an expression of the Heisenberg uncertainty principle. For example, a wide gaussian in x or t transforms to a narrow gaussian in k or ω, and vice versa, and the product $\Delta t \, \Delta x$ of the widths of the functions in each domain, Δt and Δx, is invariant. A sharp pulse in time contains a broad range of frequencies, a long-lived signal contains only a narrow range of frequencies. In the limit, a monochromatic signal (comprising a delta-function in frequency) must be infinite in time (and therefore not possible in practice).

2. Small-scale structure in the "real" domain (t or x in our notation) produces values over a wide range of (temporal, spatial or angular) frequencies in the Fourier domain. Conversely, large-scale structure in the "real" domain contributes only to low (temporal, spatial or angular) frequencies[4].

The 2D rectangular aperture was a simple case to deal with, other cases might not be so easy and might need numerical evaluation of the Fourier transform. Luckily, from the 1960s, some clever algorithms known collectively as the *Fast Fourier Transform* have been designed to compute these very rapidly (see the Wikipedia article on this for a discussion of its history).

[4] A lens acts as a *low-pass filter*, removing the high angular frequencies and eliminating small-scale structure in an image produced by the lens - see section 4.6.

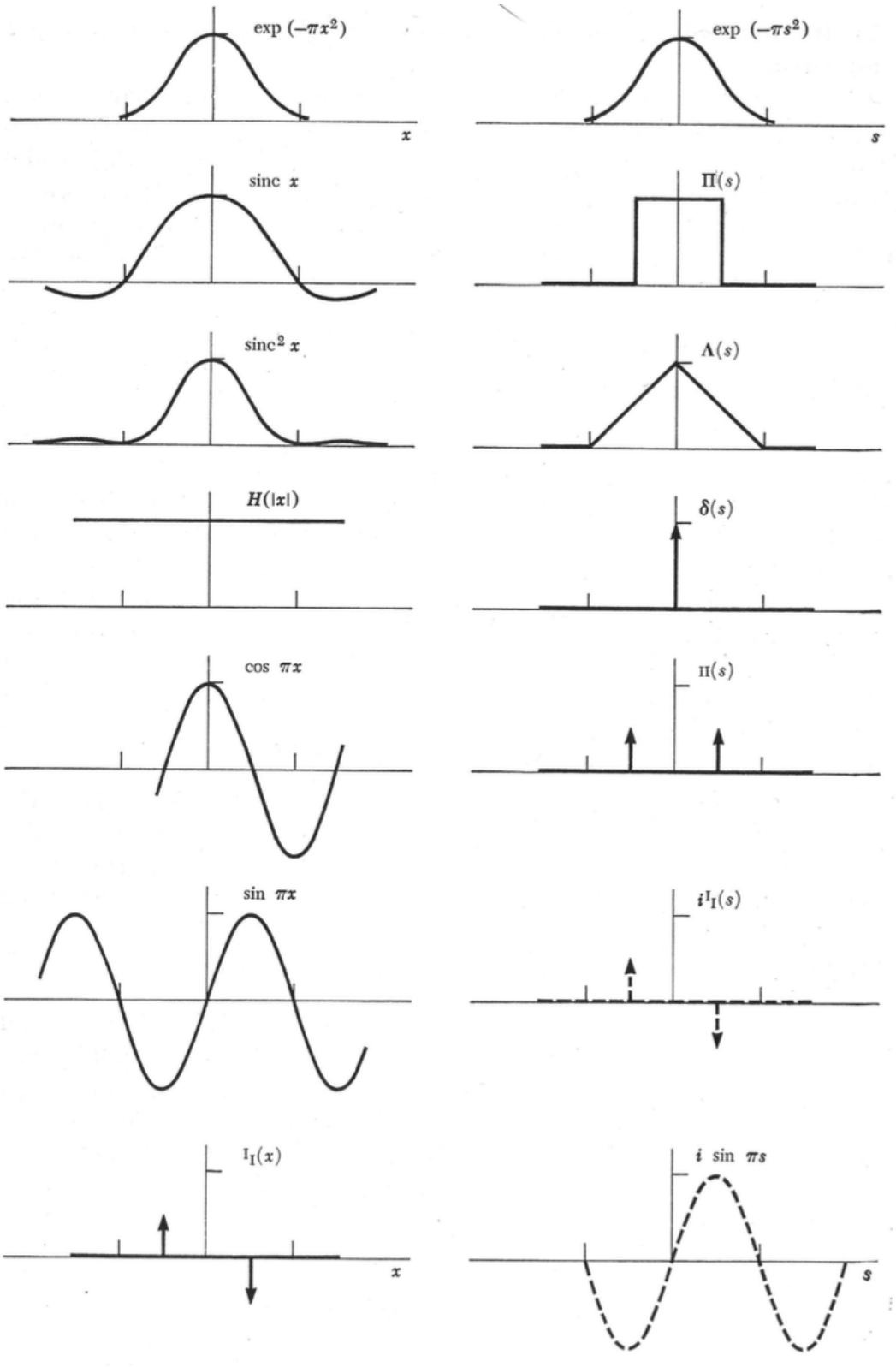


Figure 7: Examples of some useful Fourier transforms, reproduced from Bracewell, "The Fourier Transform and its applications". The functions, from top to bottom, are: (i) the Fourier transform of a gaussian is another gaussian; (ii) a sinc function transforms to a tophat function, and vice versa; (iii) a sinc^2 function transforms to a triangular function (see section 3.3); (iv) a constant extending to $\pm\infty$ transforms to a delta function, and vice versa; (v) a cosine transforms to a pair of delta functions at positive and negative frequency; (vi) a sine transforms to a pair of imaginary delta functions with opposing signs; (vii) a pair of real delta functions with opposing signs transforms to an imaginary sine. We see the second of these in the diffraction from a finite width slit. We see the fifth of these in Young's slits.

One the most important cases is the transform of a circular aperture, such as the entrance pupil of a telescope or microscope. An aperture of radius a produces a diffraction pattern whose intensity is given by the *Airy profile*,

$$I(\theta) = I(0)\left[\frac{2J_1(ka\sin\theta)}{ka\sin\theta}\right]^2 , \tag{33}$$

where J_1 is the Bessel function of the first kind (for a derivation, see e.g. Hecht - we do not need to derive this in our course). This is a circularly-symmetric function with oscillating rings around a central maximum - the radial profile as plotted by *Mathematica* is shown in Fig. 8. If we observe a star with a diffraction limited telescope, this is essentially the pattern we see in our detector, as shown in Fig. 9 for the intensity distribution expected for the Euclid space telescope. Notice the pattern of rings that oscillate with radius, known as "Airy rings".

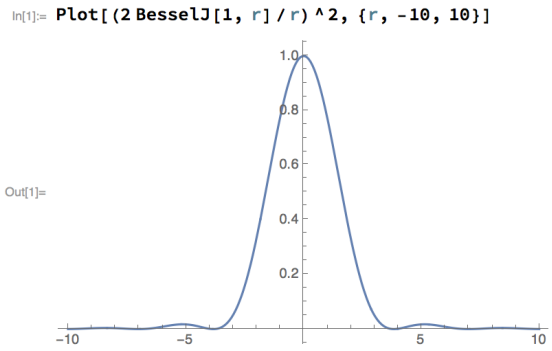


Figure 8: Plot of the radial intensity of the Airy profile as a function of $r = ka\sin\theta$, plotted using *Mathematica*.

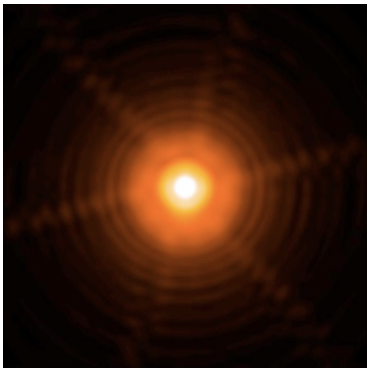


Figure 9: The 2D intensity distribution expected from the observation of a star with the *Euclid* space telescope. The colour scale is plotted logarithmically so that the fainter Airy profile circularly-symmetric ring structure is more visible. The 6 radial "spikes" arise from diffraction around the three struts that support the secondary mirror. In telescopes with a cross-shaped support structure for the secondary mirror, we get "diffraction spikes" that have the familiar cross shape seen on many astronomical images. Diffraction by quasi-linear suture structures in the eye may have led to the popular drawings of stars as "spiky" objects.

3.3 Using the convolution theorem

We can often use the convolution theorem to deduce the diffraction pattern of more complicated apertures given what we know about the Fourier transforms of simple apertures. The convolution of two functions f and g is

$$h(y) = f \star g = \int f(x)g(y-x)\mathrm{d}x\,. \tag{34}$$

The convolution theorem says that we can calculate this by multiplying the Fourier transforms of the functions, $\widetilde{f}, \widetilde{g}$, so

$$\begin{aligned} h(y) &= \mathcal{F}^{-1}[\widetilde{h}] \\ &= \mathcal{F}^{-1}[\widetilde{f}.\widetilde{g}] \\ &= \mathcal{F}^{-1}[\mathcal{F}[f].\mathcal{F}[g]] \end{aligned} \tag{35}$$

where $\mathcal{F}[g]$ denotes the Fourier transform of g and $\mathcal{F}^{-1}[\widetilde{h}]$ denotes the inverse Fourier transform[5] of $\widetilde{h}$.

3.3.1 Example 1

The triangular function shown on the right-hand column of the third row of Fig. 7 is obtained by convolving a tophat function with itself. Hence its Fourier transform is a sinc^2 function as shown in the left-hand column of Fig. 7.

3.3.2 Example 2

Consider an idealised Young's slits experiment with arbitrarily small slits - the far-field diffracted intensity should vary as equation 17. Mathematically, we can represent the two slits as being delta functions at $x = \pm d/2$ for slit separation d. We may convert the experiment into one with finite slits of width a by convolving the delta functions with a tophat of width a. Thus the expected diffracted intensity is the product of the Young's slits $\cos^2$ function and the sinc^2 tophat diffraction pattern:

$$I = I_0 \cos^2\left[\frac{kd\sin\theta}{2}\right] \text{sinc}^2\left[\frac{ka\sin\theta}{2}\right] . \tag{36}$$

3.4 The interpretation of diffraction theory

How reasonable is it to assert that we can replace a wave passing through an aperture by a continuous set of oscillators at the aperture? There are some obvious differences: the oscillators should radiate to both sides of the aperture, yet the diffracting wavefront only propagates in one direction. Huygens and Fresnel both proposed that the "secondary waves" only propagate in the same direction as the primary wave - we'll come back to this below, but this "ad hoc" assumption makes us realise that this framework is somewhat artificial - a set of coherent oscillators in the aperture is not the same as a wavefront propagating through the aperture, without some "fudging". Also, it made some sense when people thought that light waves propagate through an "aether" - an elastic medium supporting the wave motion. But after Maxwell's work in 1865 and then Poincaré's and Einstein's interpretation of the Michelson-Morley experiment, we now realise that electromagnetic waves can propagate through a vacuum, with no "luminiferous aether" being required. So now we regard these hypothetical secondary oscillators as being a mathematical convenience rather than having any physical reality.

Thomas Young (of Young's slits) had a different idea. He proposed that diffraction was caused by a boundary wave that is scattered from the surface of the obstruction around the aperture. Young found it difficult to make the theory work, and subsequently worked with Fresnel's ideas, but it has later been shown that Young's theory can also correctly reproduce diffraction effects from apertures.

So if we realise that no secondary oscillators exist, what is the status of our diffraction theory? Why do waves diffract at all? Fresnel's ideas were driven by experimental results on diffraction, trying to develop a theory that would explain the observations. In the context of electromagnetic waves (unknown to Fresnel), we could make oscillating electromagnetic field at an aperture by replacing the propagating wave at that location with a set of coherently-oscillating electric charges at that location, and we could arrange them so that we get the same oscillating fields: maybe not microscopically, because we have supposed a finite set of discrete electric charges, but would we be able to tell the difference in the resulting electromagnetic field, if we were a distant observer? Fresnel took this idea and ran with it (although he didn't know about electromagnetic fields) and much to everyone's surprise at the time, it agreed very well with experimental results.

Really, we should try to calculate the equations of diffraction from electromagnetic wave theory, however it turns out to be surprisingly difficult, with analytic solutions only for special cases. For most diffraction problems, on scales larger than the wavelength of light, Huygens-Fresnel theory works extremely well, predicted diffraction patterns may be readily calculated, and it forms the basis of almost all optical design and engineering. To derive diffraction theory mathematically from electromagnetism theory is beyond the scope of this course, but some of the steps towards this are discussed below.

[5] If we instead use the convention for the 2π factor adopted in the Mathematical Methods course, then equation 35 has a prefactor $\sqrt{2\pi}$.

3.5 Huygens-Fresnel diffraction

Huygens did not attempt to explain the diffraction of light with his theory, and Fresnel wanted to extend it to include diffraction effects. Fraunhofer diffraction is a special case of Fresnel's that applies in the far-field case. Huygens, and initially Fresnel, thought that light would consist of longitudinal waves, but it was during the course of his work that Fresnel realised that assuming a transverse wave would allow explanation of the phenomena that we now know arise because of the polarisation of light, discussed towards the end of this course.

In this section, for consistency with the convention adopted in Hecht, we write a wave as having complex exponential $\exp[\mathrm{i}(kr-\omega t)]$, and we factor out the time dependence, writing only the spatial part $\exp[\mathrm{i}kr]$.

If we are given a propagating wave, perhaps a spherical wave propagating from a point source as in Fig. 10, can we predict what the wave amplitude will be at some arbitrary location P? Fresnel's theory asserts that we can calculate the amplitude at P by considering the amplitude and phase of a secondary wave that could propagate from one part of the wavefront to P and integrating the contributions of the secondary waves over the entire wavefront:

$$u_P = C\int_S \frac{u_\rho}{r}\eta(\mathbf{n},\mathbf{r})\exp[\mathrm{i}kr]\mathrm{d}S\,, \tag{37}$$

where the integral is over the wavefront surface S, u_ρ is the wavefront amplitude at distance ρ from the source, the point P is a distance r from an element $\mathrm{d}S$ of the wavefront, k is the wavenumber, assumed the same for the wave and the hypothetical secondary wave, C is a normalisation constant and $\eta(\mathbf{n},\mathbf{r})$ is an *obliquity factor*, which is a function of the angle, θ, between the normal to the wavefront, $\mathbf{n}$ and the vector $\mathbf{r}$.

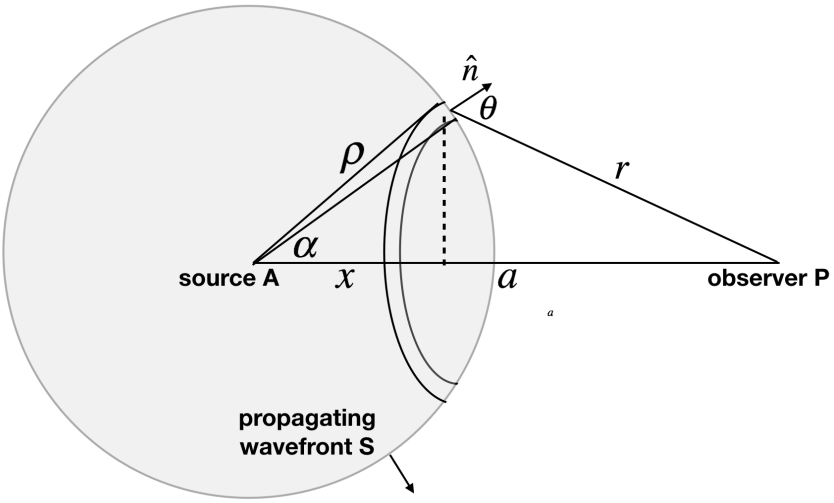


Figure 10: The geometry of a spherical wave, propagating from source A, of radius ρ at some instant of time, from which we wish to predict the amplitude at point P, showing the notation used in these notes. We will take the limit of P being very distant from A, $a \gg \rho$.

Huygens realised that to make this formalism work, we would need to suppress the backwards-propagating waves from the secondary wavelets, and Fresnel included an obliquity factor in his integral as a function of angle. This factor has been arbitrarily introduced to make the theory reproduce what we observe, but Kirchoff later was able to show that, for this geometry, the obliquity factor should be

$$\eta(\theta) = \frac{1+\cos\theta}{2}\,, \tag{38}$$

where θ is the angle between vectors $\mathbf{n}$ and $\mathbf{r}$ (see below).

We can calculate the value of the normalisation constant C for the assumed geometry in Fig. 10, because we know that the amplitude at point P should be

$$u_P = \frac{u_0}{a}\exp[\mathrm{i}ka]\,, \tag{39}$$

where a is the distance of P from the source, and hence we can find the value of C such that the Fresnel calculation gives the same answer (the following calculation is non-examinable). For simplicity let us

place the observer P very distant from S, so that the rays $\mathbf{r}$ may be assumed parallel, so $\theta \simeq \alpha$ and $r \simeq a - \rho\cos\alpha$. If we define $\mathrm{d}S$ to be an infinitesimal annulus around the axis joining A and P, then $\mathrm{d}S = 2\pi\rho^2 \sin\alpha \mathrm{d}\alpha$, $\mathrm{d}r \simeq \rho\sin\alpha\mathrm{d}\alpha$ and thus $\mathrm{d}S \simeq 2\pi\rho\mathrm{d}r$. The amplitude of the wavefront at radius ρ is

$$u_\rho = \frac{u_0}{\rho}\exp[\mathrm{i}k\rho] \tag{40}$$

and so the Fresnel integral becomes

$$\begin{aligned} u_P &\simeq C\int_{a-\rho}^{a+\rho}\frac{u_0}{\rho r}\eta\exp[\mathrm{i}k(\rho+a-\rho\cos\alpha)]2\pi\rho\mathrm{d}r \\ &\simeq 2\pi C\exp[\mathrm{i}k(\rho+a)]\frac{u_0}{a}\int_{-\rho}^{\rho}\frac{1+\cos\alpha}{2}\exp[-\mathrm{i}kx]\mathrm{d}x \\ &\simeq 2\pi C\exp[\mathrm{i}k(\rho+a)]\frac{u_0}{a}\int_{-\rho}^{\rho}\frac{1+x/\rho}{2}\exp[-\mathrm{i}kx]\mathrm{d}x \end{aligned} \tag{41}$$

,

where $x = \rho\cos\alpha$ and where we allow $r \simeq a$ in the denominator (valid for $a \gg \rho$) but retain the dependence on x in the complex exponential, as the variation of the phase with x cannot be neglected. This integral evaluates to

$$u_P = \frac{2\pi C u_0 \mathrm{i}}{k}\frac{\exp[\mathrm{i}k(\rho+a)]}{a}\left[\exp[-\mathrm{i}k\rho] - \frac{\sin(k\rho)}{k\rho}\right]. \tag{42}$$

For $\rho \gg \lambda$, $k\rho \gg 1$ and $\mathrm{sinc}(k\rho) \to 0$, so we may neglect the sinc term. But we expect $u_P = u_0\exp[\mathrm{i}ka]/a$, so we conclude that $C = k/2\pi\mathrm{i} = -\mathrm{i}/\lambda$. Thus we obtain the *Huygens-Fresnel diffraction integral* with this value of C inserted into equation 37,

$$\boxed{u_P = -\frac{\mathrm{i}}{\lambda}\int_S \frac{u_\rho}{r}\eta(\mathbf{n},\mathbf{r})\exp[\mathrm{i}kr]\mathrm{d}S\,.} \tag{43}$$

We have only obtained a value for C for this particular geometry, but it turns out to be a general result in the limit where the wavelength is small compared with the size of the system. Note the unexpected feature that C is imaginary: this is equivalent to introducing a phase shift of $\pi/2$ to the secondary wavelets (see equation 7). This doesn't seem to fit well with the hypothesis that we could replace the wave with a set of oscillating charges acting as Huygens' secondary oscillators, but we should emphasise that the Huygens-Fresnel integral is best viewed as a mathematical convenience that works, and which, on scales larger than the wavelength of light, does agree with more rigorous derivations, in cases where that is possible.

Because of its unsatisfactory origins and assumptions, as well as continuing opposition to the wave theory of light in the early 19th century, Fresnel's theory was not immediately accepted. In 1818 a physics competition was held by the French Academy, for which Fresnel submitted his diffraction theory. Poisson was one of the judges and showed that Fresnel's theory predicted a bright spot at the centre of the shadow of an opaque, circular object, which he felt disproved Fresnel's work. However, another judge, Dominique Arago, demonstrated that this was in fact observed, as shown in Fig. 11 (apparently the observation was already known but had been ignored until then), and Fresnel won the prize.

3.6 Fresnel-Kirchoff diffraction

Ultimately, we would like our theory of diffraction to be rooted in the theory of electromagnetic waves. A first step towards this was achieved by Kirchoff in 1883 assuming a scalar propagating wave (i.e. not yet a polarised, transverse electromagnetic wave) and has been subsequently extended in special cases using Maxwell's equations by Somerfeld in 1896 and others in later work (see Born & Wolf if you are interested). Derivation of Kirchoff's result is not required on our syllabus, but it would be worthwhile for you to look at the derivation in Hecht and the discussion about the assumptions of Kirchoff's approach by Brooker, to give you some faith in the Huygens-Fresnel integral. I don't normally recommend Wikipedia, but the article, "Kirchoff's diffraction formula" (https://en.wikipedia.org/wiki/Kirchhoff%27s_diffraction_formula) seems correct to me, at the time of writing these notes, and is also worth reading.

Kirchoff considered the spatial part of a solution to a scalar wave equation, such as

$$\nabla^2\psi + k^2\psi = 0\,. \tag{44}$$

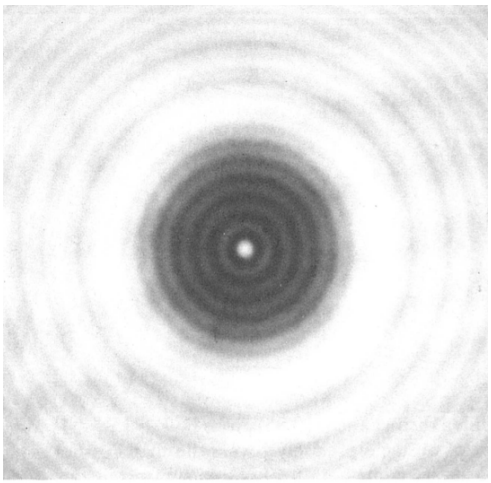


Figure 11: The shadow of a spherical ball-bearing being illuminated by a He-Ne laser beam, showing the diffraction pattern and in particular a bright spot at the centre of the shadow. Figure reproduced from Hecht.

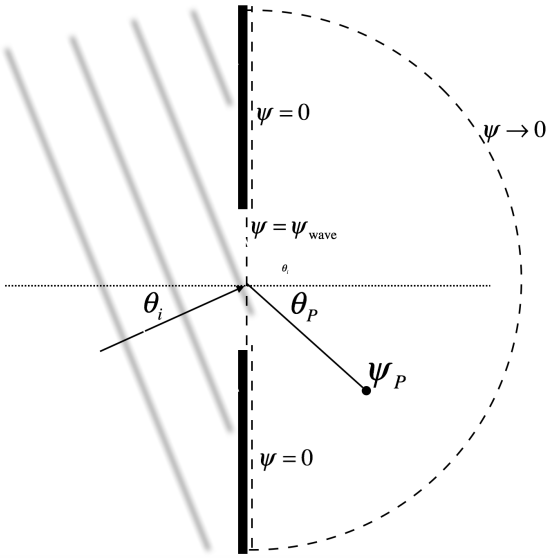


Figure 12: Kirchoff's boundary conditions for the example of the field ψ_P at any location P behind a diffracting aperture.

By imposing boundary conditions on the wave behind an aperture, as shown in Fig. 12, it is possible to find the value of ψ_P at any location in the volume shown - and that solution matches the Huygens-Fresnel integral, and is often known as the *Fresnel-Kirchoff diffraction integral*. It also shows that the obliquity factor should be

$$\eta = \frac{\cos\theta_i + \cos\theta_P}{2}, \tag{45}$$

For normal incidence, $\cos\theta_i = 1$ (note also that in the standard derivation there is usually a different sign of $-\cos\theta_P$, because of the definition of the angle with respect to the normal to a surface enclosing the point P). The boundary conditions themselves are assumptions - that the wave amplitude at the aperture is the same as would have been obtained with no aperture present; that the amplitude close to the obscuring structure that forms the aperture is zero; and that the right-hand surface as shown may be arranged to have negligible amplitude by taking that part of the surface to infinity - none of those assumed boundary conditions are exactly true, but are a good approximation in cases where the aperture is much larger than the wavelength of light.

3.7 From the Fresnel-Kirchoff integral to the Fourier transform

Fraunhofer diffraction arises as a special case of Fresnel diffraction, if we apply the Fraunhofer condition, that the wavefront phase is a linear function of position across the aperture. Consider a 1D aperture, which

has an aperture function $T(x)$, with value unity between the limits of the aperture, say $-b/2 < x < b/2$, and zero outside. The element $\mathrm{d}S$ in the Fresnel integral becomes $\mathrm{d}x$, and in the far-field limit, $1/r \to 1/a$, approximately invariant with x for $r \gg x$, and the complex exponential becomes $\exp[\mathrm{i}k(a - u'x)]$, where u' is the direction cosine, $u' = \sin\theta$ for the angle θ of the observer with respect to the axis parallel to the incident wavevector. For small variations in θ, we may neglect the variation of the obliquity factor, so $\eta \to 1$, and the Fresnel integral becomes

$$u_P(u') \simeq -\frac{\mathrm{i}}{\lambda}\frac{\exp[\mathrm{i}ka]}{a}\int_{-\infty}^{\infty} T(x)\exp[-\mathrm{i}ku'x]\mathrm{d}x \tag{46}$$

As before, let us define $k_x \equiv kx$. Then, changing variables,

$$\begin{aligned} u_P(u') &\simeq -\frac{\mathrm{i}}{\lambda}\frac{\lambda}{2\pi}\frac{\exp[\mathrm{i}ka]}{a}\int_{-\infty}^{\infty} T(k_x)\exp[-\mathrm{i}k_x u']\mathrm{d}k_x \\ &\simeq \frac{1}{2\pi a}\int_{-\infty}^{\infty} T(k_x)\exp[-\mathrm{i}k_x u']\mathrm{d}k_x\,, \end{aligned} \tag{47}$$

neglecting the benign phase prefactors, which will disappear from the intensity. With suitable definition of variables, the far-field (Fraunhofer condition) amplitude is proportional to the Fourier transform of the aperture function, as found in section 3.1.

3.8 Fresnel or Fraunhofer?

Fraunhofer diffraction is the special case of Fresnel diffraction, given by equation 43, when the Fraunhofer condition applies. How do we know when the assumption of Fraunhofer diffraction is good enough?

Let's go back to equation 24, but now consider the terms in x^2 that we dropped. For clarity, let treat this as a 1D problem.

$$\begin{aligned} r &= [r_0^2 + x^2 - 2ux]^{1/2} \\ &\simeq r_0\left[1 - \frac{ux}{r_0^2} + \frac{x^2}{2r_0^2} + \frac{1}{8}\left(\frac{2ux}{r_0^2}\right)^2\right]\,, \end{aligned} \tag{48}$$

keeping terms in x^2. For $u \ll r_0$, the next significant term is $x^2/(2r_0^2)$, so the Fresnel exponential becomes

$$\mathrm{d}u_P \propto \exp[\mathrm{i}kr_0]\,\exp[-\mathrm{i}ku'x]\,\exp\left[\mathrm{i}\frac{kx^2}{2r_0}\right]\,. \tag{49}$$

That extra exponential term can be neglected provided

$$\frac{kx^2}{2r_0} \ll \pi \tag{50}$$

or $F \equiv x^2/\lambda r_0 \ll 1$, where F is known as the *Fresnel number*. Small F means that we can assume Fraunhofer diffraction. Large F implies we should instead calculate the full Fresnel diffraction using equation 43 - but be careful what we mean by r_0: it is the distance to which rays would be brought to a focus as inferred at the diffracting aperture: i.e. ignoring any changes due to focussing optics. So the lens and screen of Fig. 13 still satisfies the Fraunhofer condition: the observer at the screen is effectively observing the diffraction pattern from the entrance aperture of the lens, as if they were at infinity.

3.9 Replacing the distant observer, the importance of Fraunhofer diffraction and the point-spread function

It wouldn't be very convenient if we had to keep absenting ourselves to infinity to study diffraction patterns. Luckily, there is a better way. If we insert a lens of focal length f followed by a screen, a distance f away from the lens, then that lens will focus any parallel rays from some direction onto the screen. Rays from a different angle would be focused at a shifted position on the screen, as indicated in Fig. 13. Hence we may view the far-field, Fraunhofer diffraction pattern by such an arrangement. The diffraction pattern will be imaged onto the screen with a scaling between angle and distance on the screen given by the lens focal length.

This concept also highlights the importance of Fraunhofer diffraction. Only light that enters the lens will form the image, and the lens therefore acts as an aperture (usually this is referred to as the *entrance aperture* of a telescope, for example). So if we make an image of a distant point-like source, such as a

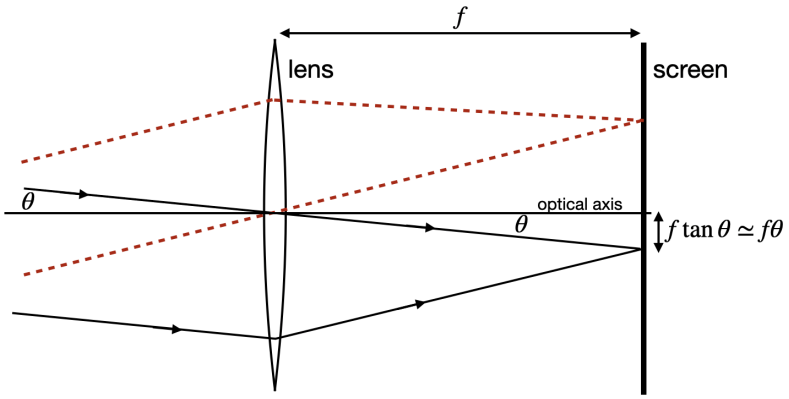


Figure 13: Imaging of parallel rays onto a screen, placed a distance f from a lens of focal length f. The coordinate on the screen of the focus of the rays is $f \tan\theta$, $\simeq f\theta$ for small θ. The dashed lines show the location of the image for parallel rays at a different angle.

star, say, where we just have a plane wavefront entering the aperture, what we will see on the screen is the projection of the diffraction pattern arising from the finite entrance aperture. If the distant object is extended (say, a galaxy), so that a range of ray angles enter the aperture, each set of parallel rays will form their own diffraction pattern. **The in-focus image formed by an optical system is convolved by the Fraunhofer diffraction intensity pattern of the entrance aperture** (or whatever aperture in the system most limits the extent of the wavefront being imaged). The diffraction pattern of the entrance aperture is often known as the "point spread function": it is how the image of a distant point source would appear. The consequences of this are discussed further in section 4.3.

4 Diffraction in optical instruments: telescope, microscope and diffraction grating spectrograph

4.1 A simple imaging lens

Fig. 14 shows schematically a ray diagram for a simple lens focussing light onto a screen, which forms an image on the screen, of an object placed at infinity, in the focal plane - remember the lens equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}\,, \tag{51}$$

for object distance u, image distance v and lens focal length f.

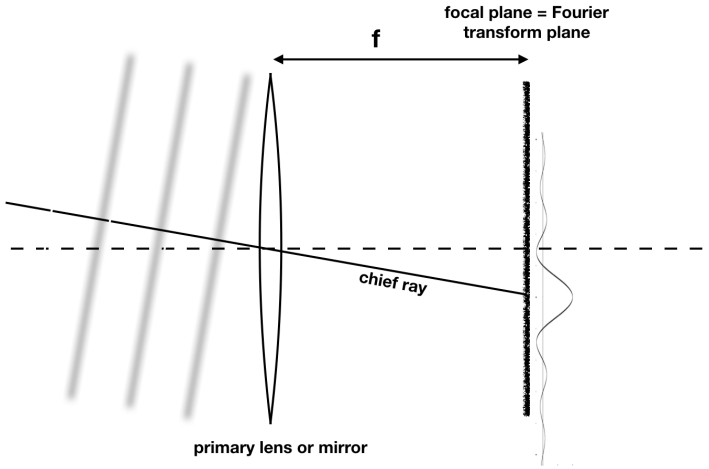


Figure 14: Schematic illustration of a simple imaging lens, showing an ideal incoming plane wavefront from a distant point source. f is the lens focal length.

The lens forms an image satisfying the Fraunhofer condition. The lens defines the *entrance pupil* of

the imaging system[6] . The *chief ray* from the source object passes through the centre of the entrance pupil and marks the centre of the resulting image on the focal plane. The image quality is limited by the finite entrance pupil, which acts as a low-pass filter, and causes the image of a point source to be blurred out into the Airy diffraction pattern of a circular entrance pupil (equation 33).

4.2 The astronomical telescope

An astronomical telescope is one designed to observe objects at infinity, without worrying about whether the image is inverted or not. Fig. 15 shows a two-lens or two-mirror astronomical telescope, now arranged

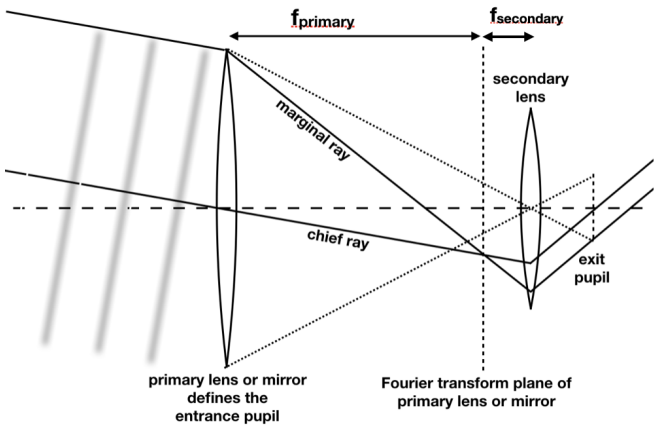


Figure 15: Schematic illustration of a two-lens astronomical telescope, adjusted to produce an image at infinity, and showing the *exit pupil*. Reflecting elements are more likely to be used in large telescopes, but the diagrams are clearer for a lens telescope.

to produce an image at infinity, suitable for viewing with the eye. In this case, parallel rays emerge. This configuration is obtained by forming an image at the focal point of the primary lens, which is also at the focal point of the secondary lens.

Recall that the angular magnification MP of an astronomical telescope is given by the ratio of the focal lengths of the primary and secondary lenses or mirrors (see Hecht or your first-year notes):

$$\mathrm{MP} = -\frac{f_{\mathrm{primary}}}{f_{\mathrm{secondary}}}\,, \tag{52}$$

where the negative sign indicates that the image is inverted.

The *exit pupil* is the image of the aperture stop (which is usually the same as the entrance pupil) formed by the secondary optical elements (here, the secondary lens). The chief ray also passes through the centre of the exit pupil. Its importance is that a *marginal ray* that enters at the edge of the entrance pupil, also passes through the edge of the exit pupil: all the light that enters the telescope leaves through the exit pupil. For optimum viewing by eye, we want to arrange the exit pupil to be close in both position and diameter to the entrance pupil of the eye (or a subsequent camera)[7].

The position along the optical axis of the telescope's exit pupil can be altered with the addition of a third *field lens* at or near the common focus of the objective and eyepiece lenses: this has no effect on the magnification perceived by the observer, but can change the location of the exit pupil so that it matches better the placement of the subsequent detection system (either a camera or the human eye).

4.3 The angular resolution and resolving power of a telescope

The angular resolution of a telescope is determined by the Fraunhofer diffraction limit applied to the entrance pupil, i.e. the width of the Airy function given by equation 33. The *Rayleigh diffraction limit* is

[6] Usually, the primary lens or mirror of an imaging system defines the entrance pupil, but in principle the entrance pupil could be smaller and/or offset in position, if an *aperture stop* is included: in this case, the entrance pupil is defined as the image of aperture stop as seen looking through any optics that precede it.

[7] In practice, the eye's imaging quality deteriorates towards the edge of its pupil, so telescope exit pupil diameters of $1-3\,\mathrm{mm}$ are generally preferred when looking through the eyepiece, even though the dark-adapted dilated pupil of the eye could be as large as $7\,\mathrm{mm}$.

reached when two stars are sufficiently close together that the maximum of one star's intensity diffraction pattern lies on the minimum of the other star's. For the Airy pattern this is reached for angular separation

$$\Delta\theta \simeq 1.22\frac{\lambda}{D} \tag{53}$$

for a circular telescope entrance pupil of diameter D, imaging at wavelength λ. Note that angular resolution is expressed in either radians or, more usually, seconds of arc ($1\,\text{arcsec} \equiv (1/3600)^\circ$).

The *resolving power* of the astronomical telescope is defined as

$$R \equiv \frac{1}{\Delta\theta} \simeq \frac{D}{1.22\lambda}\,, \tag{54}$$

with $\Delta\theta$ in radians, and in general we want R to be a large number.

Note an important point about the coherence assumptions in this. To derive the Fraunhofer diffraction pattern in the case of the astronomical telescope, we assume that the telescope receives a coherent, plane wavefront at its entrance aperture from a distant star (i.e. we treat the star as being a very distant, point source, so that its wavefront is plane to a very good approximation). Hence we calculate the diffraction pattern as a Fresnel integral of the wave amplitude, and then calculate the intensity distribution by taking the modulus squared.

A neighbouring star, at some slightly different angle of observation, creates a similar diffraction pattern, but displaced by a small amount on the focal plane. These two stars are incoherent with respect to each other, so we simply superimpose their intensity diffraction patterns to decide how close in angle they can be before their patterns overlap too much to be distinguished. There are no interference terms between the light waves from the two stars: they are incoherent light sources. The final image may be regarded as the convolution of the distribution of point sources with the diffraction pattern of the telescope (the point-spread function mentioned already in section 3.9).

4.4 The compound microscope

Fig. 16 illustrates the basic features of a compound microscope. The primary lens, known as the objective, acts also as the entrance aperture, and forms a real image of a nearby object at the location of the field stop. An eyepiece is used to form an image at infinity of that objective lens image. Hecht gives more

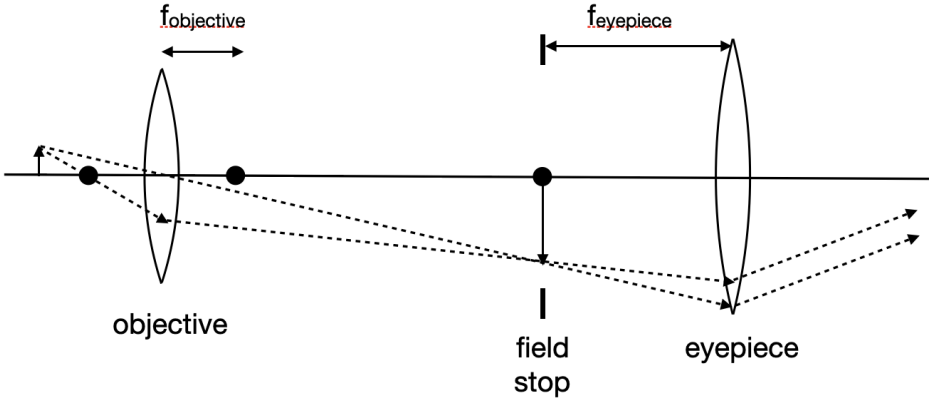


Figure 16: Schematic illustration of a compound microscope. The objective forms an image of a nearby object at the location of the field stop. The field stop is placed at the focal point of the eyepiece so that a magnified image is formed at infinity (parallel rays emerge from the eyepiece). Solid dots indicate the locations of the focal points of the lenses. In reality, the eyepiece is usually a compound lens system.

discussion of the microscope and its magnification.

4.5 The linear resolution and resolving power of a microscope with incoherent illumination

Let's now consider the effects of diffraction on the microscope image quality. The assumption of the sources being incoherent might not be true for objects being observed with a microscope, and here we will treat two cases: in the first case we assume that two closely neighbouring structures in the object

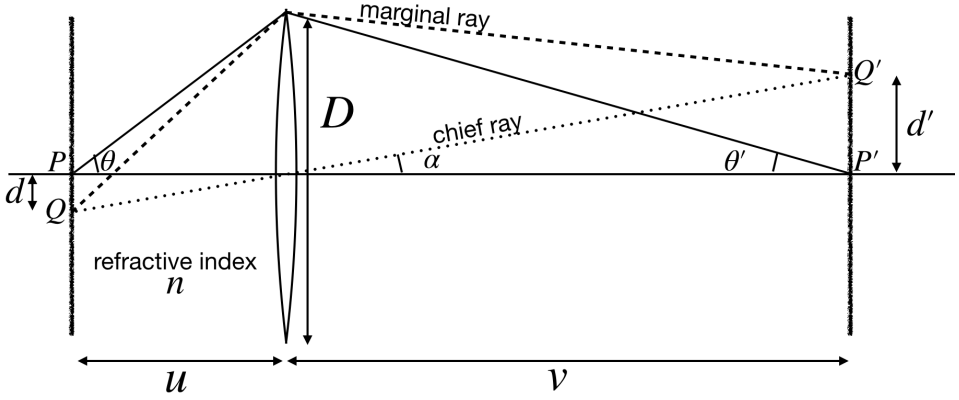


Figure 17: Ray geometry of the objective lens (with diameter D) of a microscope, observing an object a distance u from the lens. For clarity/simplicity we omit the eyepiece of a compound microscope and consider the image formed at the field stop at distance v. If two neighbouring structures, P and Q, in the object are separated by d in the object plane, their images, P' and Q', are separated by distance d' in the image plane. For high magnification, $v \gg u$ and $d' \gg d$.

are, like the astronomical telescope, incoherent with respect to each other, so we effectively repeat the calculation of section 4.3 but with a modified geometry.

Consider Fig. 17. The image formed by the lens has a diffraction pattern, and even though the object is not at infinity, we expect a similar diffraction pattern to the Fraunhofer pattern for the telescope, and for a circular lens, the first zero of the Airy pattern has an angular diameter given by $\sin\alpha \simeq 1.22\lambda/D$ (for more discussion on this point, see Born and Wolf).

We would like to find the smallest object separation d that can be measured, so now we need to consider the geometry. At the centre of the image diffraction pattern, all the wavefronts that contribute to the image have arrived in phase with each other, so that they constructively interfere and form a maximum. For the object P on the axis, the total optical path length s, equation 2, along the chief ray is

$$s_{\rm P} = \int n{\rm d}s = nu + s_{(\text{lens centre})} + v\,, \tag{55}$$

where $s_{(\text{lens centre})}$ is the optical path length introduced by the lens at its centre, and where we assume that there is a medium of refractive index n in the space between the lens and the object, and unity refractive index elsewhere. For the chief ray from location Q, s becomes

$$s_{\rm Q} = n(u^2 + d^2)^{1/2} + s_{(\text{lens centre})} + (v^2 + d'^2)^{1/2}\,. \tag{56}$$

For high magnification, $v \gg d'$ and $u \gg d$, so the variation in s is

$$s_{\rm Q} - s_{\rm P} \simeq n\frac{d^2}{2u} + \frac{d'^2}{2v} \tag{57}$$

to second-order in d.

For structure at P on the axis, the optical path length for the marginal ray, shown on the diagram, is

$$s_{\rm P}^{\rm marg} = n(u^2 + (D/2)^2)^{1/2} + s_{\rm lens} + (v^2 + (D/2)^2)^{1/2}\,, \tag{58}$$

and for structure at Q

$$s_{\rm Q}^{\rm marg} = n(u^2 + (D/2 + d)^2)^{1/2} + s_{\rm lens} + (v^2 + (D/2 - d')^2)^{1/2} \tag{59}$$

For rays at the maxima of the diffraction patterns, $s_{\rm P}$ and $s_{\rm P}^{\rm marg}$ must be equal, and similarly for Q. Given that the chief ray path lengths for P and Q are equivalent to first order in d, we therefore require the path lengths of the marginal rays to be equivalent to first order in d, which requires

$$\begin{gathered} n\frac{Dd}{(u^2 + (D/2)^2)^{1/2}} \simeq \frac{Dd'}{(v^2 + (D/2)^2)^{1/2}} \\ nd\sin\theta \simeq d'\sin\theta'\,. \end{gathered} \tag{60}$$

This is known as the *sine condition* and is discussed more thoroughly in Born & Wolf section 4.5.1. For high magnification, $v \gg D$, so $\sin\theta' \simeq D/2v$, and $v \gg d'$, so $d' = v\tan\alpha \simeq v\sin\alpha$. Combining the sine condition with the Rayleigh criterion,

$$\begin{aligned} nd\sin\theta &\simeq v\sin\alpha\frac{D}{2v} \\ &\simeq 1.22\frac{v\lambda}{D}\frac{D}{2v} \\ d &\simeq 0.61\frac{\lambda}{n\sin\theta}\,. \end{aligned} \tag{61}$$

d is the linear resolution of the microscope, i.e. smallest size of structure that can be resolved, and we see that this size is larger with increasing wavelength and smaller with increasing NA$= n\sin\theta$, where NA is known as the *numerical aperture*.

The resolving power of a microscope is defined as

$$R \equiv \frac{1}{d} \simeq \frac{\mathrm{NA}}{0.61\lambda}\,. \tag{62}$$

Now we see why it might be useful to arrange $n > 1$, for example by filling the space between the object and the lens by a high refractive index, transparent oil, as that improves the resolving power of the microscope. The importance of the numerical aperture for microscope resolving power was recognised by Abbé, who verified experimentally in 1866 that larger objective lenses produce sharper images.

4.6 Abbé's theory of the resolving power of a microscope for coherent illumination

However, Abbé also recognised that it is not satisfactory to consider the object as being incoherently illuminated. Fig. 18 (top panel) shows a ray diagram version of the imaging process for a microscope (again omitting the eyepiece of a compound microscope), where the target object is illuminated by a plane wavefront that is output from a condenser (collimating) lens. Now, the amplitude of light diffracted from one part of the object may interfere with the light diffracted from another part. The coherence between different parts of the object will not be complete, in reality there is partial coherence. We discuss partial coherence later in this course (section 7), but here we just qualitatively outline Abbé's treatment if we assume full coherence between the different parts of the object.

In this case, we can think of the object as being a diffracting screen (Fig. 18 lower panel). The objective lens collects the diffraction pattern - it forms the Fourier transform of the object in the focal plane of the light source, a distance f away from the lens, where f is the lens focal length. A smaller objective lens would miss the high-angle parts of the diffraction pattern, leading to a broader Fourier transform and a final image with worse resolution - the finite aperture of the lens causes it to act as a low-pass filter.

A calculation of the linear resolution in this case is discussed by Born & Wolf, section 8.6.3, but that derivation is not required for this course. The linear resolution limit they obtain is

$$d \simeq 0.82\frac{\lambda}{n\sin\theta}\,, \tag{63}$$

which is slightly worse than the linear resolution expected in the incoherent imaging case.

4.7 Aberrations in imaging systems

This subject is not listed on our syllabus, but it would be remiss to leave the topic of imaging systems without discussion of aberrations, which exist in all real optical instruments and which may ultimately limit their performance. The design of a microscope, telescope or camera may include obscurations of the wavefronts as they pass through, which lead to diffraction artifacts such as the diffraction spikes visible around bright stars in astronomical images (Fig. 9). Scattering of light by dust or surface defects may also affect the amplitude of the wavefront. However, some of the most important effects are caused by errors in the *phase* of the wavefront. Fig. 19 illustrates schematically what happens if an optical element such as a lens or a mirror has a non-perfect shape. The outgoing wavefront will be deformed accordingly - if the optical element is a mirror, the reflected wavefront will be perturbed by twice the perturbation in the mirror's surface shape. We say there is a *wavefront error*. There are three primary sources of wavefront error:

1. Compromises in the optical design - in astronomy we want a wide field of view, but for a telescope (or other instrument) with a finite number of optical elements, it is impossible to obtain a perfect

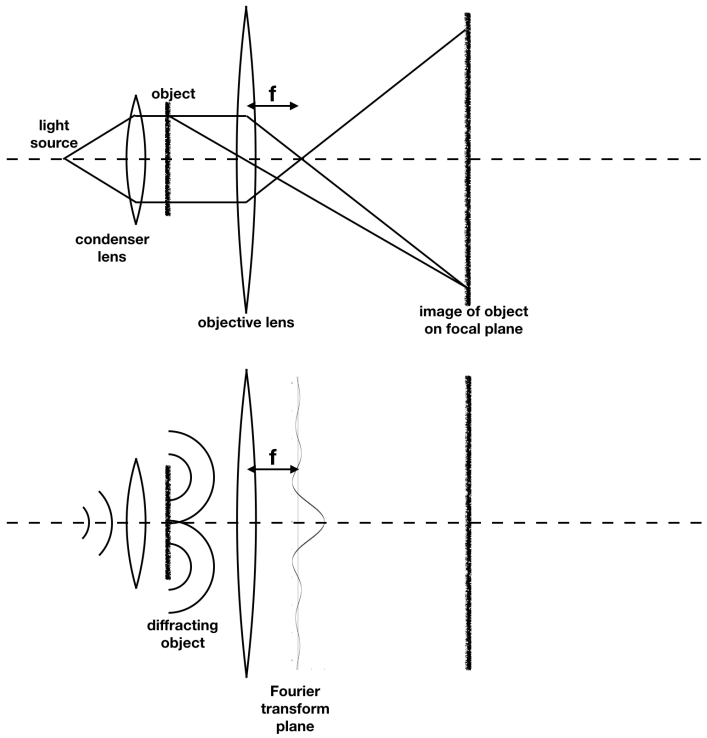


Figure 18: Illustration of the Abbé theory of imaging for a microscope (the usual eyepiece and stop of a compound microscope omitted for clarity). A condenser lens has been included which collimates the light before illuminating the object, so that a plane wave is incident. *top:* a ray diagram version of the imaging process; *bottom:* the object as a diffracting screen being imaged by the objective lens which acts as a Fourier filter. f is the lens focal length.

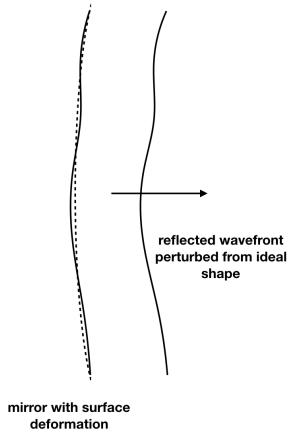


Figure 19: An optical element (lens or mirror) whose shape departs from the ideal shape introduces an error in the shape of the outgoing wavefront.

diffraction-limited image over a wide field of view - although some designs come close. Typically we find that as we observe off-axis, increasingly noticeable wavefront errors occur, arising from this design difficulty.

2. There may be wavefront errors introduced before a wavefront arrives at the instrument. In astronomy, refraction through turbulent density variations in the Earth's atmosphere cause rapidly varying wavefront errors (a refractive index variation causes a wavefront phase delay).

3. Manufacturing errors and differences between the actual optical element and its ideal optical design.

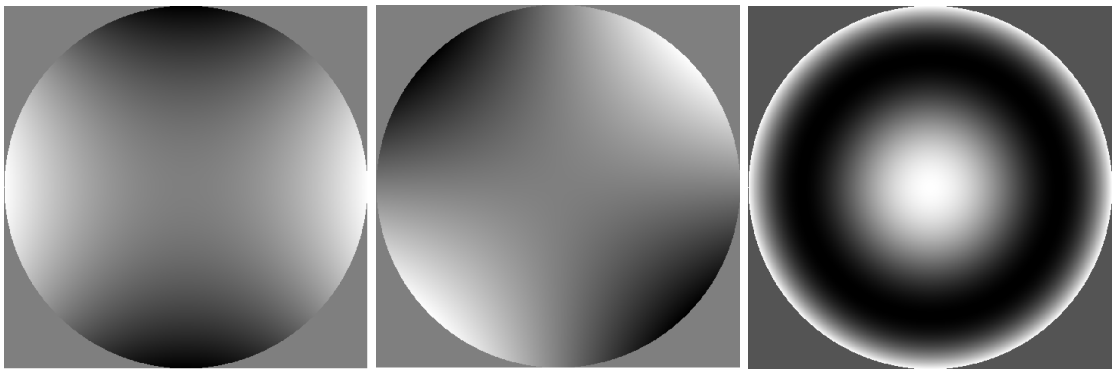


Figure 20: Zernike polynomials corresponding to pupil wavefront errors that cause astigmatism (left and centre panels) and spherical aberration (right panel)

Mathematically, a wavefront error can be represented by a phase delay in the wavefront, so in the Fraunhofer condition, the diffracted amplitude has the form

$$u_P(\mathbf{u}') \propto \int T(\mathbf{k_x}) \exp[\mathrm{i}\phi(\mathbf{k_x})] \exp[-\mathrm{i}\mathbf{k_x}.\mathbf{u}'] \mathrm{d}\mathbf{k_x} \,, \tag{64}$$

where $\phi(\mathbf{k_x})$ is the wavefront phase error as a function of vector position $\mathbf{k_x}$ on the 2D pupil (in fact, the exit pupil). Here for convenience we use of vectors: $\mathbf{k_x}$ defines coordinates on the aperture in wavenumber units, and the dot product with the vector of direction cosines $\mathbf{u}'$ gives the usual Fraunhofer condition phase term.

Wavefront errors of type 1 above are often represented mathematically by *Zernike polynomials*, which are an orthogonal set of polynomials defined on a unit circle (Frits Zernike was awarded the Nobel prize in 1953 for his invention of phase contrast microscopy). They have the nice property that individual Zernike polynomials correspond to well-known optical aberrations. For example, Fig. 20 (left and centre panels) shows the Zernike polynomials that cause astigmatism - a common defect in human eyes. The right panel of Fig. 20 shows the Zernike polynomial associated with spherical aberration, which plagued the Hubble Space Telescope when it was first launched, because of a manufacturing error in the primary mirror, and which was subsequently corrected by including corrections in the optics of the cameras that were subsequently replaced by space-walking astronauts. To avoid wavefront errors dominating the image quality, we need surface errors in mirrors to be less than about $\lambda/20$ in amplitude - a 2.4 m diameter mirror for a diffraction-limited optical telescope in space, such as the Hubble space telescope, has to be made with surface errors of less than 25 nm!

Wavefront errors caused by the Earth's atmosphere are hard to correct for, as they vary on timescales around 0.01 s. One technique is to attempt in real time to measure the distortion of the wavefront at the telescope, and apply an opposite deformation to a deformable mirror in the optical path, so as to cancel out the phase error - this problematic engineering feat is known as *adaptive optics* - see, e.g. http://www.eso.org/sci/facilities/develop/ao.html

4.8 The diffraction grating as a spectrograph

Diffraction gratings are the workhorse of spectroscopy, finding applications in every branch of experimental physics. As one example from astronomy, Fig. 21 shows a spectrum of the currently most distant known quasar, obtained with two diffraction grating spectrographs. The spectrum is redshifted by the cosmological expansion, so that the Lyman-α transition of hydrogen, normally at $\lambda = 121.6$ nm, is observed at $\lambda = 0.983\,\mu$m. The light was emitted only 770 million years after the Big Bang, but even at this early cosmological epoch, the heavy elements familiar to us on Earth were already abundant.

For the basic maths of diffraction gratings, refer back to section 2.4. First, we should modify that basic description to account for grating elements of finite width. We should convolve the ideal delta-function grating with a tophat finite-width function, as we have seen before this causes the diffraction intensity pattern to be multiplied by a sinc^2 function. If the diffracting element widths are a, the pattern becomes

$$I(\delta) = \frac{I_0}{N^2} \left[\frac{\sin(Nkd\sin\theta/2)}{\sin(kd\sin\theta/2)} \right]^2 \mathrm{sinc}^2(ka\sin\theta/2) \,. \tag{65}$$

The higher orders of the grating diffraction pattern are strongly suppressed. If $a = d/2$, as it might be for a transmission grating, then the resulting diffraction pattern is shown in Fig. 22.

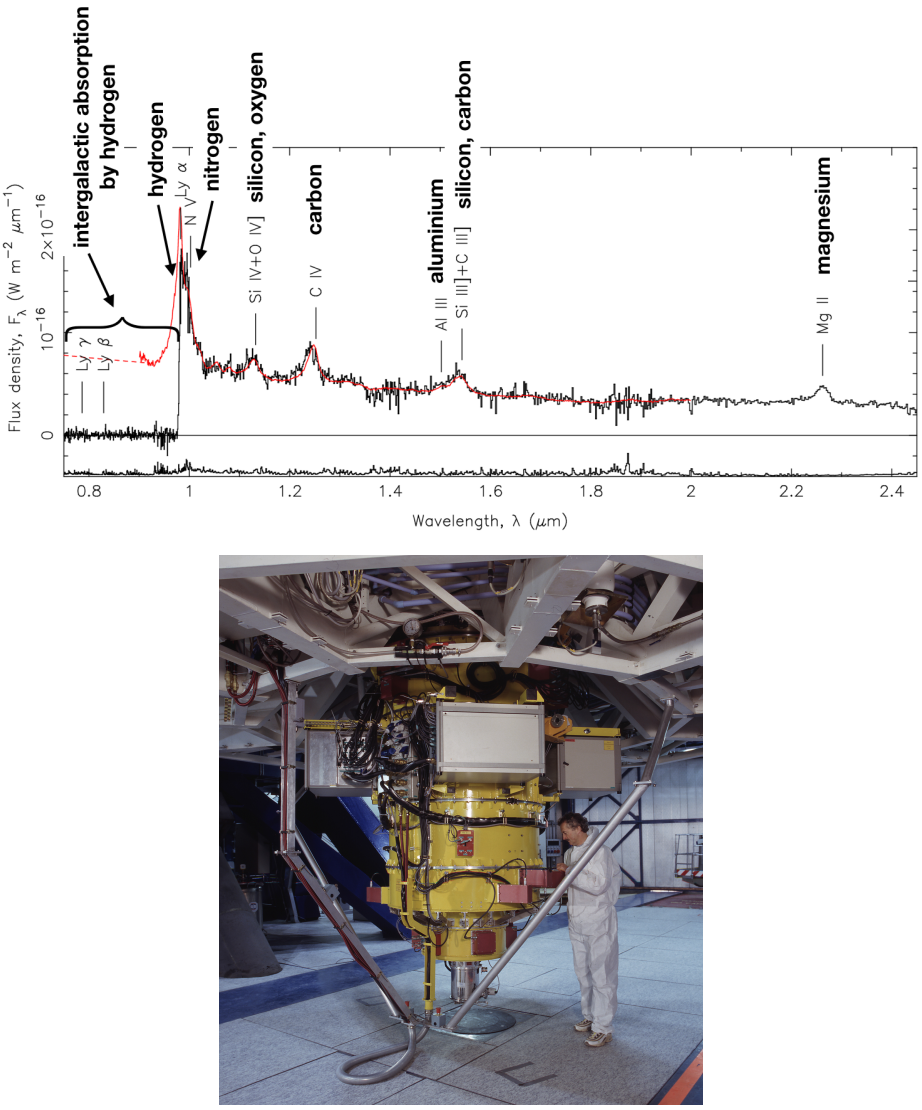


Figure 21: (upper panel) An infrared spectrum, showing flux density plotted against wavelength in μm, of the currently most distant known quasar, ULASJ112001.48+064124.3, whose light was emitted 13 billion years ago, only 770 million years after the Big Bang. The emission lines shown correspond to known atomic transitions of ionised C, N, O, Al, Si, Mg and neutral hydrogen, as well as strong absorption at wavelengths shorter than the Lyman-α transition by intervening intergalactic neutral hydrogen. Figure adapted from Mortlock et al. 2011, Nature 474, 616. (lower panel) One of the instruments used to obtain this spectrum, the FORS2 VPH-grating-prism (grism) spectrograph, on one of the Very Large Telescopes at the European Southern Observatory in Chile, for wavelengths $\lambda < 1\,\mu$m. The spectrograph is the large yellow thing. A portion of the telescope's 8-m. diameter primary mirror can be seen above the engineer's head. The spectrum at $\lambda > 1\,\mu$m was obtained using the reflection-grating Gemini North Near Infrared Spectrograph (GNIRS) on the 8-m. Gemini North telescope at Mauna Kea Observatory, Hawaii.

To use the diffraction grating in a spectrograph, we are not interested in what the pattern looks like for a single wavelength λ, but rather how the position in θ of the maximum of one chosen order n varies with λ. We say that the diffraction grating *disperses* the light to be analysed and allows us to measure its spectrum. The key characteristics of a diffraction grating spectrograph are as follows.

Diffraction order. We aim to select one particular order n, specified by the grating equation,

$$n\lambda = d\sin\theta \tag{66}$$

and measure the source spectrum as the dispersed intensity of light in that chosen order. Fig. 23 illustrates the spectra that can be observed in the various orders of a diffraction grating.

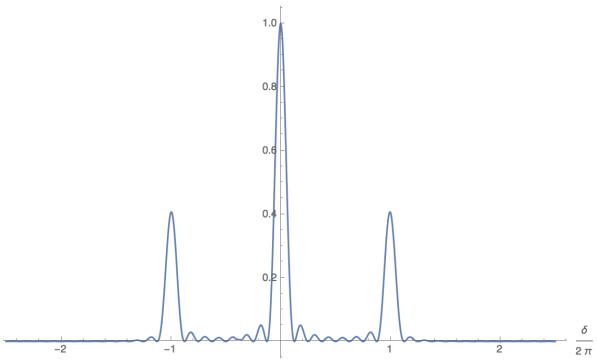


Figure 22: Suppression of higher orders of the diffraction grating response caused by finite width of the diffracting elements.

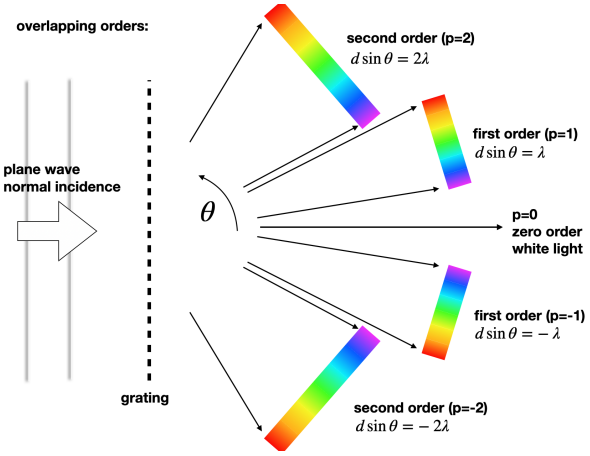


Figure 23: Schematic illustration of the diffracted spectra that may be observed from a diffraction grating, illuminated at normal incidence (incident light rays perpendicular to the grating). Note that the orders may overlap (see "free spectral range").

Angular dispersion. The separation $\mathrm{d}\theta$ between neighbouring wavelengths $\mathrm{d}\lambda$ is

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = \frac{n}{d\cos\theta}\,. \tag{67}$$

We immediately see that $n = 0$ is useless for spectroscopy, there is no dispersion of light. Higher orders have greater dispersion.

Free spectral range. We don't want to mix up diffraction from different orders in our spectrum, or we will not know whether some detected spectral feature is at some wavelength in order n or some different wavelength in order $n+1$. Suppose we want to measure a spectrum over the wavelength range λ_{blue} to λ_{red}. There will be overlap between orders if the range $\Delta\lambda_{\mathrm{FSR}} \equiv \lambda_{\mathrm{red}} - \lambda_{\mathrm{blue}}$ is sufficiently large that

$$\begin{aligned} n\lambda_{\mathrm{red}} &= n\left(\lambda_{\mathrm{blue}} + \Delta\lambda_{\mathrm{FSR}}\right) = d\sin\theta = (n+1)\,\lambda_{\mathrm{blue}}\,, \\ \Delta\lambda_{\mathrm{FSR}} &= \frac{\lambda_{\mathrm{blue}}}{n}\,. \end{aligned} \tag{68}$$

$\Delta\lambda_{\mathrm{FSR}}$ is the *free spectral range* and is the wavelength interval within which the order of diffraction can be uniquely determined. Our spectrum to be analysed must not contain wavelengths over a wider range than $\Delta\lambda_{\mathrm{FSR}}$ and filters may be needed to block wavelengths outside the range of interest. Higher orders have reduced free spectral range.

Resolving power. Two spectral features, separated in wavelength by $\Delta\lambda$, may be distinguished from one another in the grating spectrograph if the maxima of their diffraction patterns, for some chosen order n, are sufficiently far apart in angle, $\Delta\theta$. *Rayleigh's criterion* is that the diffraction maximum of one wavelength should fall on the minimum of the other - closer wavelengths than this cannot be resolved. The minima in the pattern at wavelength λ are separated by

$$\Delta\sin\theta = \frac{\lambda}{Nd}\,. \tag{69}$$

This is also the separation between a primary maximum and its neighbouring minimum. From the grating equation, or from the dispersion equation,

$$n\Delta\lambda = d\Delta\sin\theta \tag{70}$$

and so

$$\begin{aligned}\Delta\lambda &= \frac{\lambda}{Nn}\\ \frac{\lambda}{\Delta\lambda} &= Nn\,.\end{aligned} \tag{71}$$

The last expression is the *resolving power*, $\mathcal{R}$, of the grating. Resolving power increases with the number of lines on the grating, N and also with the order, n, but note that the free spectral range decreases as $1/(n+1)$: higher orders can "zoom in" to smaller spectral ranges with higher resolution. But note that the throughput of the spectrograph may also decrease with increasing n, depending on the blazing, discussed next.

Grating throughput and blazing. Spectrographs with high throughput efficiency can be made with either *transmission* or *reflection* gratings. The most efficient transmission gratings are "Volume Phase Holographic" (VPH) gratings, in which an interference pattern from a laser is copied onto a film, causing periodic refractive index variations, and thus periodic phase variations, see http://www.kosi.com/na_en/products/holographic-gratings/vph/intro-to-vph-for-transmission.php and http://adsabs.harvard.edu/abs/2000PASP..112..809B

Gratings may be used at oblique angles of incidence, in which case the grating equation becomes

$$d(\sin\theta - \sin\theta_i) = n\lambda\,, \tag{72}$$

where θ_i is the angle of incidence with respect to the normal to the grating surface, and θ, as before, is the angle in the diffraction pattern. Thus the zero order occurs when the diffraction angle equals the angle of reflection from the grating surface, as indicated in the left panel of Fig. 24. The efficiency of

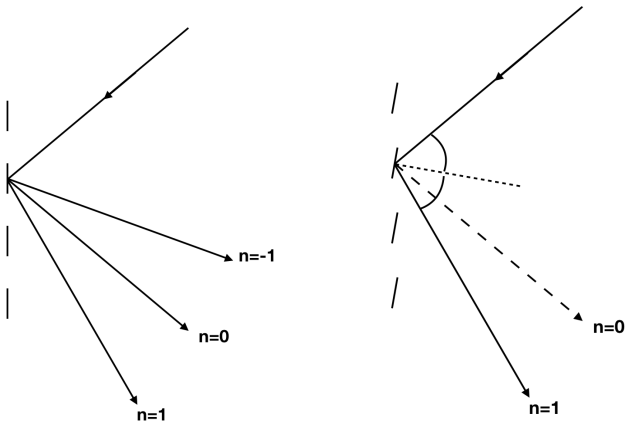


Figure 24: With a blaze angle introduced on a reflection grating, a particular order may be selected to have high throughput.

reflection gratings can be greatly increased if we arrange the diffracting elements to be angled so that the angle of reflected light from the surface coincides with the angle of diffraction for some choice of order n and wavelength λ: this maximises the throughput of the spectrograph at that order and wavelength - thus by appropriate choice of blaze angle, we can select a particular order, as indicated in Fig. 25.

4.9 A practical spectrograph design

Blazed reflection gratings are commonly used in spectrographs, owing to their high efficiency (i.e. high light intensity in the desired diffraction order for some specified central wavelength). The grating should be illuminated by plane parallel wavefronts, so collimating optics is required to ensure the image of the source slit as seen from the grating is at infinity. After light has been diffracted by the grating, it is brought to a focus at an exit aperture by further optics. For quantitative work a digital camera (usually with a Charge-Coupled Device, or CCD, as the photon detector) is used to measure the resulting spectrum. A standard design that incorporates these features is the Czerny-Turner spectrograph, illustrated in Fig. 26.

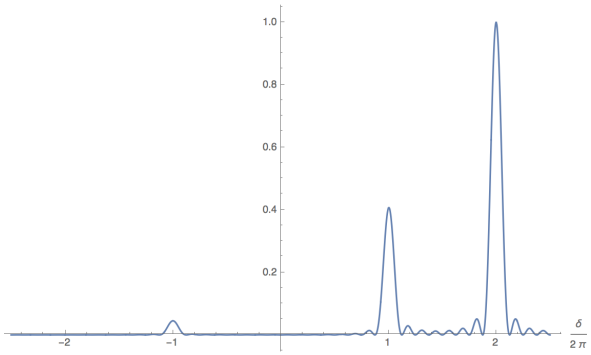


Figure 25: (left panel) With an oblique angle of incidence, the zero diffraction order appears at the angle of reflection from the grating surface. (right panel) if the individual elements are angled so that the angle of reflection from the elements equals the angle of diffraction for some desired diffraction order, the efficiency of the grating in that order can be greatly increased.

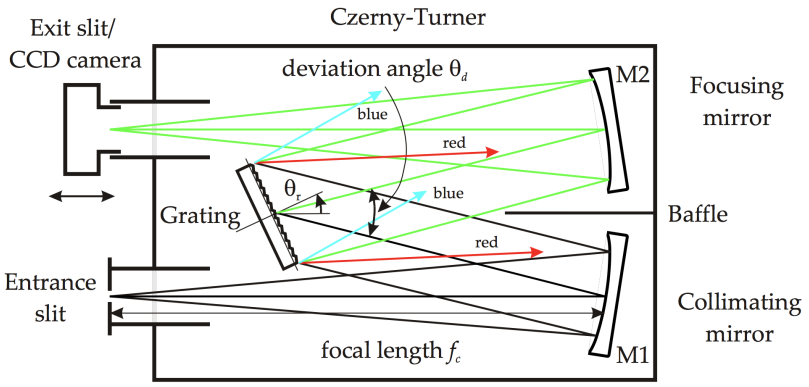


Figure 26: Schematic diagram of a standard spectrograph design, the Czerny-Turner reflection grating spectrograph. Diagram copied from Oxford Physics practical course notes, experiment OP24.

5 Interference fringes and interferometers

5.1 Interference by division of wavefront or amplitude

The experiments and instruments we have concentrated on so far are examples of *interference by division of wavefront*, in which we interfere the different parts of a wavefront - for example, in Young's slits we sample a wavefront at two locations and interfere those two wavefront components. For us to be able to see an interference pattern the wavefront needs to have *spatial* or *transverse coherence* - the different parts of the wavefront have to maintain a consistent phase with respect to each other. Transverse coherence will be discussed more later in section 7.

In the following sections we extend our ideas to *interference by division of amplitude*, in which the same portion of a wavefront is replicated by sharing the wave amplitude between two separate beams, which may then be recombined to form an interference pattern. Splitting a wavefront in this way can be achieved using surfaces which are partially transmitting and partially reflecting, such as a piece of plain glass or the surface of a transparent film, and in general may be achieved using a glass surface called a *beamsplitter*. In order to produce interesting effects, we often introduce a time delay in one of the beams before we recombine them. For us to see interference fringes in this case, we need the divided wavefront to maintain its phase for the duration of the time delay - the wavefront needs to have *temporal coherence*, also discussed later in section 7.

5.2 Localisation of fringes

Young's slits produce a fringe pattern in which the fringes are said to be *nonlocalised*, meaning that the fringes are not restricted to one particular plane, as illustrated in Fig. 27. The fringes shown in the figure

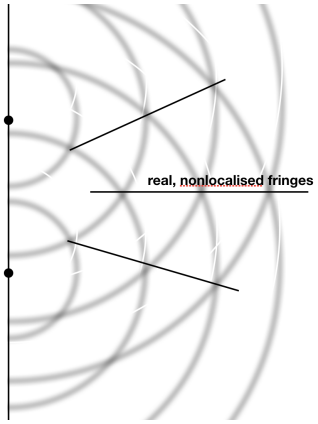


Figure 27: Illustration of real, nonlocalised fringes formed by Young's slits.

are nonlocalised because they form in the diffraction pattern of the slits - hence nonlocalised fringes are formed by point or slit sources. Nonlocalised fringes are also *real*, meaning that they can be viewed directly on a screen placed anywhere in the nonlocalised fringe region.

In the next section we will meet interference patterns where the fringes are *localised* to a particular plane. Localised fringes may be real or *virtual*, meaning that they can only be viewed if brought to a focus by a lens system (which includes the eye, of course). The plane of localisation might be at infinity: for example, if we view either Young's slits or a diffraction grating in the Fraunhofer condition using a lens system (so here we see an example of one diffracting system producing either type of localisation, depending on how we observe it). Extended light sources always produce localised fringes, but point sources may also produce localised fringes. This will become apparent in the following examples.

5.3 Interference fringes from reflections off two surfaces

5.3.1 Fringes of equal inclination.

Consider a point source illuminating two parallel, partially-reflecting surfaces, as in Fig. 28. The optical path distance for the ray reflected off the front surface between points O and A is $l_A = 2n_1 d \tan\theta_r \sin\theta_i$. The optical path distance for the ray reflecting from the rear surface to point B is $l_B = 2n_2 d/\cos\theta_r$. Hence the optical path difference between the rays is

$$\Delta l = \frac{2d}{\cos\theta_r}\left(n_2 - n_1 \sin\theta_i \sin\theta_r\right) . \tag{73}$$

But $n_1 \sin\theta_i = n_2 \sin\theta_r$, so

$$\begin{aligned} \Delta l &= \frac{2n_2 d}{\cos\theta_r}\left(1 - \sin^2\theta_r\right) \\ &= 2n_2 d\cos\theta_r \,. \end{aligned} \tag{74}$$

To form interference fringes that are *localised at infinity*, the rays may be brought to a focus with a lens or with the eye, approximately perpendicular to the emerging rays, and then the path difference between A and B causes interference fringes. The path difference results in a phase difference $\delta = k_0\Delta l - \pi$, where there is an extra relative phase shift of π because of the internal reflection at one of the surfaces (see Appendix A.6) and where k_0 is the wavenumber in vacuum. We use k_0 in this expression because Δl is the optical path length difference, which already includes the refractive index $n_{1,2}$ in its definition. Note that $n_2\lambda_2 = n_1\lambda_1 = \lambda_0$, where λ_0 is the wavelength in vacuum, and also therefore $k_2/n_2 = k_1/n_1 = k_0$. As in Young's slits, the interference of the two beams causes fringes whose intensity varies as

$$I \propto \cos^2\left(\frac{\delta}{2}\right) \tag{75}$$

and we get interference intensity maxima when $\delta/2$ is a multiple of π, that is when

$$\begin{aligned} 2k_0 n_2 d\cos\theta_r - \pi &= 2m\pi \\ d\cos\theta_r &= \frac{(2m+1)\lambda_2}{4} \,, \end{aligned} \tag{76}$$

for integer m. If we vary the angle at which we view the rays, we vary θ_r and hence the path difference varies, so the fringe intensity varies between bright and dark as a function of the inclination angle. Any one fringe traces a locus of equal inclination angle and such fringes are named *fringes of equal inclination*.

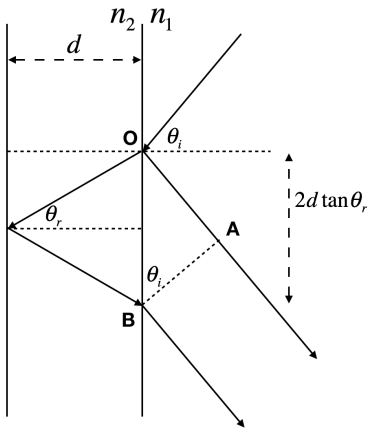


Figure 28: The geometry of fringes of equal inclination formed from parallel partially-reflecting surfaces, where the refractive index of the region of incident light is n_1 and of the region between the surfaces is n_2, and the separation of the surfaces is d.

We would only see the fringes if both rays enter the lens or eye, so for a point source we only observe fringes over a small region. An extended source produces fringes localised at infinity over a wider area. The different parts of an extended source are incoherent with respect to each other, so there is no interference due to division by wavefront, but the two reflections have split the amplitude of rays from each part of the source and these two components have been interfered by division of amplitude, even in an extended source. If the transverse separation between the emerging rays is larger than the lens system, we will not see fringes. Thus either the separation d must be small (so we tend to see these effects from thin films) or we must view them at close to normal incidence.

Point source illumination of this system may also produce *real, non-localised* fringes. Each reflecting surface produces a virtual image of a point source as indicated in Fig. 29, these virtual sources are coherent with each other and thus the system acts like virtual Young's slits. However, if the source is extended, the non-localised fringes randomly overlap and no pattern is seen.

In either case, we require the temporal coherence of the light source to be longer than the time delay between the light paths from the two surfaces.

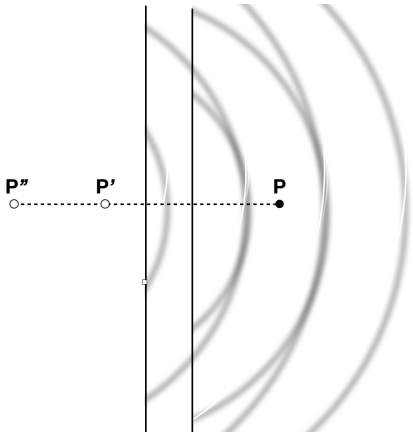


Figure 29: The virtual point sources P’ and P” created by the reflection of a point source P in two surfaces. Real, non-localised fringes may be formed on the source side of the surfaces.

5.3.2 Fringes of equal thickness.

More commonly in nature (such as light reflecting from a thin film of oil), we observe interference fringes from the two surfaces of thin films which vary in thickness (i.e. the surface separation d varies), known as *Fizeau fringes*. The thickness variations do not need to be much, a thickness variation of $\lambda/2$ would be sufficient to turn an interference maximum into a minimum (the path length is twice the thickness). So for any thin film or twin-surface arrangement whose thickness/separation variation is larger than a quarter-wavelength, the fringe pattern we see is dominated by the thickness variations, and following any one fringe traces out a locus of equal surface separation. Thus these are *fringes of equal thickness*.

Consider a wedge made of two plane surfaces (these could be the air gap between two plane glass surfaces with a small piece of dust creating a wedge whose separation variation is of order λ). First, consider illumination by a point source. For the very small wedge angle α, the wedge surface separation $d = x \tan\alpha \simeq x\alpha$ as a function of distance x away from the apex of the wedge. Substituting into equation 76, then assuming the wedge is filled with air and we are viewing at normal incidence to the surfaces, we will find maxima at regularly-spaced x locations:

$$x_m \simeq \frac{(2m+1)\lambda}{4\alpha}\,. \tag{77}$$

With parallel surfaces, the localised fringes were at infinity, as the outcoming rays were parallel. For the wedge, the rays may either be converging, forming real, localised fringes on the same side of the wedge as the source, or diverging, virtual fringes on the opposite side of the wedge from the source, depending on the orientation as indicated in Fig. 30. The virtual point sources also produce real, non-localised fringes

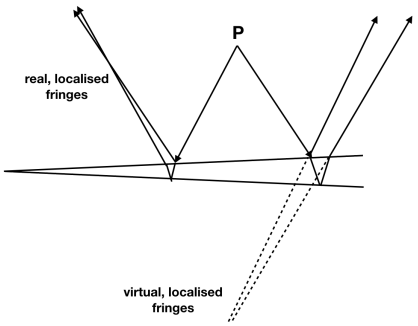


Figure 30: The localised fringes formed by a point source illuminating a wedge may be either real or virtual.

on the source size of the wedge, as in the parallel-surfaces case. However, if the source is extended, these fringes will again disappear, except when the path delay is small, so we can see fringes localised close to the plane of the wedge. Such fringes are equivalent to what is seen in the classic "Newton's Rings" experiment, where we essentially form a circularly-symmetric wedge (see e.g. Hecht).

5.4 The Michelson interferometer

5.4.1 The Michelson interference pattern

The Michelson interferometer interferes two beams which are produced by division of amplitude at a 50:50 beamsplitter, as indicated in Fig. 31. The beamsplitter is coated on one surface to enhance its reflectance and achieve the 50:50 split. As indicated on the diagram, the beams combined at the detector appear to

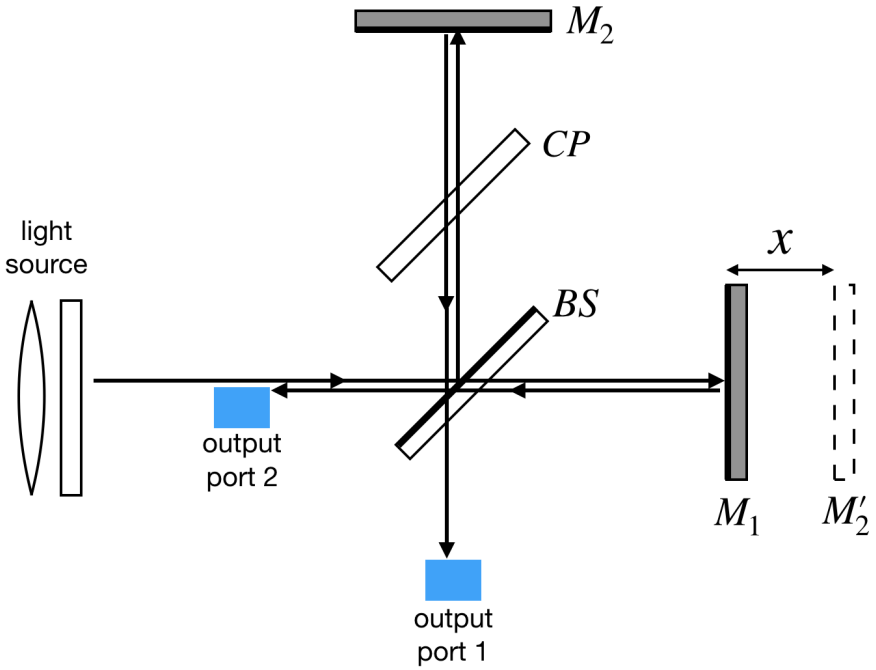


Figure 31: The Michelson interferometer. Beamsplitter BS sends light to mirrors M_1 and M_2 along two arms at right angles to each other. M_2' is the image of M_2 in BS when looking towards M_1. The compensating plate CP ensures the beams traverse equal thickness of glass along the two arms - important particularly for non-monochromatic light, where otherwise there would be a wavelength-dependent path difference. Interference fringes may be detected when the two beams are recombined at either of the two output ports.

have come effectively from reflections by the two surfaces M_1 and M_2', the latter being the image of M_2 in beamsplitter BS when looking towards M_1. Thus all of the results of section 5.3 immediately apply to the Michelson interferometer. If the two mirrors are aligned so that M_1 and M_2' are perfectly parallel, the system looks like section 5.3.1, and by symmetry the resulting fringes of equal inclination, localised at infinity, are circular. If the mirrors are not so aligned, the system looks like section 5.3.2 with curved fringes depending on the misalignment.

In the figure, the distance between M_1 and M_2' is x[8]. The phase difference between the beams is

$$\delta = 2kx\cos\theta \tag{78}$$

for rays at angle θ to the axis[9]. Thus there are maxima when $2x\cos\theta = m\lambda$ for integer m. The circular fringes produced when the mirrors are aligned are shown in Fig. 32. For aligned mirrors, the central intensity at $\theta = 0$ will be a maximum for $2x = m_0\lambda$, where m_0 is the order of that maximum for that value of t. Even for moderate x, m_0 could be large (e.g. 100,000). If we move outwards in radius, the next ring will occur when $2x\cos\theta_1 = (m_0 - 1)\lambda$ and the p^{th} ring when $2x\cos\theta_p = (m_0 - p)\lambda$. Combining with the expression for the central maximum we have $2x(1 - \cos\theta_p) = p\lambda$ and the order of the p^{th} ring is

[8] Note that mirror movement x introduces a path difference $2x$, so be careful to keep a consistent definition of what "x" is, some textbooks may choose to work with path difference rather than mirror movement.

[9] The beamsplitter introduces a phase shift of π between a beam which is internally reflected (at a glass-to-air interface) and a beam which is externally reflected (at an air-to-glass interface) - see Appendix A.6. Depending on which output port is used, there may be a net phase shift of π between the beams, in which case the locations of maxima and minima will be interchanged compared with that predicted by equation 78. Thus a setup in which the fringes are bright at one output port will result in dark at the other, and vice versa: energy is conserved.

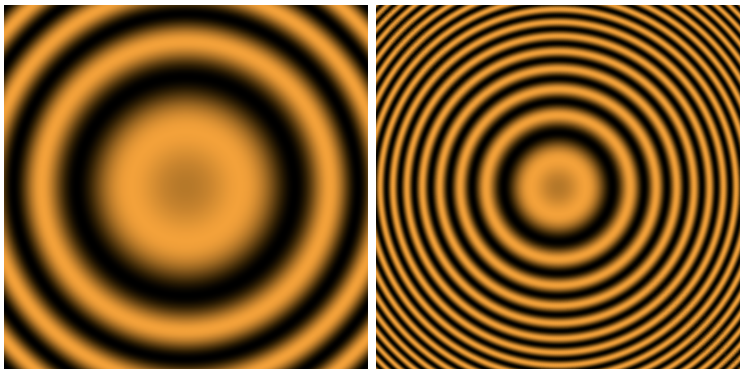


Figure 32: The Michelson interferometer ring pattern for aligned mirrors, shown for small path length difference (left panel) and larger path difference (right panel). Figure reproduced from Mathematica demonstration http://demonstrations.wolfram.com/MichelsonInterferometerAndHaidingerFringes/

$m_0 - p$. For small θ_p, $\theta_p^2 \simeq p\lambda/x$, and the relationship between the angular radii of neighbouring fringes is

$$\theta_{p+1}^2 - \theta_p^2 \simeq \frac{\lambda}{x} \tag{79}$$

or if we project the fringes onto a detector a distance f away from a lens of focal length f,

$$r_{p+1}^2 - r_p^2 \simeq \frac{f^2\lambda}{x}\,. \tag{80}$$

On axis ($\theta = 0$), the amplitude is

$$u(x) \propto (1 + \exp[2\mathrm{i}kx]) \tag{81}$$

and so the intensity is

$$I(x) = \frac{I_{\max}}{2}(1 + \cos(2kx)) \tag{82}$$

($k = 2\pi/\lambda$ is often written $k = 2\pi\bar{\nu}$ where $\bar{\nu} \equiv 1/\lambda$. Both k and $\bar{\nu}$ are often called "wavenumber").

5.4.2 The Michelson interferometer for length measurement

Michelson used the interferometer to measure the wavelength of light. By moving one mirror a measured distance and counting how many fringes cross some chosen radius, we can measure λ, although we might need to be inventive to get an accurate answer without counting zillions of fringes.

5.4.3 The Michelson as a spectrometer

To use the Michelson as a spectrometer we normally illuminate it with a plane wave at normal incidence and consider the on-axis intensity, after alignment of the mirrors. A monochromatic source produces a sinusoidal variation in fringe intensity as one of the mirrors is moved (equation 82). Suppose we observe a source comprising two narrow spectral lines at wavelengths λ_1 and λ_2, with equal intensity in the two components. These two wavelengths do not interfere of course, so we just add the intensities:

$$\begin{aligned} I(x) &= \frac{I_{\max}}{2}\left(1 + \cos(2k_1x) + 1 + \cos(2k_2x)\right) \\ &= I_{\max}\left(1 + \cos[(k_1 + k_2)x]\cos[(k_1 - k_2)x]\right)\,, \end{aligned} \tag{83}$$

using the usual manipulations for beats between two waves: the resulting "interferogram", a plot of intensity versus mirror displacement x, shows the beat-frequency envelope modulation characteristic of the presence of two wavelengths, as illustrated in Fig. 33. Thus the presence of two wavelengths is apparent from the envelope of the fringe pattern, and we can detect clearly the presence of the two wavelengths if we measure out to a maximum displacement, $x_{\max}$, such that we detect the pattern varying between its maximum and minimum values (*exercise: what happens to the interferogram if the two wavelengths do not have equal intensity?*),

$$\Delta k\, x_{\max} = \pi\,, \tag{84}$$

where $\Delta k = k_1 - k_2$. Thus

$$\Delta k = \frac{\pi}{x_{\max}} = \frac{2\pi}{L_{\max}} \tag{85}$$

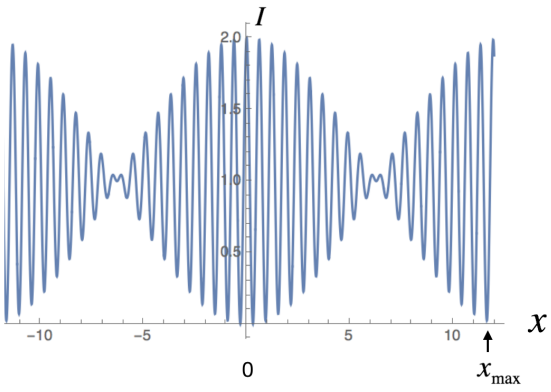


Figure 33: Schematic of Michelson interferogram from the measurement of a source with two closely neighbouring monochromatic spectral lines, showing the characteristic beats pattern as a function of mirror displacement, measured to a maximum displacement $x_{\max}$. Note that if the beamsplitter introduces a half-wavelength phase shift into one of the beams, the beams interfere destructively for $x = 0$, and the pattern is shifted so that its *minimum* occurs at $x = 0$.

or

$$\Delta\bar{\nu} = \frac{1}{2x_{\max}} = \frac{1}{L_{\max}} \tag{86}$$

where $L_{\max} = 2x_{\max}$ is the maximum path length difference in the interferometer. Thus the resolving power is

$$\mathcal{R} = \frac{\lambda}{\Delta\lambda} = \frac{k}{\Delta k} = \frac{\bar{\nu}}{\Delta\bar{\nu}} = \frac{L_{\max}}{\lambda}\,, \tag{87}$$

i.e. *the instrument resolving power is given by the maximum path length difference measured in units of the wavelength.* We saw this also for the diffraction grating, and is a general principle for spectrometers.

5.4.4 The Michelson interferometer as a Fourier Transform Spectrometer

Suppose now we use a light source with a broad spectrum of light, where the source spectral intensity $S(\bar{\nu})$ (measured in power per unit area per unit wavenumber) is some unknown function of wavenumber. Our task is to use the interferometer to infer the spectrum. Adding intensities across the spectrum is equivalent to integrating over $\bar{\nu}$,

$$\begin{aligned} I(x) &= \frac{1}{2}\int_0^\infty S(\bar{\nu})(1+\cos(4\pi\bar{\nu}x))\mathrm{d}\bar{\nu} \\ &= S + \frac{1}{2}\int_0^\infty S(\bar{\nu})\cos(4\pi\bar{\nu}x)\mathrm{d}\bar{\nu}\,, \\ I(L) &= S + \frac{1}{2}\int_0^\infty S(\bar{\nu})\cos(2\pi\bar{\nu}L)\mathrm{d}\bar{\nu}\,, \end{aligned} \tag{88}$$

where S is the integral of $S(\bar{\nu})$ over $\bar{\nu}$ and thus is the total power in the spectrum. In practice the detector will only be sensitive to some finite range of wavenumbers (or wavelengths) - we may simply set S to zero outside that range and allow the integration range to extend from 0 to infinity. The integral is the *cosine transform* of S and is closely related to the Fourier transform. In fact, the interferogram is a function of x which is both *real* and *even*. The spectrum $S(\bar{\nu})$ is also real. If we allow ourselves to define negative wavenumbers such that the spectrum also is an *even* function, $S(-\bar{\nu}) = S(\bar{\nu})$, then we may use the property that the Fourier transform of a real, even function is itself a real, even function. There are no odd components to either function, so the imaginary part of their Fourier transforms is zero. Thus we may write

$$\begin{aligned} I(L) &= \int_{-\infty}^\infty S(\bar{\nu})\exp(-2\pi\mathrm{i}\bar{\nu}L)\mathrm{d}\bar{\nu} \\ S(\bar{\nu}) &= \int_{-\infty}^\infty I(x)\exp(2\pi\mathrm{i}\bar{\nu}L)\mathrm{d}L\,, \end{aligned} \tag{89}$$

apart from some constant offsets, where we adopt the path length difference $L = 2x$ as the interferogram variable, and where the 2π terms have been retained in the exponents and hence there is no normalising

2π factor. The source spectral power (as a function of wavenumber or frequency) and the interferogram intensity form a Fourier transform pair, and using a Michelson in this way is known as a *Fourier transform spectrometer* (FTS). An FTS has the disadvantage that the Michelson must be scanned in x to obtain the interferogram, and must remain stable during the scan, which is unfavourable compared with using, say, a diffraction grating where the spectrum is obtained in one measurement with a static system. However, it has the advantage of potentially higher resolution than a grating, and in the case where the noise in the detector is independent of the signal, because all parts of the spectrum contribute simultaneously to the interferogram, there is a *multiplex advantage* (first pointed out by Peter Fellgett in his 1949 Ph.D. thesis) which leads to improved signal-to-noise ratio of the measurement. That condition on the detector noise is more often encountered at infrared wavelengths than optical or ultraviolet, so successful FTS have been used in the infrared. There is also the possibility of replacing the detector with an imaging camera, so that we end up with a "data cube" with each plane of the cube being an image of, say, an extended astronomical source such as a galaxy, and the third axis of the cube being the interferogram axis. A Fourier transform of the cube along that axis yields a new data cube containing a spectrum for every point on the image - "spatially resolved spectroscopy".

5.4.5 Autocorrelation

The Michelson interferometer adds together a light wave and a delayed version of the same wave, then takes the modulus-squared when the intensity is measured. As we average over time, this may be written as an integral over time,

$$\begin{aligned} I(\tau) &\propto \int [u(t) + u(t+\tau)][u^\star(t) + u^\star(t+\tau)]\mathrm{d}t \\ &= 2\int |u(t)|^2 \mathrm{d}t + 2\int u(t)u^\star(t+\tau)\mathrm{d}t\,, \end{aligned} \tag{90}$$

where delay time $\tau = L/c$ for path difference L and where we assume $u(t)$ falls to zero at small and large t. The last integral is known as the *autocorrelation function* of u. It looks a bit like the convolution of u with itself (compare with equation 34) apart from a sign change in the argument of the delayed u function. Apart from an additive constant, its Fourier transform is

$$\widetilde{I}(\omega) = \widetilde{u}\widetilde{u}^\star(\omega) = |\widetilde{u}(\omega)|^2 \propto S(\omega)\,, \tag{91}$$

which is a statement that the Fourier transform of the signal's time autocorrelation function is the spectral power density of the signal, an example of the *Wiener-Khinchin theorem*.

5.4.6 Aether drift experiments

In the 19th century it was thought that light propagates through a "luminiferous aether", so in Galilean mechanics, if the light source is moving with respect to the aether, there should be a change in the measured speed of light (Galilean addition of velocities). Fizeau attempted to measure this in 1851 using a symmetric interferometer with light passing through water running in opposite directions along the two arms. He measured a velocity shift, but smaller than expected. Fresnel had proposed in 1818 that the aether is partially dragged along with glass and thus presumably water also - although this required the drag to depend on the wavelength of light. Michelson and Morley improved the experiment in 1886, using the interferometer shown in Fig. 34[10]. The results confirmed that the velocity of light and velocity of the flowing water do not add linearly. At the time this was interpreted as confirming "partial aether drag" but was recognised as being problematic. Einstein quoted this experiment as being particularly important to his development of special relativity.

In 1881 Michelson attempted to detect the aether drift due to the Earth's motion using his (now-called) Michelson interferometer - he obtained a null result but it was realised he had overestimated the expected effect by a factor 2, and his initial experiment was insufficiently sensitive.

In 1886 Michelson and Morley repeated it with an improved interferometer, with folded arms of total path length 11 m, constructed in a noise free basement on a stone slab floating on a bed of mercury! They again obtained a null result. This could not easily be explained by the aether being dragged along with the Earth, because the Fizeau experiment showed that the "measured" aether drag was only partial - some change in the speed of light in the moving water was measured, but not as much as expected from Galilean addition of velocities. The null result led in around 1887 to Lorentz's phenomenological derivation of the Lorentz transformations (named by Poincaré) and then to Einstein's interpretation of those as arising from an invariant speed of light.

[10] This design, where two beams pass in opposite directions around the same path, is known as a *Sagnac interferometer*. It has the advantage of good stability, because external perturbations affect both beams by the same amount.

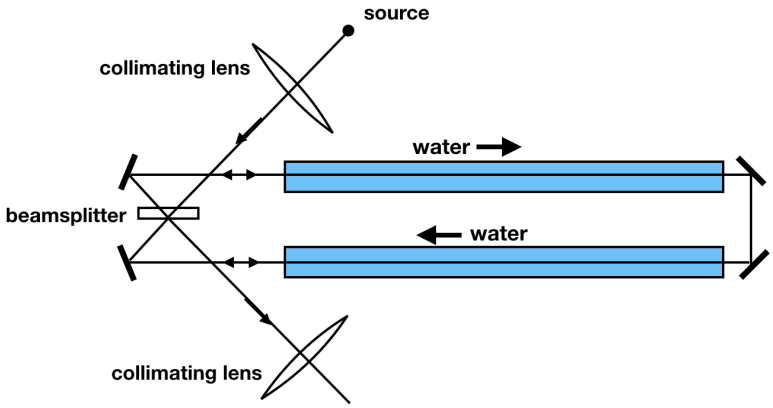


Figure 34: Schematic ray diagram of Michelson and Morley's improved interferometer for the Fizeau experiment. Light is split at the beamsplitter and the two beams pass in opposite directions through flowing water, and are then recombined to form an interference intensity.

5.5 The Fabry-Perot etalon and spectrometer

5.5.1 The Fabry-Perot intensity function

In this section we consider interference of multiple reflections inside two surfaces, known as a Fabry-Perot etalon, invented by Charles Fabry and Alfred Perot[11] around 1899. Schematic diagrams are shown in Figs 35 and 37. An etalon has fixed, reflecting surfaces, the separation and alignment of which is controlled by adjustable spacers. In an interferometer, the spacing may be varied over some range. The gap between the surfaces may be air or may be some other medium, depending on the application. The surfaces must be coated to obtain high reflectivity.

In our treatment, the two reflecting surfaces are the same (in a real system they may not be), but even then, there will be a phase shift on reflection depending on whether the reflection is internal (from glass back into glass, reflected at a glass-to-air boundary) or external (from air back into air, reflected at an air-to-glass boundary). We designate the external reflection and transmission coefficients as r, t respectively, and the internal coefficients as r', t'. There will always be a phase shift between r and r' regardless of the detailed geometry of the reflecting surfaces, and for dielectric materials (i.e. non-conducting) that relative phase shift is always π radians, i.e. $r = -r'$. The transmission coefficients are the same, $t = t'$. On each double reflection within the etalon, the wave picks up an extra phase $\delta = 2kd\cos\theta$, so after one double reflection the emerging ray that is transmitted through the etalon has complex amplitude $u_0 t^2 r^2 \mathrm{e}^{-\mathrm{i}\delta}$, the next has $u_0 t^2 r^4 \mathrm{e}^{-2\mathrm{i}\delta}$ and so on (writing $tt' = t^2$).

If we bring the emerging rays to a focus with a lens, they interfere, forming fringes of equal inclination localised at infinity, and the total amplitude of light reflected from the etalon is

$$u_r = -u_0 r\left(1 - t^2\mathrm{e}^{-\mathrm{i}\delta}\left(1 + r^2\mathrm{e}^{-\mathrm{i}\delta} + (r^2\mathrm{e}^{-\mathrm{i}\delta})^2 + \ldots\right)\right) \tag{92}$$

(after writing $r' = -r$) and the total amplitude of light transmitted through the etalon is

$$u_t = u_0 t^2\left(1 + r^2\mathrm{e}^{-\mathrm{i}\delta} + (r^2\mathrm{e}^{-\mathrm{i}\delta})^2 + \ldots\right). \tag{93}$$

This is a similar geometric series to equation 20, but for $r < 1$ and allowing the terms to extend to infinity, we find

$$u_t = u_0 t^2 \frac{1}{1 - r^2\mathrm{e}^{-\mathrm{i}\delta}}. \tag{94}$$

Thus

$$u_t u_t^\star = |u_0|^2 t^4 \frac{1}{(1 - r^2\mathrm{e}^{-\mathrm{i}\delta})(1 - r^2\mathrm{e}^{\mathrm{i}\delta})}, \tag{95}$$

[11] In his scientific papers, he used the spelling Pérot, but apparently his official record was spelled without an accent (https://en.wikipedia.org/wiki/Alfred_Perot).

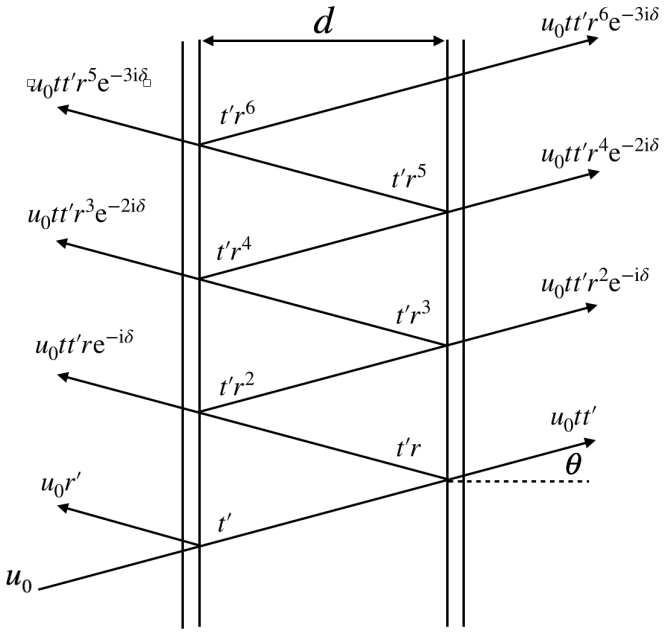


Figure 35: Schematic of multiple reflections within a Fabry-Perot etalon. Light is incident from the left with amplitude u_0. The inner surfaces of the etalon are coated to enhance their reflection coefficients. On an internal reflection, light is reflected (from glass back into glass) with amplitude r' and transmitted (from glass into air) with amplitude t'. On an external reflection, light is reflected (from air back into air) with amplitude r and transmitted (from air into glass) with amplitude t. To keep track of the factors of r, t, r', t' on each reflection they are written above the relevant ray inside the etalon. The final transmitted or reflected amplitude of each ray that emerges also includes the phase factor arising from the path length of the round trip inside the etalon, relative to the first ray.

and hence the transmitted intensity I_t, given incident intensity I_0, is

$$\begin{aligned} I_t &= \frac{I_0 T^2}{1 + R^2 - 2R\cos\delta} \\ &= \frac{I_0 T^2}{1 + R^2 - 2R(1 - 2\sin^2\delta/2)} \\ &= \frac{I_0 T^2}{(1-R)^2 + 4R\sin^2\delta/2} \\ &= \frac{I_0 T^2}{(1-R)^2}\frac{1}{(1 + \Phi\sin^2\delta/2)}, \end{aligned} \tag{96}$$

where we define the intensity transmission and reflection coefficients $T = t^2$, $R = r^2$ and where

$$\Phi = \frac{4R}{(1-R)^2} \tag{97}$$

is known as the *coefficient of finesse*, not to be confused with the *finesse*, which is coming soon. If there are no losses then $T = 1 - R$. The right-hand term $\mathcal{A} = 1/(1 + \Phi\sin^2(\delta/2))$ is known as the *Airy function*, and is plotted in Fig. 36. Maxima occur when $\delta = 2m\pi$ for integer m, so all the emerging rays are interfere constructively. Higher values of reflectivity and hence Φ mean that there are more reflections within the etalon before the amplitude diminishes, creating sharper maxima. The Fabry-Perot is acting as a *resonant cavity* with some Q-factor, in resonance when δ is an integer number of wavelengths.

5.5.2 The fringe pattern

Fringes may be observed from an extended source[12] of quasi-monochromatic light, focusing the fringes on to a screen as indicated in Fig. 37.

[12]Remember that light from different parts of the extended source are incoherent with respect to each other, but that each element of the source is responsible for a fringe pattern on the screen, and for fringes of equal inclination, the intensity patterns from those elements have maxima and minima at the same locations.

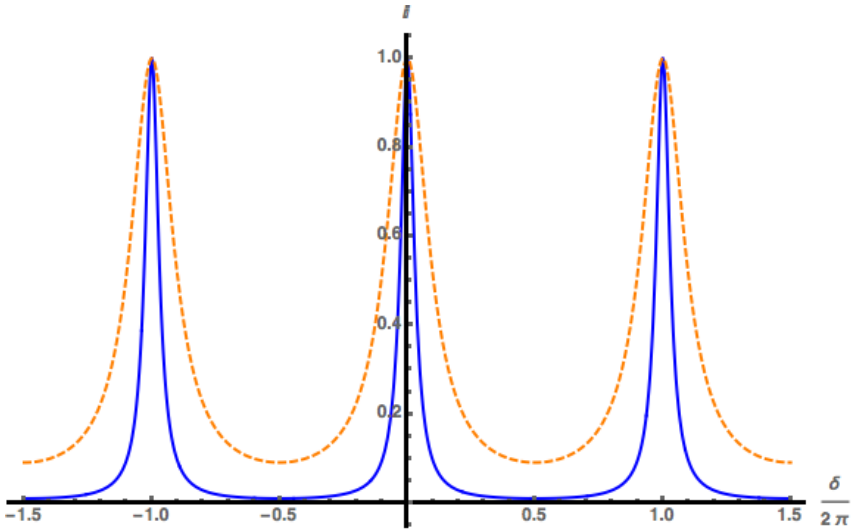


Figure 36: The Fabry-Perot Airy function, intensity plotted against $\delta/2\pi$, for coefficient of finesse $\Phi = 100$ (solid curve) and $\Phi = 10$ (dashed curve).

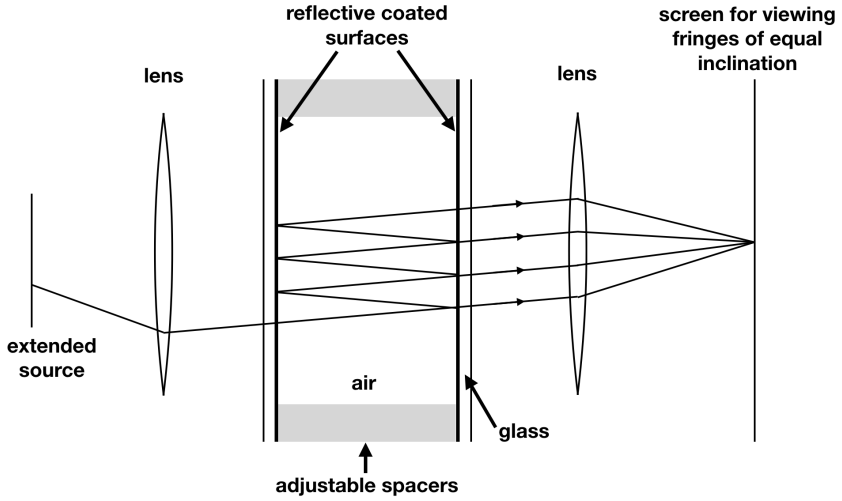


Figure 37: The setup for observing circular fringes with an air-gap Fabry-Perot etalon. Light from an extended source is collimated onto the etalon, and the resulting fringes of equal inclination appear as a circular pattern (for parallel, aligned etalon surfaces) on the screen (fine adjustment of the spacing is often made by increasing or decreasing the force with which the plates are pushed together by an external spring).

Fig. 38 shows the ring pattern for two values of R. The angular radii of the maxima, θ_m, occur when

$$\cos\theta_m = \frac{m\lambda}{2d} \tag{98}$$

so the radii of fringes on the screen, placed at the lens focal length f, is

$$r_m \simeq f\cos^{-1}\left(\frac{m\lambda}{2d}\right). \tag{99}$$

Note how the fringes become closer together at larger angles. For small angles, $\cos\theta \simeq 1 - \theta^2/2$, so

$$\begin{aligned} \cos\theta_m &\simeq 1 - \frac{\theta_m^2}{2} = \frac{m\lambda}{2d} \\ \theta_m^2 - \theta_{m+1}^2 &\simeq \frac{\lambda}{d}, \end{aligned} \tag{100}$$

i.e. for small angles the fringes are spaced evenly in angle-squared (or radius-squared if focused onto a screen).

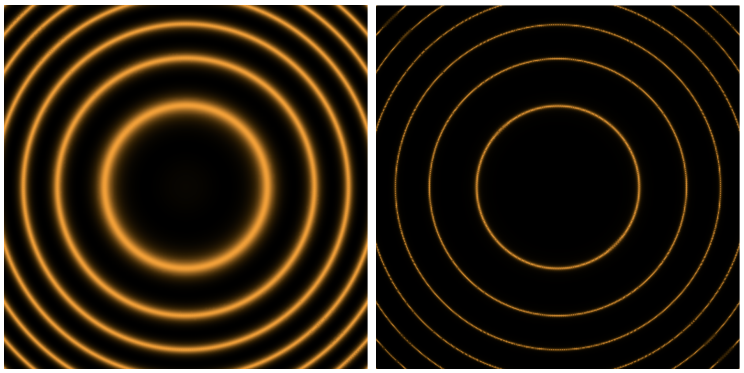


Figure 38: The Fabry-Perot ring pattern, shown for $R = 0.5, \Phi = 8$ (left panel) and $R = 0.9, \Phi = 360$ (right panel). Notice the sharpness of the rings compared with the Michelson rings. Figure reproduced from Mathematica demonstration http://demonstrations.wolfram.com/FabryPerotInterferometerAndRings/

Light of differing wavelengths appear with differing radii on the screen, allowing interferometric use of the FP. We may also use the FP as a scanning interferometer, if we illuminate it with a plane wave so that all the light stays on-axis, and measure how the intensity varies as we vary d.

5.5.3 Finesse

Finesse F is defined as the ratio of the separation between maxima to the width of the maxima, on axis. The fringe separation is $\Delta\delta = 2\pi$. The width of the fringes may quoted as the value of δ at which I falls to half its maximum value, i.e. when

$$\Phi \sin^2 \delta_w/2 = 1\,. \tag{101}$$

For high Φ,

$$\sin \delta_w/2 \simeq \delta_w \simeq 2\Phi^{-1/2}\,. \tag{102}$$

The full width at half maximum (FWHM) is twice this value, and so the finesse is

$$F = \frac{\pi\Phi^{1/2}}{2} = \frac{\pi R^{1/2}}{(1-R)}\,. \tag{103}$$

Don't get "finesse" and "coefficient of finesse" mixed up.

5.5.4 The Fabry-Perot interferometer as a spectrometer: resolution, free spectral range and resolving power

The resolution in δ is FWHM $= 2\delta_w = 2\pi/F$. If we vary the wavenumber, $\bar{\nu}$, $\delta = 2\pi\bar{\nu}2d$ on axis, so, differentiating, $\mathrm{d}\delta = 4\pi d\,\mathrm{d}\bar{\nu}$ and thus the FWHM in wavenumber is

$$\Delta\bar{\nu} = \frac{1}{2dF}\,. \tag{104}$$

The free spectral range is the interval between maxima in wavenumbers, so from the definition of F is just $\Delta\bar{\nu}_{\mathrm{FSR}} = F/2dF = 1/2d$.

On axis maxima occur when $m\lambda = 2d$, or $\bar{\nu} = m/2d$, so the resolving power is

$$\mathcal{R} = \frac{\bar{\nu}}{\Delta\bar{\nu}} = \frac{m2dF}{2d} = mF\,. \tag{105}$$

5.5.5 The effective number of reflections

We assumed the geometric series in equation 93 extended to infinity, but the terms in that series fall off in amplitude as powers of the reflection coefficient R. Thus we could say that there is some "effective" number of reflections, N_{eff}, which occur before the ray amplitude becomes insignificant. We saw in equation 87 that $\mathcal{R} \simeq L_{\mathrm{max}}/\lambda$, so if we apply that principle here we find

$$\mathcal{R} \simeq \frac{L_{\mathrm{max}}}{\lambda} \simeq mF \tag{106}$$

so

$$L_{\mathrm{max}} \simeq 2dN_{\mathrm{eff}} \simeq \lambda mF \simeq 2dF \tag{107}$$

so $N_{\mathrm{eff}} \simeq F$, i.e. the finesse is the effective number of reflections that contribute to the interference.

5.5.6 Use of the Fabry-Perot

Because it has a sharp response, the FP finds widespread use as some form of narrow band filter, in telecommunications and in applications where we want to filter a narrow range of wavelengths, perhaps if we want to form an image in some narrow spectral line. An extension of these ideas leads to *antireflection coatings* and *interference filters*, created by multiple thin layers of dielectric material deposited onto a glass surface. They also find extensive use as resonant cavities in lasers, for "mode selection" of laser light, and in the detection of gravitational waves, discussed below (Sections 5.6, 6 and 7.2).

Because the etalon produces multiple internal reflections, it is crucial that the optical surfaces are flat. We want total wavefront errors to be much smaller than $\lambda/2$, but each surface has around F reflections, so we need the surfaces to be flat with surface errors much smaller than $\lambda/2F$.

Our treatment above ignores the shape of the wavefront, we have effectively assumed plane waves. But even lasers do not generate truly plane waves, and in laser cavities we can improve efficiency by adopting concave reflecting mirror surfaces.

5.6 Interferometric detection of gravitational waves

A gravitational wave incident on a Michelson interferometer causes one of the perpendicular arms to shorten and the other to lengthen with respect to their rest state, oscillating in time at a few hundred Hz. This changes the time taken for light to make the return journey down each arm, and hence there is a phase shift when the beams are recombined, which may be measured as a variation in intensity of the interferometer output. The relevant quantity we want to measure is the *strain*, which is defined as the fractional change in length of an arm, $s = \Delta L/L$ for arm length L, and the cosmic sources detected to date have $s \simeq 10^{-21}$. The recent detections have used measurements from the two interferometers of the LIGO collaboration[13] and the one interferometer of the Virgo collaboration[14]. A schematic diagram of the Advanced LIGO interferometer is shown in Fig. 39.

Suppose we can detect and measure accurately a shift that turns an interferometer maximum into a minimum, then the relative mirror shift would be $\Delta x = \lambda/8$, if both arms are affected equally in opposite directions, so for the LIGO laser's $\lambda = 1064\,\text{nm}$ and arm length 4 km, $s \simeq 3 \times 10^{-11}$, nowhere near accurate enough to detect gravitational waves.

There are two parts to the solution to this difficulty. First, arrange the interferometer so that the beams interfere destructively and produce a dark central fringe, then instead of trying to detect an entire fringe cycle, measure the intensity variation away from zero that is caused by a much smaller movement, some very small fraction of a fringe cycle.

Second, amplify the signal by converting each arm into a Fabry-Perot "resonant cavity", as shown in the figure. The finesse of each cavity is $F \simeq 300 - 500$: a hand-waving way to think about this is that the light makes the round trip along each arm effectively F times, so a mirror movement of x may be considered to cause a light path variation $\Delta L \simeq 2Fx$. Smaller intensity fluctuations may be detected if higher effective laser power is used (because the ratio of the signal to the noise is higher), so a key engineering feature is the "power recycling" which feeds the emerging laser power back into the Fabry-Perot resonant cavities and boosts the effective power as indicated on the figure.

To reduce noise and improve stability, the arms are in highly evacuated tubes, and the mirrors are suspended on complex suspension systems that both passively damp vibrations and which also have active "noise-cancelling", using feedback from the interferometer output to keep the Fabry-Perot arms in resonance in the presence of low-frequency seismic noise. The latest design also features "signal recycling". Reliable detections of gravitational waves are made by searching only for coincident oscillations between widely spaced multiple interferometers (so as to avoid coincident vibrational/seismic noise sources - the LIGO ones are 3000 km apart, in Washington state and Louisiana). Some directional information is also available from the timing of the signals and from the weak directional dependence of the interferometers. In the August 2017 detection and identification of merging neutron stars (https://www.ligo.caltech.edu/page/press-release-gw170817), the signal was weakest at the Italian Virgo site because the arms were oriented close to 45° to the gravitational wave direction, where the fringe variation is weakest, and for the first time the electromagnetic emission associated with a gravitational wave event was identified, using the correspondence of both the timing and the inferred direction in the sky, with a burst of gamma rays and other electromagnetic emission. As well as giving new insight into the physics of merging neutron stars and the nucleosynthesis of heavy elements, this result demonstrated that gravitational waves travel at the speed of light, to high accuracy. For those interested, more details are given in the publications cited in the figure caption, and on the collaboration web sites.

[13] https://www.ligo.org/
[14] http://www.virgo-gw.eu/

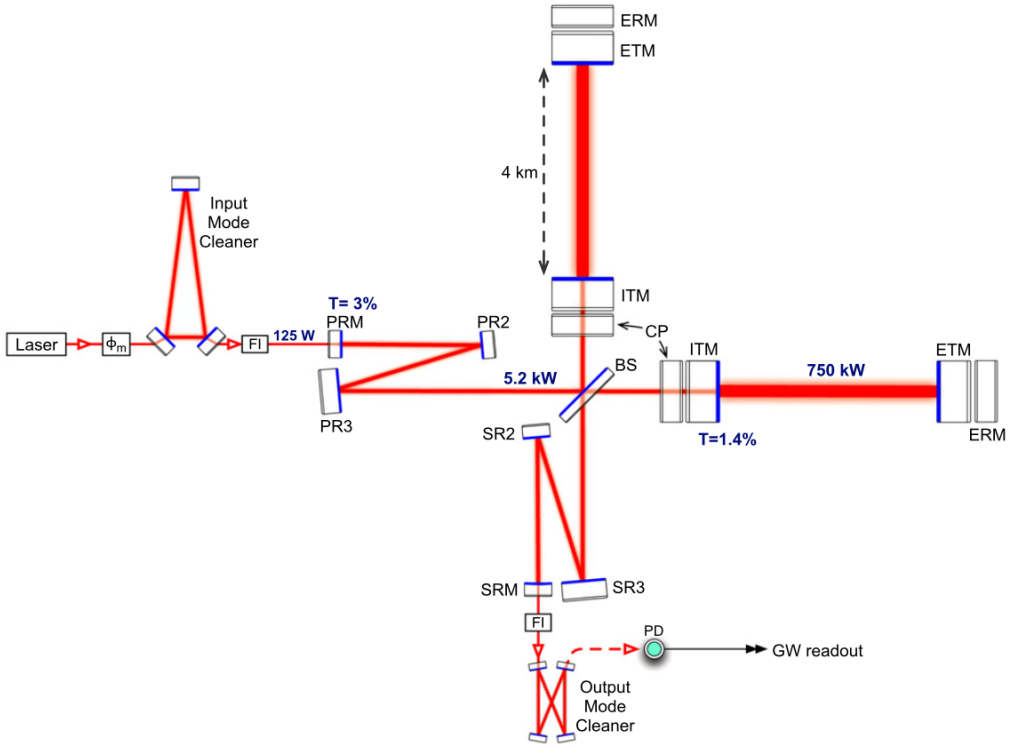


Figure 39: Schematic of the Advanced LIGO optical configuration. Key - ITM: input test mass; ETM: end test mass; ERM: end reaction mass; CP: compensation plate; PRM: power recycling mirror; PR2/PR3: power recycling mirror 2/3; BS: 50/50 beam splitter; SRM: signal recycling mirror; SR2/SR3: signal recycling mirror 2/3; FI: Faraday isolator; ϕ_m: phase modulator; PD: photodetector. The laser power numbers correspond to full-power operation. All of the components shown, except the laser and phase modulator, are mounted in the LIGO ultra-high vacuum system on seismically isolated platforms. Figure reproduced from "Advanced LIGO", The LIGO Scientific Collaboration et al. (2015), Class. Quantum Grav. 32, 074001 https://arxiv.org/abs/1411.4547. See also https://arxiv.org/abs/0711.3041

6 Basics of Lasers

The physics of lasers is covered more extensively in later years, B3 and C2, but here we give a brief introduction.

6.1 Interaction of atoms with radiation and the Einstein coefficients

It is observed that atoms both emit and absorb light at discrete wavelengths ("lines") that depend only on the atomic species being considered. Fig. 21 showed an astronomical spectrum displaying emission lines associated with a variety of atomic ions (and with the lines broadened by Doppler line broadening). It is now understood that these lines are associated with transitions of electrons between atomic energy levels, for emission lines at optical, ultraviolet or X-ray wavelengths[15]. In the presence of an electromagnetic wave, an atom may be excited to a higher energy level, in the process absorbing a photon of energy equal to the difference in energy between the atomic states: this process is known as *absorption*. An incident electromagnetic wave may also cause an excited atom to drop to a lower energy state, emitting a photon of the relevant energy - a process known as *stimulated emission*. An important point is that the electromagnetic wave associated with the stimulated emission is in phase with the incident wave. An excited atom may also spontaneously decay to a lower energy state and emit a photon in a process known as *spontaneous emission*. These possibilities are illustrated in Fig. 40.

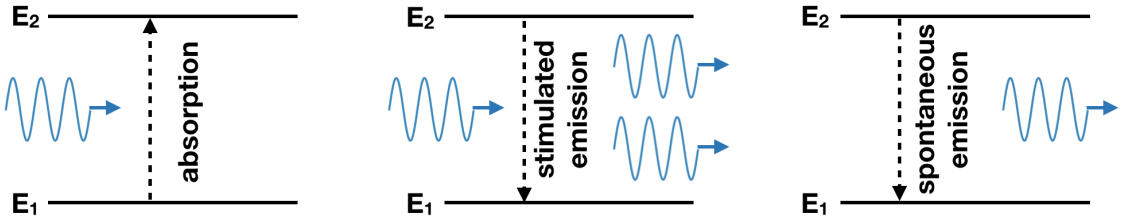


Figure 40: Schematic illustration of the interaction of radiation with two levels of an atom by absorption, stimulated emission or spontaneous emission of a photon whose energy matches the difference in the atomic energy levels.

A quantum mechanical treatment of these phenomena is beyond the scope of this optics course, but is covered in the third year of our course. A simple, semi-classical description is given in Chapter 14.4 of the course textbook, Lipson, Lipson & Lipson. A full treatment of spontaneous emission, in particular, requires Dirac's *quantum electrodynamics*. Note that radiative transition is not the only mechanism that may cause electronic transitions, *collisional* processes or the *Auger*[16] *effect* may also cause non-radiative transitions.

Einstein postulated that radiative transitions could be parameterised by two coefficients, A for spontaneous emission and B for stimulated emission or absorption, and that these coefficients were fundamental to the atoms, dependent only on the atomic structure and excitation, and not on the radiation field or the external environment. Consider a population of atoms in the presence of an isotropic radiation field, whose spectral energy density (energy per unit volume per unit angular frequency per steradian) is $u(\omega)$ for photons of angular frequency ω and photon energy $E = \hbar\omega = h\nu = hc/\lambda$ for frequency ν or wavelength λ. The atomic population has number densities n_1 and n_2 atoms per unit volume in each of two energy levels with energies E_1 and E_2, where $E_2 > E_1$, respectively. The rate per unit volume of radiative "downward" transitions from E_2 to E_1 is postulated to be given as a linear relation:

$$r_{21} = A_{21}n_2 + B_{21}u(\omega)n_2\,, \tag{108}$$

where the first term is the spontaneous emission rate and the second term is the stimulated emission rate. The dimensions of the terms in this equation are (number of transitions per unit time per unit volume). Each transition is for a photon of energy $E = \hbar\omega = E_2 - E_1$, so if we wish to quote the energy rate per unit volume, we simply multiply by $\hbar\omega$. For transitions from E_1 to E_2 there is no spontaneous term (we need an incident photon to be able to absorb it), so the "upward" transition rate is postulated to be

$$r_{12} = B_{12}u(\omega)n_1\,. \tag{109}$$

[15] Discrete emission and absorption lines also arise from transitions between rotational or vibrational excitations of molecules, typically at lower photon energies in the infrared, sub-mm or radio regimes.

[16] pronounced as a French name

In both cases we require the photon energy $E = \hbar\omega = E_2 - E_1$ (we shall discuss later the effects of broadening of the energy range over which transitions may occur). A has units of s^{-1} and B has units of $\mathrm{s}^{-1}(\mathrm{J\,s\,m^{-3}\,sr^{-1}})^{-1} \equiv \mathrm{J^{-1}m^3\,sr\,s^{-2}}$ (note the different definition of $u(\omega)$ and hence B compared with Lipson, Lipson & Lipson).

For atoms in thermal equilibrium at temperature T, the population number densities are related by the Boltzmann factor (see the Statistical Physics course):

$$\frac{n_2}{n_1} = \frac{g_2}{g_1}\exp\left[-\frac{E_2 - E_1}{k_\mathrm{B}T}\right], \tag{110}$$

where g_1, g_2 are the degeneracies of each level. For blackbody radiation, where matter and radiation are in thermal equilibrium,

$$u(\omega) = \frac{\hbar\omega^3}{\pi^2c^3}\frac{1}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1}. \tag{111}$$

In this two-level atom, there is only a single A coefficient, $A_{21} \equiv A$. The difference between the energies of the levels equals the photon energy, and in equilibrium we expect $r_{21} = r_{12}$, so

$$\begin{aligned} An_2 + B_{21}un_2 &= B_{12}un_1 \\ \left(A + \frac{\hbar\omega^3}{\pi^2c^3}\frac{B_{21}}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1}\right) g_2\exp\left[-\frac{\hbar\omega}{k_\mathrm{B}T}\right] &= \frac{\hbar\omega^3}{\pi^2c^3}\frac{g_1B_{12}}{\mathrm{e}^{\hbar\omega/k_\mathrm{B}T} - 1} \\ g_2A\left(1 - \exp\left[-\frac{\hbar\omega}{k_\mathrm{B}T}\right]\right) &= \frac{\hbar\omega^3}{\pi^2c^3}\left(g_1B_{12} - g_2B_{21}\exp\left[-\frac{\hbar\omega}{k_\mathrm{B}T}\right]\right). \end{aligned} \tag{112}$$

If A and B are indeed only dependent on the atomic structure and not on the temperature, then this relation between A and B should hold for any temperature, and hence we require

$$g_1B_{12} = g_2B_{21} \tag{113}$$

In our simple treatment, we shall ignore degeneracy, by setting $g_1 = g_2 = 1$, and hence $B_{12} = B_{21} = B$ and

$$\frac{A}{B} = \frac{\hbar\omega^3}{\pi^2c^3}. \tag{114}$$

Again, if A, B are dependent on the properties of the atom alone, then this relation must hold for any radiation field (not only blackbody) and regardless of whether the atoms are in thermal equilibrium, or not.

6.2 Conditions for laser gain and population inversion

We can now imagine a system in which radiation stimulates emission from atoms in some region, and thus we can *amplify* the initial radiation field by some factor, known as the *optical gain*. The electromagnetic waves associated with those stimulated photons are in phase with the initial radiation, so if we can ensure that the stimulated emission photons are produced at a rate faster than they are absorbed, we are able to make a *coherent* light source (see section 7). Because the stimulated emitted waves are all in phase, high light intensity values can be obtained.

The condition for successful maser or laser action is that the stimulated photons are produced more rapidly than they are absorbed, so we require

$$\begin{aligned} n_2B_{21}u(\omega) &> n_1B_{12}u(\omega) \\ n_2 &> n_1, \end{aligned} \tag{115}$$

known as the condition for optical gain, where we again ignore the possibility of degenerate energy levels. Thus we require the higher energy level to be more densely populated than the lower energy level, in contradiction to what would be expected for a population in thermal equilibrium (in equation 110, the exponent must be negative, as $E_2 > E_1$ and $T > 0$), and such a non-equilibrium state is known as *population inversion*.

For efficient generation of laser light, the rate of stimulated emission should dominate over spontaneous emission.This is achieved for $Bu(\omega) > A$, which can be achieved with sufficiently high $u(\omega)$. This is easier to attain at microwave energies, and the first *Maser* (Microwave Amplification by Stimulated Emission of Radiation) device was successful in 1953[17]. The first device at optical wavelengths, the *Laser* (Light Amplification...), was demonstrated in 1960.

[17] Naturally occurring astronomical masers are also found in giant molecular clouds in galaxies.

6.3 Pumping

Population inversion is achieved by *pumping*, in which the higher energy level is either temporarily or continuously populated by either radiative transitions from lower levels or from higher levels above E_2, or by non-radiative processes such as collisional excitation. Let us suppose that electrons have a typical lifetime τ_2 in level E_2, where the decay out of level 2 might be a radiative transition to E_1, a non-radiative transition or a transition to some other level. Similarly, electrons in level E_1 have a typical lifetime τ_1. Let us assume that these levels are populated at *pump rates* R_2, R_1, respectively. Then the time rate of change of the populations in each level are:

$$\begin{aligned} \frac{\mathrm{d}n_2}{\mathrm{d}t} &= R_2 - \frac{n_2}{\tau_2} \\ \frac{\mathrm{d}n_1}{\mathrm{d}t} &= R_1 + n_2 A - \frac{n_1}{\tau_1}\,, \end{aligned} \tag{116}$$

where, for our purposes here, the rate of population of E_1 due to spontaneous emission from E_2, given by the term $n_2 A$, has been separated from all other effects in the pump rate R_1. For steady-state population inversion we require these time derivatives to be zero, and hence

$$\begin{aligned} n_2 &= R_2\tau_2 \\ n_1 &= R_1\tau_1 + n_2 A\tau_1 = R_1\tau_1 + R_2\tau_2 A\tau_1\,. \end{aligned} \tag{117}$$

For optical gain, we require $n_2 > n_1$, so

$$\begin{aligned} R_2\tau_2 &> R_1\tau_1 + R_2\tau_2 A\tau_1 \\ R_2\tau_2\left[1 - A\tau_1\right] &> R_1\tau_1 \\ \frac{R_2\tau_2}{R_1\tau_1}\left[1 - A\tau_1\right] &> 1\,. \end{aligned} \tag{118}$$

Since $A\tau_1 > 0$, $[1 - A\tau_1] < 1$. To satisfy equation 118 we require $R_2\tau_2 > R_1\tau_1$ - implying we at least need either the pump rates to be favourable, $R_2 > R_1$, or the lifetimes to be favourable, $\tau_1 < \tau_2$. But $A\tau_1$ may be large, such that the square bracket is negative, in which case steady-state population inversion could never be achieved, and so we also require $A\tau_1 < 1$, as a necessary but not sufficient condition for stead-state laser action (transient laser action could be obtained without satisfying this condition).

6.4 Laser design

6.4.1 Laser oscillators

A *laser amplifier* may simply consist of a region containing population-inverted atoms, such that a light beam of the correct wavelength is amplified as it passes through. More usefully, however, the inverted atoms are stored inside a resonant cavity (like a Fabry-Perot cavity) to create a *laser oscillator*. The cavity can be tuned so that it traps radiation with angular frequency ω that matches the relevant atomic transition, greatly enhancing the production of a beam of coherent light at that frequency (see also Section 7.2).

6.4.2 Optical pumping

Many lasers achieve population inversion by optical pumping to a third, higher, energy level, which then decays (either radiatively or non-radiatively) to the upper level of the laser transition, as illustrated in Fig. 41. We can achieve population inversion between the upper and lower levels of the laser transition, E_1 and E_0, if the lifetime $\tau_2 < \tau_1$, so that electrons are depleted from the lowest level, E_0 by the pumping, then rapidly decay from E_2 to E_1, but then stay in the state E_1 sufficiently long to achieve the population inversion requirement, $n_1 > n_0$. The first ruby laser essentially used such a scheme. In practice, it can be difficult to maintain sufficient depletion of the lowest level, and in four-level laser systems, the lowest level of the laser transition is not the ground state of the atom.

6.4.3 Collisional pumping

The helium-neon laser comprises a mixture of He and Ne with about 10:1 ratio of partial pressures. The Ne is pumped using collisional excitation from collisions between He and Ne atoms. The carbon dioxide laser has a similar operational principle, using nitrogen as the excitation gas.

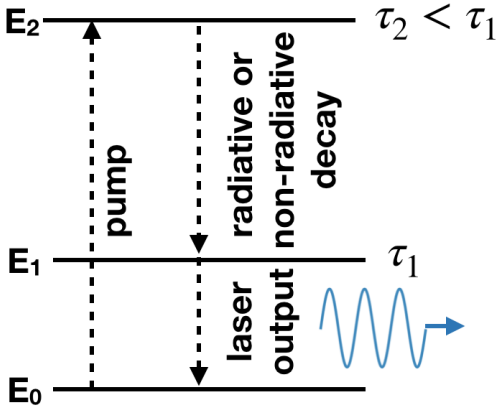


Figure 41: Schematic illustration of a three-level laser. Electrons are radiatively pumped to the highest level, E_2, where they rapidly decay to the intermediate level E_1, with lifetimes $\tau_2 < \tau_1$. Provided the lowest level E_0 is sufficiently depleted, population inversion is achieved and laser emission obtained between E_1 and E_0.

7 Coherence

7.1 Temporal coherence

So far we've only considered interference between rays or wavefronts assumed to have the same wavelength λ: the spectrum is a *delta-function* at some angular frequency ω. But the Fourier transform of a delta function is a sinusoidal wave of that frequency extending to time $t = \pm\infty$. A pure monochromatic wave cannot exist.

A light source with a finite frequency range $\Delta\nu$ can be considered as emitting a stream of waves, each of finite duration Δt, with each wave having a random phase with respect to the others. The finite duration of each wave produces the finite range of frequency in the source spectrum, $\Delta\nu \simeq 1/\Delta t$. The "random phase" element is crucial, and it leads to such a source being known as *incoherent*, as opposed to a light source such as a laser. The combined amplitude of these waves varies in a non-sinusoidal, random way, and as the waves combine the phase also jumps around randomly on some timescale which we now investigate.

Let's think about the time autocorrelation function, which we first met in Section 4.4.5. The autocorrelation function, $I(\tau)$, can be measured with the Michelson interferometer, when we realise that a mirror movement of x results in a path time delay $\tau = 2x/c$. *The Fourier transform of the time autocorrelation function is the spectral power density of the source.* Thus, from the usual result of a Fourier transform, if the intensity spectrum has width $\Delta\nu$, we expect the autocorrelation function to have width $\tau_c \simeq 1/\Delta\nu$. As a specific example, suppose the spectrum is a tophat function as indicated in Fig. 42: the autocorrelation function is a sinc function, with a maximum at time lag $\tau = 0$, which falls to zero at $\tau_c = 1/\Delta\nu$. The temporal "width", τ_c, of the autocorrelation function is inversely proportional to the frequency width of the spectrum. That autocorrelation width in time delay is known as the *temporal* (or *longitudinal*) *coherence time* and characterises the timescale on which the wave has a consistent phase.

Suppose we split a light beam from this source with a beamsplitter, introduce a time delay τ_D and then interfere the beams. If the time delay is longer than the coherence time, $\tau_D > \tau_c$, then interference effects are strongly suppressed: for a long time delay, the two beams are effectively incoherent and they only interfere coherently for a short time delay. This is illustrated in Fig. 43, where a piece of glass is introduced into one of the paths in a Young's slits experiment, thus introducing a time delay into one beam (right panels). When a laser is used as a light source (upper panels), its coherence time is long (section 7.2) and the fringe pattern is unaffected by the time delay. When an incoherent light source is used (lower panels), the time delay is longer than the coherence time and the fringes disappear.

The eye is sensitive to light over a wavelength range about $450 - 650\,\text{nm}$, corresponding to $\Delta\nu \simeq 2 \times 10^{14}\,\text{Hz}$, so the coherence time for white light is about $5 \times 10^{-15}\,\text{s}$. In that time, light travels a distance $\Delta x = c\tau_c \simeq 1\mu\text{m}$, which is the *coherence length* for white light. If we pass white light through a Michelson interferometer, or other interference experiment, the fringes disappear for any optical path difference greater than that length. We can use a Michelson interferometer to investigate the coherence time of light sources.

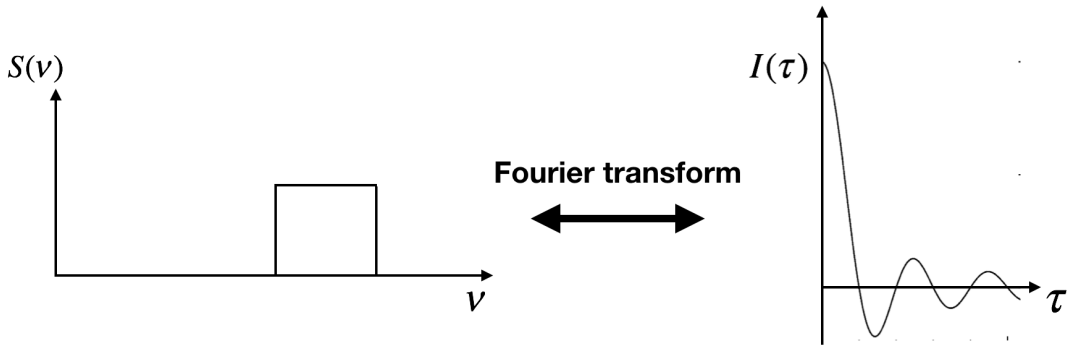


Figure 42: The Fourier transform of the spectrum is the temporal autocorrelation function of the light source.

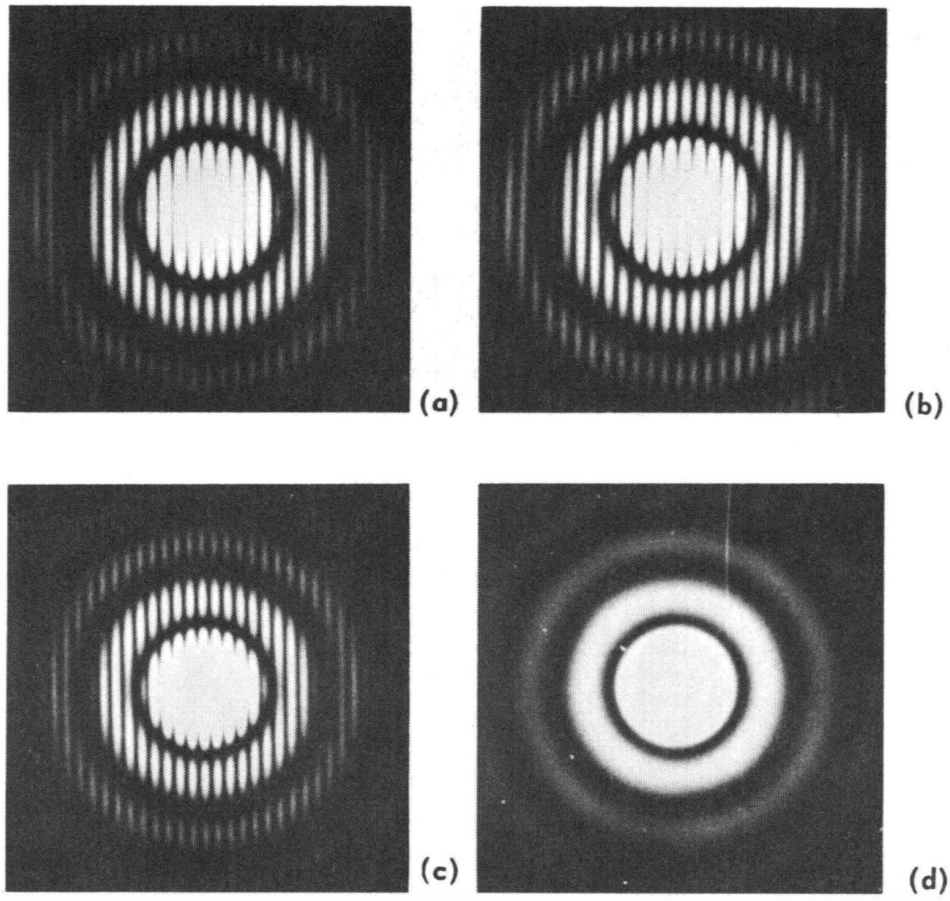


Figure 43: Fringes observed in double-beam interference using two circular apertures. The circular pattern arises from the diffraction by the individual circular apertures, the vertical fringes from the interference between the two beams. Top panels, illumination by He-Ne laser; lower panels, illumination by (incoherent) mercury arc. Right panels, with a glass plate placed over one aperture, 0.5 mm thick. The fringes disappear in (d), when the path difference exceeds the coherence length for the light source. Figure from Thompson, B.J., Optical Engineering, 4, 107 (1965), also discussed in Hecht.

7.2 The temporal coherence of lasers

We have already suggested that lasers produce light with a long coherence time. Inside the laser, photons are produced by stimulated emission such that their wavefronts are in phase (see Section 6). The timescale for the line emission is typically about $10^{-8} - 10^{-9}$ s (depending on the type of laser device), so the basic

atomic transition has a non-negligible frequency width. However, in a laser oscillator. the cavity is terminated by reflecting surfaces which act as a Fabry-Perot resonant cavity, and the sharp spectral response selects a much sharper range in frequency.

To select a single frequency, one method would be to ensure that the free spectral range is large enough that neighbouring orders lie outside the wavenumber range of the electronic transition being produced, as indicated in Fig. 44. Usually, however, the FSR is sufficiently small that there there are multiple modes within the natural linewidth of the atomic transition, but in this case the mode nearest the peak of the line is preferentially amplified by the laser oscillator.

In fact, even sharper linewidths can be obtained than would be expected from a simple FP, because there is a net optical gain in a round trip of light between the FP mirrors (i.e. as much light is gained by stimulated emission as is lost both by atomic transitions and by removal of light from the cavity). In that case the effective reflectance of the cavity $R \to 1$ and, referring to equations 103 & 104, we see that we expect the laser oscillator linewidth $\Delta\bar{\nu} \to 0$ and hence the coherence time $\tau_c \to \infty$. However, there is a theoretical upper limit on τ_c set by the occurrence of spontaneous emission, which has random phase and therefore degrades the coherence of the laser output, and in practice the most common limitation on the coherence time is set by acoustic noise and vibration, which perturbs the FP cavity away from its resonant condition.

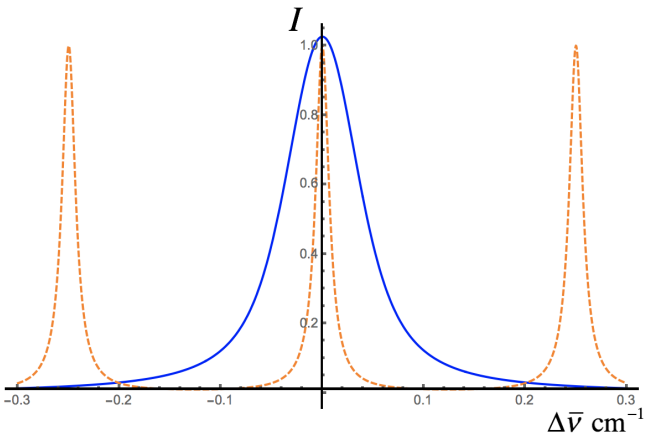


Figure 44: The wavenumber response of a Fabry-Perot resonant cavity of width 2 cm and $\Phi = 100$ at $\lambda = 600\,\text{nm}$ (dashed curve), superimposed on the profile of an atomic emission line (solid curve), showing how a FP cavity has a significantly narrower width than the natural emission line. In fact, laser oscillators achieve even sharper emission lines as explained in the text.

7.3 Interference fringe visibility and coherence

Michelson defined the *visibility* of an interference fringe pattern as

$$\mathcal{V} = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}\,, \tag{119}$$

where $I_{\text{max}}, I_{\text{min}}$ are the maximum and minimum intensity observed in the fringe pattern.

Referring back to equation 12, the interference term for the interference of two beams was written

$$\langle u_A u_B^\star + u_B u_A^\star \rangle = 2\text{Re}\left[u_A u_B^\star\right]. \tag{120}$$

The time-averaged cross-term $\Gamma = \langle u_A u_B^\star \rangle$ is known as the *mutual coherence function* and

$$\begin{aligned} I &= I_A + I_B + 2\text{Re}\left[\Gamma\right] \\ &= I_A + I_B + 2\sqrt{I_A I_B}\text{Re}\left[\gamma\right], \end{aligned} \tag{121}$$

where γ is known as the *degree of coherence*: note that γ is a complex number and thus may have a non-zero phase. If $\gamma = 0$, the beams are fully incoherent, $|\gamma| = 1$ corresponds to fully coherent interference. Values between these limits correspond to *partial coherence*.

If we consider the visibility of a set of fringes, then we expect $I_{\text{max}} = I_A + I_B + 2\sqrt{I_A I_B}|\gamma|$, $I_{\text{min}} = I_A + I_B - 2\sqrt{I_A I_B}|\gamma|$, so

$$\mathcal{V} = \frac{2\sqrt{I_A I_B}}{I_A + I_B}|\gamma| \tag{122}$$

and, if $I_A = I_B$,

$$\mathcal{V} = |\gamma| \,. \tag{123}$$

The modulus of γ gives information about the fringe visibility, the phase of γ gives information about the position of the fringes. We may therefore also define a *complex visibility* where again the complex part describes the placement of the fringe pattern with respect to the nominal phase centre of the interference pattern.

7.4 Transverse coherence

We may extend the concept of coherence to the visibility of fringes formed with extended sources. *Spatial* or *transverse incoherence* arises in general with extended light sources, where different parts of the source produce wavefronts that are uncorrelated with each other.

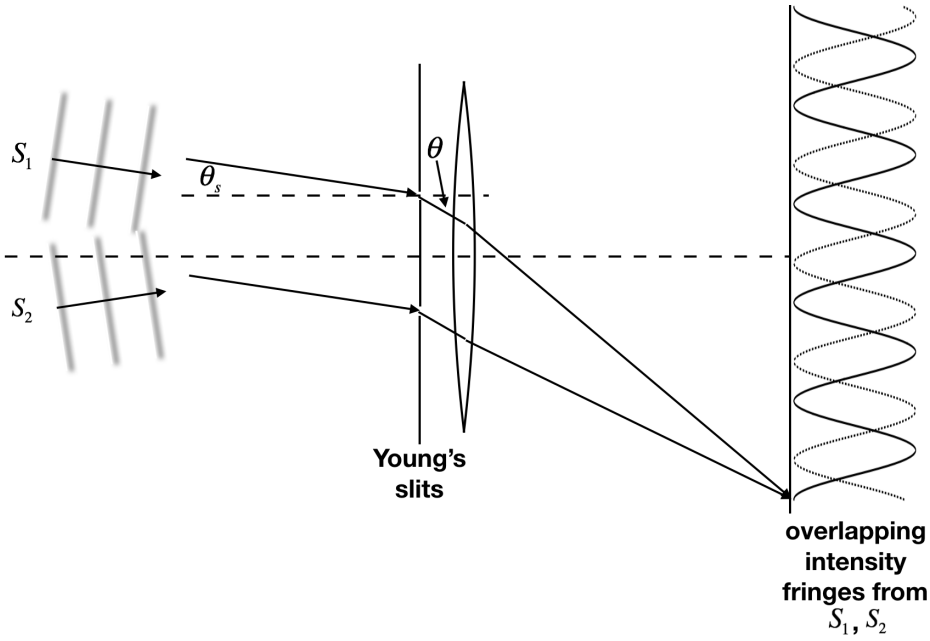


Figure 45: Illustration of the geometry of Young's fringes arising from an off-axis source S_1, with angle of incidence θ_s and angle to the interference pattern θ. When a second off-axis source S_2 is included, a pair of overlapping intensity fringes are created, offset from each other, as described in the text.

Consider two incoherent sources, equal distances either side of the axis of symmetry, which are illuminating a Young's slits experiment with plane wavefronts (Fig. 45). Each source produces a cosine-squared interference pattern in intensity, and because the sources are incoherent we add the intensities of the two patterns. For angle of incidence θ_i, and slit separation a, the path delay between the slits for a coherent wavefront from one of the sources is $\delta = ka(\sin\theta - \sin\theta_i)$, as we saw for the path delay for an inclined diffraction grating. Hence the sources at $\theta_i = \pm\theta_s$ on either side of the axis each produce an intensity pattern

$$\begin{aligned} I(\theta)_\pm &= I_0 \cos^2\left(\frac{ka(\sin\theta \mp \sin\theta_s}{2}\right) \\ &= \frac{I_0}{2}\left(1 + \cos\left(ka[\sin\theta \mp \sin\theta_s]\right)\right) \,. \end{aligned} \tag{124}$$

Adding the intensities of the two patterns,

$$I(\theta) = I_0(1 + \cos(ka\sin\theta)\cos(ka\sin\theta_s)) \,, \tag{125}$$

from which we can see that we maintain a cosine fringe pattern, but with reduced visibility. In a single-source Young's slit experiment (or when $\theta_s = 0$) maximum and minimum intensities occur at $\sin\theta = 0$ and $ka\sin\theta = \pi$: if we calculate the total intensity at these same angles in the two-source case, and insert into the (complex) visibility definition, we find

$$\mathcal{V} = \cos(ka\sin\theta_s) \,. \tag{126}$$

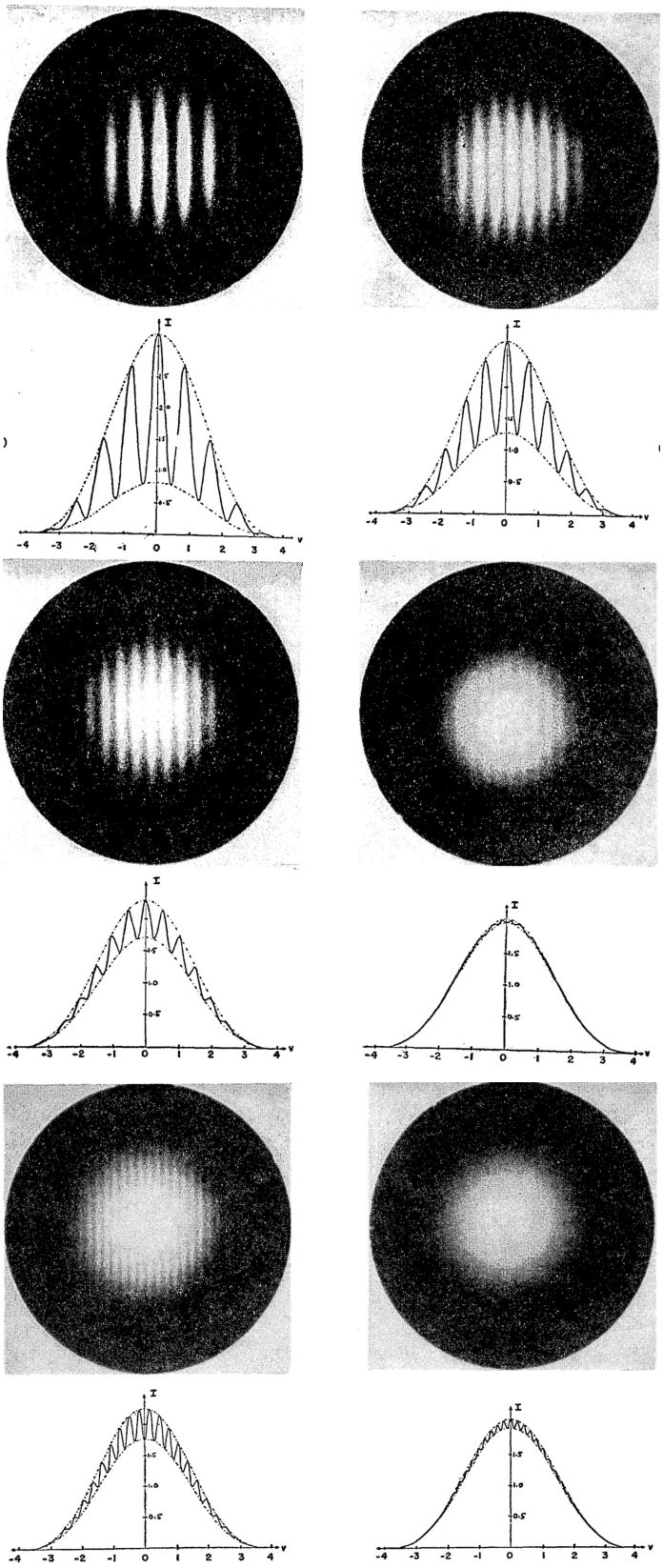


Figure 46: Double-beam interference pattern using an extended source, showing the variation in fringe visibility as the separation between the apertures is varied, showing data above and calculated fringe pattern below. From left to right and top to bottom, the aperture separation is $a = 0.6, 0.8, 1.0, 1.2, 1.7, 2.3\,\text{cm}$. Note how the visibility decreases to a minimum at $a = 1.2\,\text{cm}$ but increases again (with a change of sign) for $a = 1.7\,\text{cm}$. Figure from Thompson, B.J., J. Opt. Soc. of America, 47, 895 (1957), also discussed in Hecht.

Note that the interference pattern has not disappeared, even though we have illuminated the slits with two incoherent sources. This is because the electric fields at the two slits are still correlated with each other, but to a reduced extent. What matters for the visibility of the fringe pattern is the mutual coherence of the electric field at the two slits, and from equation 123 we can see that the fringe visibility on the screen and the degree of coherence at the slits are equivalent. If the two sources are coalesced into one, then the amplitudes at the slit become fully coherent ($\mathcal{V} = 1$) and for finite source separation, partial coherence in the range $0 < |\mathcal{V}| < 1$ is obtained. In this expression, $\mathcal{V}$ may be negative, which corresponds to exchange of the maxima and minima of the interference pattern.

We can extend this concept to the case of an extended source. Suppose the source intensity distribution may be described by a uniform (top-hat) function over the range $-\Theta_s < \theta_s < \Theta_s$. We can calculate the expected pattern by integrating equation 125 for a symmetric pair of sources, with respect to $k' = k \sin\theta_s$[18], over the range $0 \leq \theta_s \leq \Theta_s$, to find

$$\begin{aligned} I(\theta) &\propto \int_0^{k'=k\sin\Theta_s} (1 + \cos(ka\sin\theta)\cos(k'a))\mathrm{d}k' \\ &\propto k\sin\Theta_s + \frac{\cos(ka\sin\theta)\sin(ka\sin\Theta_s)}{a}\,, \end{aligned} \tag{127}$$

from which the visibility is

$$\mathcal{V}(a) = \mathrm{sinc}(ka\sin\Theta_s)\,. \tag{128}$$

This is illustrated in Fig. 46, which shows the variation in the fringes as the separation of the two apertures is varied in a double-beam (Young's slits) interference experiment, with an extended source. Note how the visibility decreases to a minimum, but then increases again. When it increases, there is a change of sign, corresponding to an exchange of the locations of the maxima and minima of the pattern.

Looking at the above equations, you might guess that in fact the (complex) visibility of the fringes as a function of the slit separation is given by the Fourier transform of the angular distribution of the source intensity as seen from the slits - the Fourier transform of a pair of point sources is a cosine function, the Fourier transform of a tophat is a sinc function. Note that we are considering the Fourier transform of the source *intensity*, not the amplitude - this is not simply a repeat of the mathematics of diffraction of coherent wavefronts. This Fourier transform relation for an extended, incoherent source is the *van Cittert-Zernike theorem*. Derivation of the van Cittert-Zernike theorem is not required on our syllabus.

7.5 The astronomical imaging interferometer

In modern astronomy, we make images of the sky across the electromagnetic spectrum, from radio to X-ray wavelengths. The resolution of a telescope of diameter D is $\Delta\theta \simeq 1.22\lambda/D$, so at radio (cm) wavelengths, the imaging angular resolution we can obtain from an individual telescope is seriously limited by the maximum size of parabolic reflector that we can build and operate. In the 1950s researchers in Australia and the UK built the first *imaging radio interferometer*s, in which separate radio dishes (antennas) are connected in pairs, and the visibility of the fringes produced by interfering the signals from each pair is measured. The concept of the "imaging interferometer" follows on from our discussion of transverse coherence in section 7.4. In a modern radio interferometer, with N dishes, we can construct $N(N-1)/2$ such pairs of baselines and thus measure the complex visibility $\mathcal{V}(a)$ as a function of dish separation a, due to sources of radio emission distributed across the sky. From the van Cittert-Zernike theorem, we can Fourier transform that distribution of visibilities to obtain an image of the sky. The smallest structures on the sky that can be measured (strictly speaking, the smallest scale Fourier modes) are given by $\theta \simeq \lambda/a$, for dish separation a. Since $a \gg D$, the imaging interferometer can return much high angular resolution images of the sky than can be obtained with a single radio dish. Fig. 47 shows the Jansky Very Large Array in the New Mexico desert, an array of 27 dishes (351 pairs) which can be configured to have spacings as large as $a = 35\,\mathrm{km}$, operating at wavelengths $0.6 < \lambda < 30\,\mathrm{cm}$, to obtain angular resolutions as small as 1.7×10^{-7} radian, or $0.04''$ ($1''$ is one arcsecond $= 1/3600$ degree). "Very Long Baselines" spanning the Earth have been connected to form a radio interferometer achieving milli-arcsecond (4×10^{-9} radian) resolution, and even at optical wavelengths it is now possible to measure complex visibilities from small numbers of closely-spaced optical telescopes connected together (see https://www.eso.org/sci/facilities/paranal/telescopes/vlti.html) to achieve comparable resolution (the challenge is to preserve the relative phase of the waves being combined from each telescope, which is difficult at optical wavelengths).

[18] This change of variables implies an assumption that the source intensity is uniform in $\sin\theta$ over the range $-\sin\Theta_s < \sin\theta_s < \sin\Theta_s$, which is not too important for small angles θ_s.

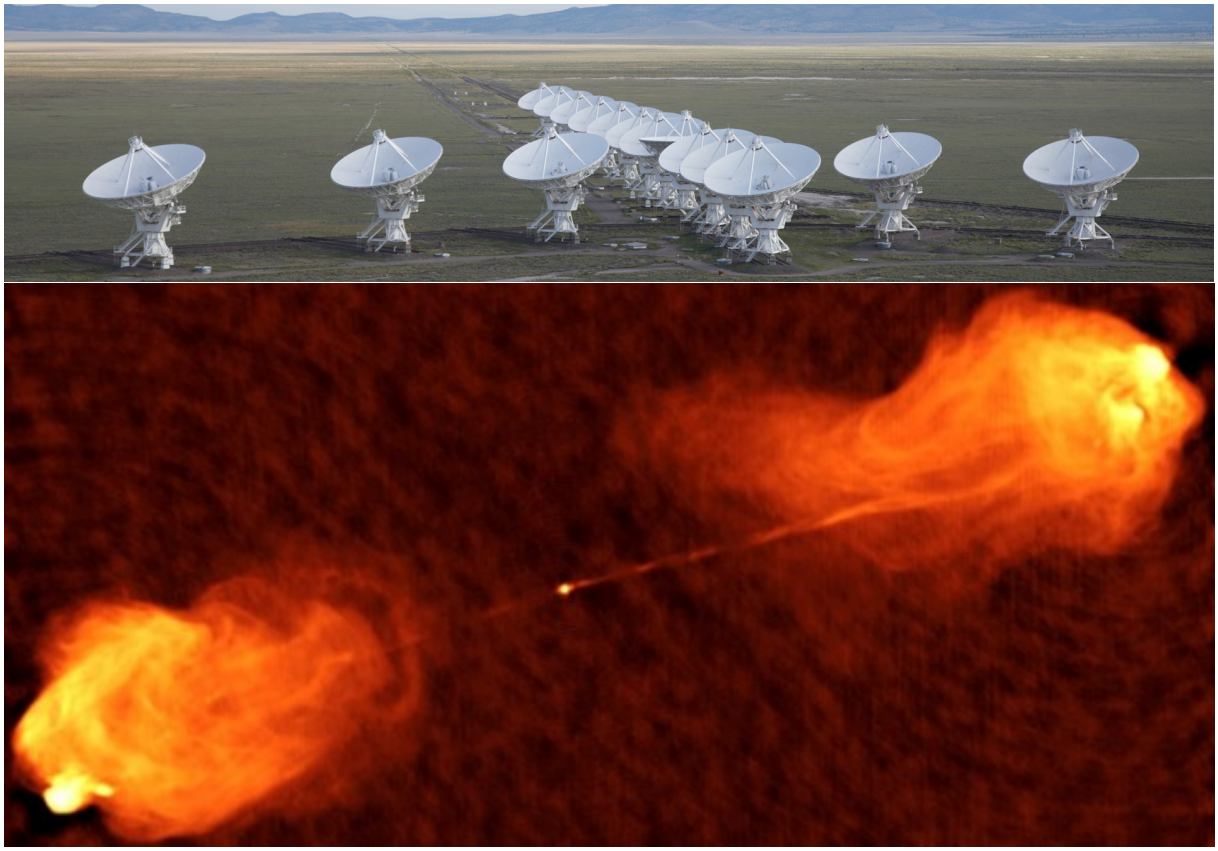


Figure 47: (top) The Jansky Very Large Array (JVLA) radio interferometer in a compact arrangement of its 27 dishes. (bottom) A JVLA image of the powerful radio source Cygnus A at 6 cm wavelength, showing radio emission from a region about 400,000 light years across, comprising radio-wavelength synchrotron emission by relativistic plasma being ejected from the region around a massive black hole located at the bright dot near the centre. See https://public.nrao.edu/telescopes/vla/.

8 Polarised light

8.1 Introduction to polarised light

8.1.1 Linear, circular and elliptical polarised light

So far in this course, we have not not used our knowledge from electromagnetism that light is a transverse wave in which the E and H fields are vectors at right angles to each other. Let us now consider the orientation, or distribution of orientations, of E or H in the plane of an electromagnetic wavefront, i.e. perpendicular to the direction of travel.

Consider first a plane electromagnetic wave travelling in the z direction with an E field that has an invariant orientation with time, e.g. in the x-direction defined by the unit vector $\hat{\mathbf{i}}$:

$$\mathbf{E}_x = \hat{\mathbf{i}} E_{0x} \cos(kz - \omega t)\,. \tag{129}$$

Such a wave is said to be *linearly polarised* or *plane-polarised*.

Now consider a beam of light comprising two such linearly-polarised waves which are coherent (i.e. have a time-invariant phase relationship) but whose E field directions are orthogonal to each other, in the x and y directions defined by unit vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$ respectively:

$$\begin{aligned} \mathbf{E}_x &= \hat{\mathbf{i}} E_{0x} \cos(kz - \omega t) \\ \mathbf{E}_y &= \hat{\mathbf{j}} E_{0y} \cos(kz - \omega t + \phi)\,, \end{aligned} \tag{130}$$

and where we have allowed there to be a phase difference ϕ between these two components and where the two components may have differing amplitudes. The resultant E-field is $\mathbf{E} = \mathbf{E}_x + \mathbf{E}_y$. If the phase difference $\phi = 0$, then we see that

$$\mathbf{E} = (\hat{\mathbf{i}} E_{0x} + \hat{\mathbf{j}} E_{0y}) \cos(kz - \omega t) \tag{131}$$

and hence the resultant is also a linearly polarised wave, with an E field that is orientated at an angle α in the plane of the wavefront, given by

$$\tan\alpha = \frac{E_{0y}}{E_{0x}}\,, \tag{132}$$

as illustrated in Fig. 48. As $(kz - \omega t)$ varies, the E field vector oscillates sinusoidally along that orientation. In fact, if the phase difference ϕ is any multiple of π, we obtain a linearly polarised resultant, with a flip

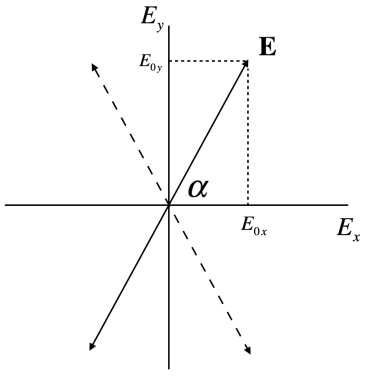


Figure 48: The coordinate system for linearly polarised light at angle α, with either zero phase difference, $\phi = 0$ (solid vector), or with phase difference $\phi = \pi$ (dashed vector).

in the sign of one component for odd multiples of π, so that $\alpha \to -\alpha$, as illustrated.

In a real physical problem we are unlikely to create two such coherent, orthogonal polarised waves. But this maths is showing how a given polarisation state may be understood mathematically as the superposition of two orthogonal states, and hence what types of polarised light may exist and how polarised light of one type may be manipulated into another type, as described below.

In general, the phase difference ϕ may not be zero or a multiple of π, in which case

$$\begin{aligned} \mathbf{E} &= \hat{\mathbf{i}} E_{0x} \cos(kz - \omega t) + \hat{\mathbf{j}} E_{0y} \cos(kz - \omega t + \phi) \\ &= \hat{\mathbf{i}} E_{0x} \cos(kz - \omega t) + \hat{\mathbf{j}} E_{0y} \cos(kz - \omega t)\cos\phi - \hat{\mathbf{j}} E_{0y} \sin(kz - \omega t)\sin\phi\,. \end{aligned} \tag{133}$$

and the E vector traces out a closed path on the xy plane. To see what that path is, eliminate the $(kz - \omega t)$ terms:

$$\begin{aligned}
\cos(kz-\omega t) &= \frac{E_x}{E_{0x}} \\
\sin(kz-\omega t) &= \left(1-\left(\frac{E_x}{E_{0x}}\right)^2\right)^{\frac{1}{2}} \\
\frac{E_y}{E_{0y}} &= \frac{E_x}{E_{0x}}\cos\phi - \left(1-\left(\frac{E_x}{E_{0x}}\right)^2\right)^{\frac{1}{2}}\sin\phi \\
\left(1-\left(\frac{E_x}{E_{0x}}\right)^2\right)\sin^2\phi &= \left(\frac{E_x}{E_{0x}}\right)^2\cos^2\phi + \left(\frac{E_y}{E_{0y}}\right)^2 - 2\frac{E_xE_y}{E_{0x}E_{0y}}\cos\phi \\
\sin^2\phi &= \left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - 2\frac{E_x}{E_{0x}}\frac{E_y}{E_{0y}}\cos\phi\,,
\end{aligned} \tag{134}$$

where we have used $\cos^2\phi = 1 - \sin^2\phi$ in the last step. If $\sin\phi = 0$ then $\cos\phi = \pm 1$ and

$$\frac{E_x}{E_{0x}} = \pm\frac{E_y}{E_{0y}}\,, \tag{135}$$

so the E vector traces out a linear variation on the xy plane: the wave is linearly polarised, as already discussed. If $\sin^2\phi = 1$, then $\cos\phi = 0$ and

$$1 = \left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2\,, \tag{136}$$

which is the equation of an ellipse aligned with the xy coordinate axes: the light is *elliptically polarised*. If we also have $E_{0x} = E_{0y}$ the locus is a circle on the xy plane and we say the light is *circularly polarised*. As $(kz - \omega t)$ varies, then at fixed z, the E field rotates around that circle with angular velocity ω. If the rotation is clockwise, **looking backwards towards the source**, the polarisation is *right-circularly polarised*, and if anti-clockwise looking backwards towards the source, it is *left-circularly polarised* (this convention is for historical reasons).

If ϕ has some arbitrary value, the locus is again an ellipse, but now not aligned with the xy coordinates: equation 134 has the form of an ellipse rotated with respect to the coordinate system. To see this, consider an elliptically-polarised wave as in equation 136, in a rotated coordinate system where the E fields are $E_{x'}$,$E_{y'}$:

$$1 = \left(\frac{E_{x'}}{E_{0x'}}\right)^2 + \left(\frac{E_{y'}}{E_{0y'}}\right)^2\,, \tag{137}$$

If the coordinate system is rotated by angle α, the fields are related by a standard rotational transformation:

$$\begin{aligned}
E_{x'} &= E_x\cos\alpha + E_y\sin\alpha \\
E_{y'} &= -E_x\sin\alpha + E_y\cos\alpha\,.
\end{aligned} \tag{138}$$

Substituting in equation 137,

$$1 = E_x^2\left(\frac{\cos^2\alpha}{E_{0x'}^2} + \frac{\sin^2\alpha}{E_{0y'}^2}\right) + E_y^2\left(\frac{\cos^2\alpha}{E_{0y'}^2} + \frac{\sin^2\alpha}{E_{0x'}^2}\right) - E_xE_y\sin 2\alpha\left(\frac{1}{E_{0y'}^2} - \frac{1}{E_{0x'}^2}\right)\,, \tag{139}$$

which has the same dependence on E_x, E_y as equation 134, for $\sin\phi \neq 0$, if we equate terms:

$$\begin{aligned}
\frac{1}{E_{ox}^2\sin^2\phi} &= \left(\frac{\cos^2\alpha}{E_{0x'}^2} + \frac{\sin^2\alpha}{E_{0y'}^2}\right) \\
\frac{1}{E_{oy}^2\sin^2\phi} &= \left(\frac{\cos^2\alpha}{E_{0y'}^2} + \frac{\sin^2\alpha}{E_{0x'}^2}\right) \\
\frac{2\cos\phi}{E_{0x}E_{0y}\sin^2\phi} &= \sin 2\alpha\left(\frac{1}{E_{0y'}^2} - \frac{1}{E_{0x'}^2}\right)\,.
\end{aligned} \tag{140}$$

Subtracting the first two equations and noting that $\cos 2\alpha = \cos^2\alpha - \sin^2\alpha$, we find

$$\frac{1}{\sin^2\phi}\left(\frac{1}{E_{0x}^2} - \frac{1}{E_{0y}^2}\right) = \cos 2\alpha \left(\frac{1}{E_{0x'}^2} - \frac{1}{E_{0y'}^2}\right) \tag{141}$$

and dividing into the third equation we obtain

$$\tan 2\alpha = \frac{2E_{0x}E_{0y}\cos\phi}{(E_{0x}^2 - E_{0y}^2)}\,, \tag{142}$$

thus yielding the rotation angle of the elliptical polarisation in our coordinate system. As $\cos\phi \to 1$, the expression for $\tan 2\alpha$ becomes consistent with equation 132, but the axis ratio of the ellipse tends to zero and the polarisation becomes linear.

Thus we see that the most general form of polarised light is elliptical polarisation, with linear polarisation and circular polarisation being special cases of elliptical polarisation. All forms may be obtained mathematically from the combination of two coherent, orthogonal linear polarised waves, given some choice of amplitudes of those waves, E_{0x}, E_{0y} and their relative phase difference, ϕ.

8.1.2 The Jones vector

We may describe a polarised wave by the electric field amplitude, including its phase, on two orthogonal axes that are perpendicular to the direction of wave motion. Denoting these components as E_x, E_y, we may write

$$\mathbf{E} = \begin{pmatrix} E_x \\ E_y \end{pmatrix}\,. \tag{143}$$

We may factor out E_x:

$$\mathbf{E} = E_x \begin{pmatrix} 1 \\ (|E_y|/|E_x|)\mathrm{e}^{\mathrm{i}\phi} \end{pmatrix}\,, \tag{144}$$

where ϕ is the phase difference between E_y and E_x, and if we normalise this vector in the complex plane to unit length, we obtain the *Jones vector*, $\mathbf{E_J}$. For example, linearly-polarised light has $\phi = 0$ and its Jones vector is

$$\mathbf{E_J} = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix} \tag{145}$$

for plane of polarisation angle α. Left or right circularly polarised light has Jones vector

$$\mathbf{E_J} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ \pm\mathrm{i} \end{pmatrix}\,, \tag{146}$$

since $\phi = \pi/2$, where the sign of the $y-$component is positive for left circular polarisation. Jones vectors are not explicitly required on our syllabus, but they are a convenient way to denote polarisation states. Note that Jones vectors cannot be used to describe unpolarised or partially polarised light (see next section): for that we need to introduce *Stokes parameters*, but these are not on our syllabus.

8.1.3 Unpolarised and natural light

In nature, most light sources display either absent or only weak polarisation effects. *Unpolarised light* might be better described as "randomly polarised": it may be considered to comprise a sequence of individual waves whose polarisations are randomly oriented. Mathematically, the two orthogonal states, that formed the basis set in the above analysis, are incoherent, so that no consistent polarisation orientation is evident.

Commonly, light is in a state of *partial polarisation*, which may be regarded as a mixture of unpolarised light with some fraction of polarised light. The *degree of polarisation* tells us what the fraction of polarised light is.

Partially polarised light is produced naturally by reflections from surfaces (see next section) and by scattering. The sky appears blue because of *Rayleigh scattering* in the atmosphere, which is strongly wavelength dependent, and scattering preferentially scatters light whose E field is perpendicular to both the incident direction and the scattered light direction. The scattered light from the daytime sky is most highly polarised at an angle 90° away from the direction to the Sun.

8.2 Dichroism

Dichroism is the selective absorption of one polarisation state in a beam of light. The effect may be used to polarise light.

An aligned grid of fine, closely spaced conducting wires can act as a dichroic polariser. An oscillating E-field parallel to the wires drives motion of electrons and leads both to enhanced reflection of that polarisation and to dissipation of wave energy. The E-field that is perpendicular to the wires is more favourably transmitted. A more effective, manufactured dichroic polariser is polaroid, which is composed of polymers that are stretched in one orientation.

Some natural crystals such as tourmaline also have wavelength-dependent dichroic properties. For more information, read the relevant section in Hecht, not required on our syllabus.

Some people are able to detect polarisation of light by eye. Laptop screens generally produce polarised light. Display a blank, white screen and, away from other light sources, look at the screen while you rotate it clockwise and anticlockwise. You may be able to see a faint, coloured pattern known as "Haidinger's brush" rotating with the screen, as illustrated in Fig. 49, from the Wikipedia page on this phenomenon. It is thought to arise from (wavelength-dependent) dichroic absorption in the retina (not on our course syllabus).

Figure 49: Artist's impression of "Haidinger's brush" for vertically polarised light, comprising a yellow horizontal feature and a blue vertical feature. By Dpbsmith at the English language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=693281

8.3 Birefringence

Optical anisotropy is the property that the refractive index in a crystal may depend on the direction of propagation of light with respect to the ordered structure of the crystal. In *birefringence*, there are two differing orientation-dependent refractive indices (or velocities of light), which can lead to two separate refracted images being formed when an object is viewed through the crystal. The effect was first noted in 1669 by Bartholinus, when he was studying a transparent variant of calcite (which is a form of calcium carbonate) known as "Iceland spar". Information on the relevant crystal structures is given by Hecht, "Optics", details are not required on our syllabus.

In *uniaxial* birefringent crystals, electrons may oscillate more (or less) easily along one direction, known as the *optic axis*, than in a plane perpendicular to that axis, such that the speed of light in the crystal, and hence the refractive index, has one value when the **E** field is parallel to the optic axis, and a different value when the **E** field lies in the plane perpendicular to the optic axis. Note that it is the orientation of **E** with respect to the optic axis that matters, and not directly the direction of ray propagation. In calcite, the speed for **E** parallel to the optic axis is the faster, so that axis is known as the *fast axis*, the refractive index is the smaller of the two values, and calcite is known as a *negative uniaxial* crystal. In quartz, the refractive index for **E** parallel to the optic axis is the larger, that axis is known as the *slow axis* and the crystal is known as *positive uniaxial*. Other crystals, e.g. mica, may have more than one optic axis, in differing orientations. We will only consider uniaxial crystals.

For some orientations of the crystal, rays with differing orientations of **E** are refracted differently, thus forming a double image when looking through it. When the two refracted rays are inspected, one appears to have the properties of normal refraction, in particular if the crystal is rotated about the normal to its surface, the position of the refracted image does not change. That ray is known as the *ordinary ray* or *o*-ray. Refraction of the other image depends on the orientation of the crystal and is known as the *extraordinary ray*, or *e*-ray.

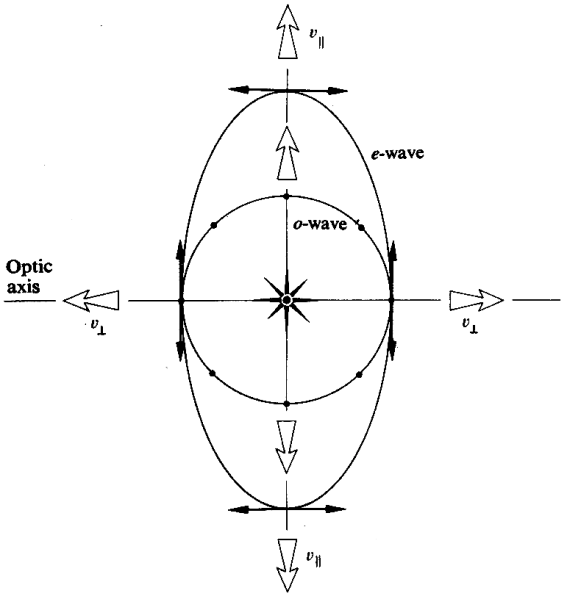


Figure 50: The propagation of Huygens wavelets in a negative uniaxial crystal such as calcite. The optic axis is depicted in the plane of the page. An *o*-ray has polarisation such that **E** (shown here normal to the page) is perpendicular to the optic axis everywhere around the wavefront and the wavefront propagates isotropically in the plane of the page. An *e*-ray has polarisation such that **E** is perpendicular to the optic axis for some portion of the wavefront, but parallel to the optic axis for another portion, and hence the propagation is anisotropic. Figure from Hecht, "Optics".

o-rays have Huygens wavefronts with the **E** vector everywhere perpendicular to the optic axis, so that they experience the same refractive index irrespective of direction, and they propagate isotropically through the crystal, and refract "ordinarily". *E*-rays have some component of their **E** vector along the optic axis and some component perpendicular to it, so that an *e*-ray wavefront propagates faster when **E** is parallel to the optic axis than it does when **E** is perpendicular to it (for a negative uniaxial crystal). If we think of Huygens wavefronts propagating through a crystal, the *o*-ray wavefronts are spherical in shape, but the *e*-ray wavefronts are ellipsoidal, as illustrated in Fig. 50.

The wavefront shapes and the relationship of the vectors is shown in Fig. 51. Note in particular that, for an *e*-ray, the ray direction defined by the Poynting vector is at some angle with respect to the normal to the wavefront, which is why the refracted images appear separate when the rays emerge, for suitable orientations of the crystal. More diagrams and explanation are available in Hecht, "Optics".

8.4 Waveplates and analysis of polarised light

8.4.1 Birefringent waveplates

We have seen how polarised light may be expressed as the sum of two orthogonal, linearly polarised components and how the various manifestations of polarised light (linear, circular, elliptical) depend only on the relative amplitudes of those components and the phase shift between them. We can use birefringent materials to create *waveplates* that change the phase of the two components and hence are able to change the polarisation. The idea is to arrange the two orthogonal, linearly-polarised components to experience differing refractive indices, so that they experience a relative phase shift on passage through the crystal, and the light beam emerges with an altered polarisation state. However, we do not want the two components to become spatially separated during this process. This can be achieved by cutting the crystal so that its optic axis lies in the plane of the surface of the crystal (front and rear surfaces should be parallel). If plane-polarised light is incident normal to the surface, both components travel through the crystal undeflected. But the component with **E** parallel to the optic axis experiences a refractive index n_e and the component with **E** orthogonal to the optic axis experiences n_o (we are free to rotate our mathematical coordinate system to align with the optic axis). Hence for crystal thickness d, there is a phase shift

$$\delta = kd(n_o - n_e)\,. \tag{147}$$

The *relative retardation* is the optical path length difference, usually expressed in wavelengths.

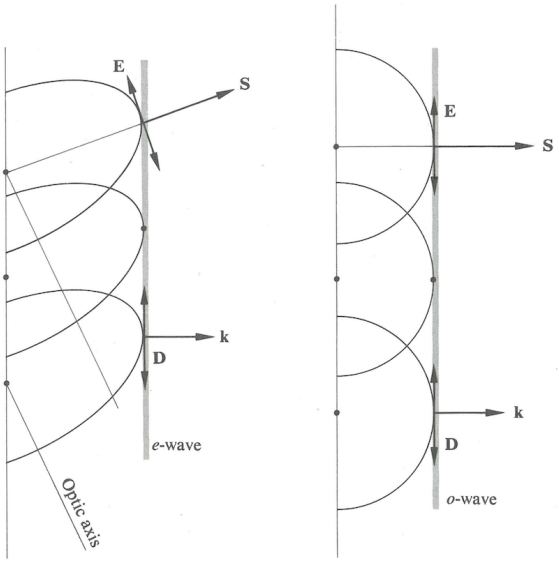


Figure 51: The relationship between the **E** and **D** fields, the wavevector **k** and the Poynting vector **S** in a birefringent material. In the left panel, the optic axis lies in the plane of the page, as indicated, and the *e*-ray wavefronts propagate at differing speeds when **E** is parallel to or perpendicular to that optic axis: the energy flow is in the direction of **S**, which is perpendicular to **E**, but the normal to the wavefronts is in the direction of **k**, which is perpendicular to **D**. In the right panel, the optic axis is perpendicular to the plane of the page, the *o*-ray wavefronts propagate isotropically in that plane, and **E** and **D** are parallel, as are **k** and **S**. Figure from Hecht, "Optics".

Note that unpolarised light is unaffected by a waveplate - the orthogonal components have random phase with respect to each other, so modifying the relative phase still leaves them uncorrelated.

8.4.2 Jones matrices

We may define the *Jones matrix*, **J**, as a 2×2 matrix that operates on a Jones vector, converting the polarisation state of light from an input state to an output state. A linear polariser such as polaroid has Jones matrix

$$\mathbf{J} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{148}$$

if oriented horizontally, and

$$\mathbf{J} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{149}$$

if oriented at 45°. Other examples of Jones matrices for birefringent waveplates are given below. A feature of a valid Jones matrix is that has one eigenvalue that is zero. Knowledge of Jones matrices is not required on our syllabus, and we will not discuss the mathematics of these any further, but they can be a convenient way to describe the effect of a birefringent waveplate, and the Jones matrix of a combination of optical elements may be obtained by multiplication of their individual Jones matrices.

8.4.3 The full-wave plate

If we choose d such that $\delta = 2\pi$, the relative retardation is one wavelength, and there is no net effect on the polarisation of the output beam at the stated wavelength. Its Jones matrix is the 2×2 identity matrix at that wavelength. However, wavelengths other than the one for which $\delta = 2\pi$ experience a phase shift and hence a polarisation modification.

8.4.4 The half-wave plate

If a waveplate has $\delta = \pi$ at some wavelength, and we input linearly polarised light with polarisation orientation at angle α to the optic axis, the output remains linearly polarised but rotated to angle $-\alpha$ (i.e. a change in angle of 2α). Any elliptically-polarised light is similarly flipped in angle, and also the sense of rotation is flipped, so that right-polarised (circular or elliptical) light becomes left-polarised and vice versa. Note that we are unable to distinguish between linearly polarised light with orientation α or

orientation $\alpha + \pi$. Linearly polarised light along or perpendicular to the optic axis appears unchanged. One possible Jones matrix is

$$\mathbf{J} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \tag{150}$$

8.4.5 The quarter-wave plate

If $\delta = \pi/2$, the waveplate converts linearly polarised light to elliptical and vice versa, depending on the orientation of the polarisation to the optic axis. Light linearly polarised along or perpendicular to the optic axis appears unchanged. If linear light has orientation $\alpha = \pi/4$ (45°), the orthogonal components have equal amplitude and the phase shift δ converts it to circularly polarised light. Similarly, circularly polarised input light is converted to linearly polarised light at $\alpha = 45°$. One possible Jones matrix is

$$\mathbf{J} = \begin{pmatrix} 1 & 0 \\ 0 & \pm\mathrm{i} \end{pmatrix} . \tag{151}$$

8.4.6 The circular polariser

From the above, we can create circularly polarised light using a combination of a linear polariser such as polaroid followed by a quarter-wave plate with optic axis at 45° to the polaroid. The Jones matrix of this circular polariser is given by product of the Jones matrices of its individual components,

$$\mathbf{J} = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & \pm\mathrm{i} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} . \tag{152}$$

Note that such a circular polariser is *inhomogeneous*, which effectively means that it has the desired effect only for light that propagates first through the polaroid and then through the quarter-wave plate. We would obtain linear polarisation, not circular polarisation, for light propagating in the opposite direction (if we swap the multiplication order above, we get a different total Jones matrix). For completeness, but not required on our syllabus, a *homogeneous* circular polariser has Jones matrix

$$\mathbf{J} = \frac{1}{2} \begin{pmatrix} 1 & \mp\mathrm{i} \\ \pm\mathrm{i} & 1 \end{pmatrix} \tag{153}$$

(try multiplying various input Jones vectors by the above, to see the effect).

8.4.7 Achromatic waveplates

The retardation phase shift δ of equation 147 depends on wavelength, both through the k factor and the wavelength-dependence of the refractive indices. Hence a simple waveplate only provides a given retardation δ value (e.g. a quarter wave) at a finite set of wavelengths (note that multiple factors of 2π do not change the effect on the polarised light, so more than one wavelength can satisfy the condition for a given retardance). The problem is exacerbated because the crystal materials have to be many wavelengths thick, in order to be able to be manufactured. The phase dependence on wavelength increases with the total thickness (see this by differentiating equation 147 with respect to wavelength, as a function of total thickness).

The simplest way of reducing the chromaticity is to construct a waveplate from two layers of the same material, with their optic axes at 90°, so that the net retardance from each layer cancels, apart from the minimum amount required to obtain the target retardance. This reduces but does not eliminate the wavelength dependence of δ.

Better cancellation can be obtained by exploiting the wavelength dependence of the refractive index of different materials, e.g. quartz and magnesium fluoride, making the two layers out of different material, with carefully chosen thicknesses, so that the phase δ is constructed to be approximately invariant with wavelength over some useful range.

8.4.8 Analysis of polarised light

We may also use the above to deduce the state of polarisation of a beam of light, by seeing how the intensity of a beam of light varies as we insert various retarding and polarising elements into the beam.

Suppose we pass linearly polarised light through a linear polariser. If the polarisation angle is aligned with the polariser, the light is fully transmitted. If they are aligned at 90° with respect to each other, no light is transmitted. At intermediate angles, θ, the *amplitude* of the transmitted light varies as $u = u_0 \cos\theta$, and hence the intensity of transmitted light varies as $I = I_0 \cos^2\theta$. This is known as *Malus's*

law. If we rotate the polariser and measure the dependence on θ, we can determine both whether the incoming light is consistent with being linearly polarised, or not, and if so, what the position angle is. If the light instead is elliptically polarised, the transmitted intensity would not fall to zero at $\theta = \pi/2$, and the relative intensities at maximum and minimum may be used to infer the ellipticity of the polarisation. Thus we could adopt the following steps to establish the polarisation of a polarised beam:

1. Pass the light through a linear polariser. Rotate the polariser to determine approximately the orientation of the major/minor axes of the polarisation ellipse – this is the angle at which a maximum and minimum transmission is obtained.

2. Set the linear polariser for maximum transmission. Remember that if the coordinate axes are chosen to coincide with the principal axes of the ellipse there is then a phase difference of $\delta = \pm\pi/2$ between the components. Insert a quarter-wave plate before the linear polariser (the axes of the quarter-wave plate are known.) Align the axis of the quarter-wave plate with the approximate ellipse axis. If it is exactly along the axis, then linearly polarised light results.

3. Rotate the linear polariser to check for complete extinction, and if necessary, iterate the orientations of quarter-wave plate and the linear polariser to obtain total extinction. The angle for total extinction is perpendicular to the orientation of the now linearly polarised E-vector. If the angle between this vector and the axis of the quarter-wave plate is θ, the ratio of the E-vector components is $\tan\theta$.

If instead, the beam is unpolarised, the measured intensity should be invariant with any rotations of polarisers and waveplate. If the beam of light is partially polarised, being the sum of some unpolarised light and some polarised light, in the last step, complete extinction would not be attainable.

8.4.9 Study of crystals in polarised light

Waveplates are commonly used to image crystal structure in geology. For example, suppose we make an analyser comprising in order:

1. linear polariser;

2. thin, transparent sample of crystalline rock;

3. optional waveplate;

4. linear polariser at 90° with respect to the first polariser at (1).

At wavelengths where the total retardance through waveplate and rock is a multiple of 2π, the net transmitted intensity is zero. At other wavelengths, light will be transmitted, so coloured images of the crystalline structure may be viewed, which depend on the orientation of the waveplate optic axis and the optic axis in the rock.

8.5 Induced birefringence

Some materials exhibit birefringence properties when induced by some external influence. *Photoelasticity* is the effect of birefringence following mechanical stretching of a material. In the *Faraday effect*, the orientation of linearly polarised light may be rotated when a strong magnetic field is applied.

The *Kerr effect* is birefringence induced by an electric field, which varies as the square of the applied field. A Kerr cell may be used as a very fast switch, by arranging a cell to have crossed linear polarisers on either side of a variable-delay waveplate whose retardance can be controlled by an electric field. For zero electric field, no light is transmitted, but when a field is applied, the resulting retardance rotates the polarisation and allows transmission. This may be operated at high switching frequencies. The *Pockels effect* occurs in some crystals, in which birefringence is observed that depends linearly on an applied electric field. Pockels cells may also be used as fast switches.

Study of these devices is beyond our course syllabus.

Appendices

This appendix discusses useful additional material that has previously been in the A2 optics syllabus, but parts of which are now instead included in the A2 electromagnetism syllabus.

A The reflection of polarised light

A.1 Terminology

When a light wave is incident on a plane surface, the *plane of incidence* is the plane which is perpendicular to the surface and which contains the direction of propagation of the light.

We may (instantaneously) resolve the electric field in the wave into two orthogonal polarisation components. The component where the electric field is parallel to the plane of incidence is known as the *p-mode* (for "parallel"), also known as the *TM* (transverse magnetic) mode. The component where the electric field is perpendicular to the plane of incidence is known as the *s-mode* (from the german "senkrecht", meaning perpendicular), also known as the *TE* (transverse electric) mode. The fields associated with these modes are shown diagrammatically in Fig. 52.

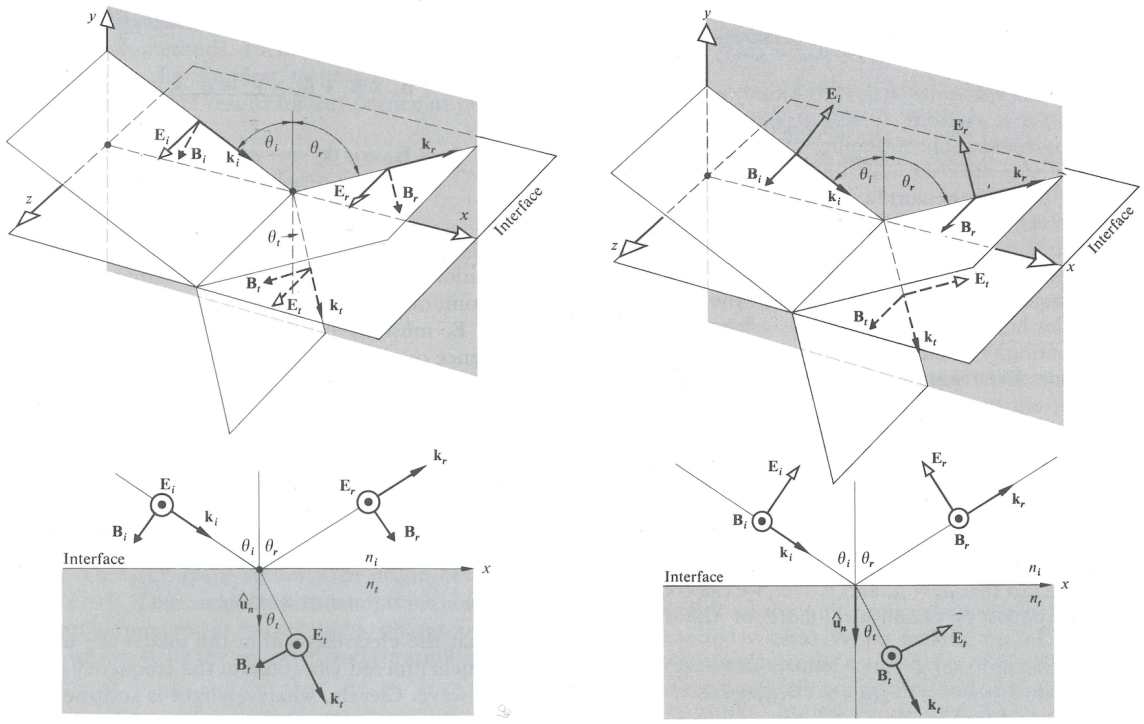


Figure 52: Definition of the **E** and **B** fields on reflection and refraction at an interface between two media, for the case of *s*-mode polarisation (left panel) and *p*-mode polarisation (right panel). Figures from Hecht, "Optics".

A.2 s-mode reflection and refraction

At the interface between the media with refractive indices n_i and n_t in the incident and transmitting medium respectively, the tangential components of the E and H fields should be continuous. Thus,

$$\begin{aligned} E_i + E_r &= E_t \\ -H_i \cos\theta_i + H_r \cos\theta_r &= -H_t \cos\theta_t \,, \end{aligned} \tag{154}$$

where subscripts i, r, t refer to incident, reflected and transmitted waves, respectively. In the following we shall assume the media to be non-magnetic, so that $\mathbf{B} = \mu_0 \mathbf{H}$, and $B = E/v = nE/c$ where v is the velocity of light in the medium and n is the refractive index. The incident and reflected waves are in the same medium, $n_r = n_i$ and $v_r = v_i$, and also $\theta_r = \theta_i$. Thus

$$n_i(E_i - E_r)\cos\theta_i = n_t E_t \cos\theta_t \,. \tag{155}$$

Defining the amplitude reflection coefficient as $r = E_r/E_i$ and the amplitude transmission coefficient as $t = E_t/E_i$, then for the s-mode,

$$\begin{aligned} r_s &= \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \\ t_s &= \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \,. \end{aligned} \tag{156}$$

These are the *Fresnel equations* for the s-mode. You should recognise these from the A2 Electromagnetism course. Using Snell's law, $n_i \sin\theta_i = n_t \sin\theta_t$,

$$r_s = \frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)} \,. \tag{157}$$

A.3 p-mode reflection and refraction

For the p-mode,

$$\begin{aligned} E_i \cos\theta_i - E_r \cos\theta_r &= E_t \cos\theta_t \\ H_i + H_r &= H_t \,, \end{aligned} \tag{158}$$

so

$$n_i(E_i + E_r) = n_t E_t \tag{159}$$

and hence,

$$\begin{aligned} r_p &= \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_t \cos\theta_i + n_i \cos\theta_t} \\ t_p &= \frac{2n_i \cos\theta_i}{n_t \cos\theta_i + n_i \cos\theta_t} \,, \end{aligned} \tag{160}$$

yielding the Fresnel equations for the p-mode. Again using Snell's law,

$$\begin{aligned} r_p &= \frac{\sin\theta_i \cos\theta_i - \sin\theta_t \cos\theta_t}{\sin\theta_i \cos\theta_i + \sin\theta_t \cos\theta_t} \\ &= \frac{\sin 2\theta_i - \sin 2\theta_t}{\sin 2\theta_i + \sin 2\theta_t} \\ &= -\frac{\tan(\theta_t - \theta_i)}{\tan(\theta_t + \theta_i)} \,, \end{aligned} \tag{161}$$

where the last two expressions may be seen to be equivalent using the usual trigonometric relationships for $\sin(\theta_t - \theta_i)$ etc.

A.4 The Brewster angle

The left panel of Fig. 53 shows the r_s and r_p amplitude reflection coefficients for light incident from air onto glass with $n = 1.5$. A particular property of equation 161 is that it goes to zero when $\theta_i + \theta_t = \pi/2$, as seen in the figure. When this condition holds, the reflected ray would be at a right-angle to the transmitted ray. Applying Snell's law again to equation 160, an alternative formulation of the condition $r_p = 0$ is

$$\tan\theta_i = \frac{n_t}{n_i} \tag{162}$$

and this angle of incidence is the *Brewster angle*. For glass with $n = 1.5$ this occurs when $\theta_i \simeq 56°$, and for water with $n = 1.33$ when $\theta_i \simeq 53°$. If we shine unpolarised light onto a plane glass surface at that angle, only s-mode linearly polarised light is reflected - oblique reflections can create linearly polarised light, completely polarised for reflection at the Brewster angle. The effect was first noted by Malus in 1808 but measured more accurately by Brewster in 1815 (Brewster also invented the kaleidoscope!). Light scattered obliquely from natural surfaces may be highly polarised: polaroid sunglasses filter out the horizontal component of polarised light, so that the glare caused by scattered or reflected light from horizontal surfaces, such as the sea, is reduced.

The intensity of the reflected, s-polarised ray is rather weak and its intensity is low. The effect may be used more efficiently as a polariser by stacking surfaces, so that multiple reflections increase the total reflected, polarised intensity.

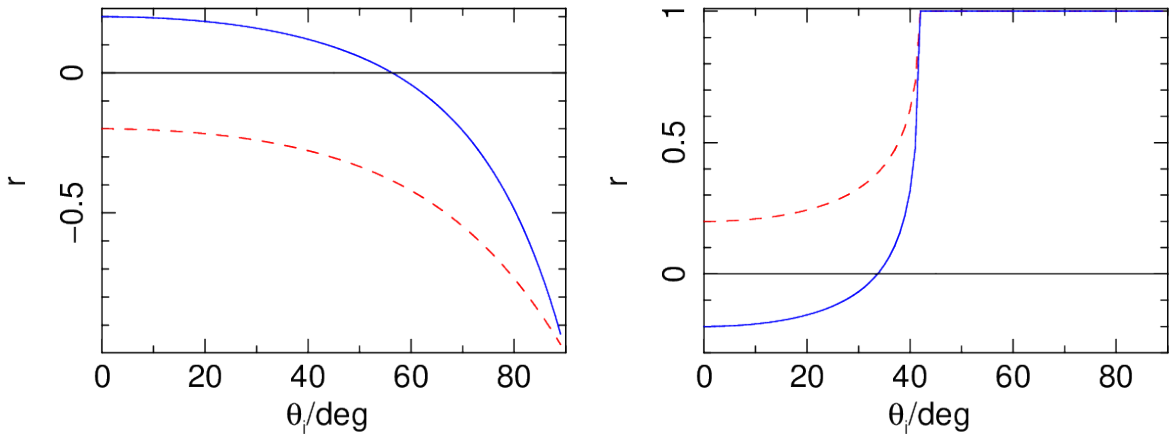


Figure 53: The variation in amplitude reflection coefficient r for s-mode (dashed curve) and p-mode (solid curve), (left panel) for light incident at angle θ_i from air onto glass and (right panel) for light incident from glass onto air.

A.5 Total internal reflection

The right panel of Fig. 53 shows the amplitude reflection coefficients for a reflection of a wave inside glass at an air boundary, i.e. an internal reflection. *Total internal reflection* occurs for both modes when $\sin\theta_t = 1$, i.e. when

$$\sin\theta_i = \frac{n_t}{n_i} \tag{163}$$

where $n_i > n_t$.

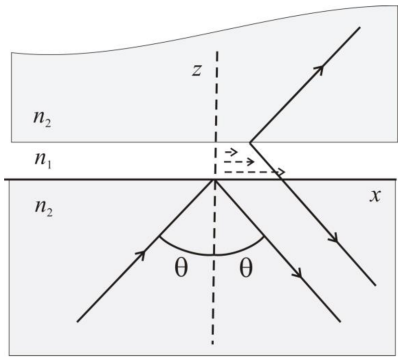


Figure 54: Schematic illustration of frustrated total internal reflection. Light incident on a glass-air interface in the lower piece of glass (refractive index $n_2 > n_1$), at angle θ larger than the total internal reflection angle, generates an evanescent wave (dashed arrows) between the two pieces of glass, but a transverse wave propagates again through the upper piece of glass. The reflected ray in the lower piece of glass is weakened, which may be thought of as arising from an out-of-phase reflected wave that originates at the surface of the upper piece. Diagram from notes by by P. Ewart.

However, for θ_i larger than the total internal reflection angle, the reflection coefficient is positive, so the reflected and incident waves are in phase. Thus it would be impossible to satisfy the boundary condition that we used to derive the reflection coefficient, unless the fields propagate across the boundary. In fact, in a full solution there is a time-oscillatory E-field parallel to the boundary, with imaginary wavenumber, so that the E field does not form a conventional transverse wave, but instead falls off exponentially with distance away from the boundary. Its Poynting vector is zero. This is known as an *evanescent wave*. If we now place a piece of material beyond the boundary, only a short distance away, then we find that the total internally-reflected ray weakens and a transverse wave does propagate through the additional material, as illustrated schematically in Fig. 54. This effect is known as *frustrated total internal reflection* and can even be used in optical systems to obtain a beam-splitter with a controlled, variable reflectance and transmittance. The effect is directly analogous to tunnelling in quantum (wave) mechanics.

A.6 The phase shift of reflected light

The sign of r in Fig. 53 is worth noting: negative r corresponds to a π phase shift. For reflection at a glass surface (left panel) the s-mode always undergoes a phase shift, whereas the p-mode does not, for angles less than the Brewster angle. At larger angles both modes experience a π phase shift.

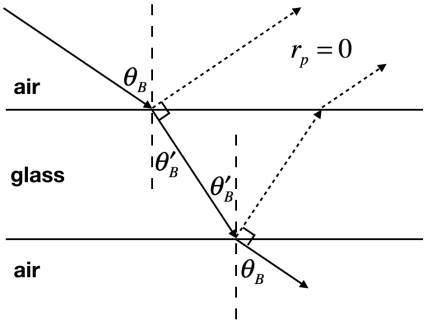


Figure 55: Illustration of the ray geometry for light incident at the Brewster angle onto a piece of glass with parallel sides. Angle $\theta'_B = \pi/2 - \theta_B$.

The case of reflection from two plane, parallel surfaces of a piece of glass is important: we have already met this configuration in Section 4. Consider the case when the incident angle onto the glass is the Brewster angle, when $\theta_B + \theta'_B = \pi/2$, where θ_B is the Brewster angle of incidence and θ'_B is the corresponding angle of refraction. as indicated in Fig. 55. For these parallel surfaces, the rays are at the respective Brewster angles for both the external and the internal reflections. Since the change in sign of r_p occurs at the Brewster angle, inspection of Fig. 53 shows that rays reflected from the two surfaces always have the opposite sign of reflection coefficient to each other, for any angle of incidence, for both s-mode and p-mode. Transmission preserves the phase of the wave (t is always positive), so there is always a relative phase shift of π for rays reflected from the two parallel sides of a plain piece of glass or other material (a metal or dielectric coating on one surface would change this conclusion).