IB Physics A
Oscillations, Waves and Optics

Staff of the Cavendish Laboratory

Michaelmas 2021

Contents

1 Oscillations

1.1 Damped, Driven Oscillations

Many systems exhibit oscillatory behaviour. We start by considering the most simple example which still shows the major features of the behaviour we see in the more complex cases.

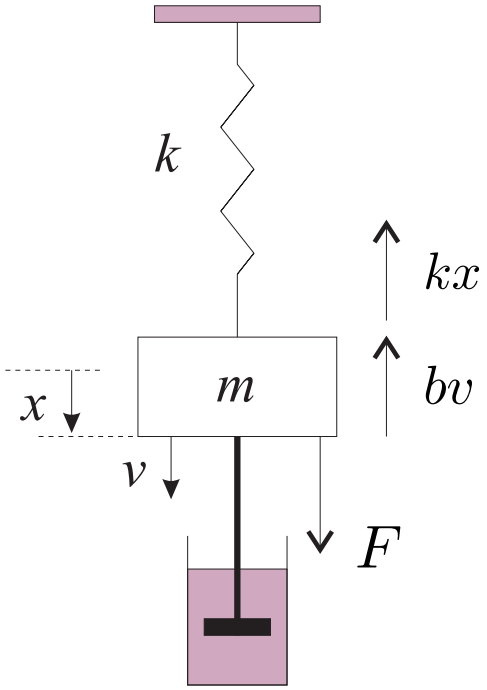


Figure 1.1: Driven, damped oscillator.

Our example is a damped oscillator consisting of a mass m suspended by a spring of spring constant k and attached to a fixed surface via a "damper" of constant b. We drive the mass with a time-varying external force $F(t)$ and the mass displaces by an amount $x(t)$. For both x and F we define the quantity to be positive if it is in a downwards direction. With this setup, then the net force on the mass is given by

$$F_{\text{net}}(t) = F(t) - kx(t) - b\dot{x}(t).$$

where $\dot{x}$ denotes differentiation with respect to time. Using Newton's second law, and dropping the explicit time dependence, we derive an equation of motion

$$m\ddot{x} = F - kx - b\dot{x},$$

which can be rearranged to

$$m\ddot{x} + b\dot{x} + kx = F \tag{1.1}$$

This is a second-order linear differential equation and can be solved using standard mathematical methods.

We can optionally choose to recast the equation in terms of a set of coefficients which are generic to such systems and not specific to the damped oscillator. Conventionally, this is done by writing $\omega_0 = \sqrt{k/m}$ and $\gamma = b/m$ (where we note that both coefficients have units of s^{-1} i.e. frequency or angular frequency), so that we get the equation in its *canonical form*:

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = F/m. \tag{1.2}$$

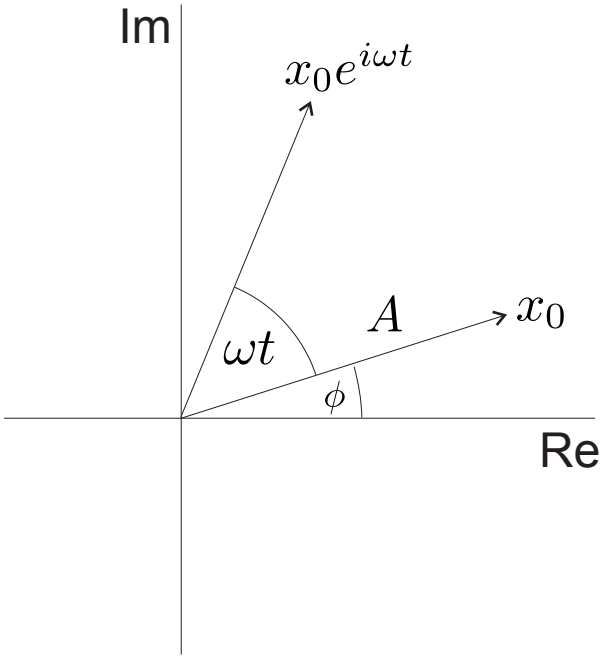


Figure 1.2: Complex notation.

1.1.1 Complex Notation

We will at first restrict ourselves to looking at the response to a *sinusoidal* driving force. As you have already seen last year, it is often convenient to describe sinusoidal waveforms using complex notation. For example, we can describe the motion represented by

$$x = A\cos(\omega t + \phi)$$

by writing instead

$$x = \Re\left(x_0\, e^{i\omega t}\right) \tag{1.3}$$

where the complex coefficient x_0 is given by

$$x_0 = A\, e^{i\phi}, \tag{1.4}$$

since

$$\begin{aligned} x &= \Re\left[A\, e^{i\phi} e^{i\omega t}\right] \\ &= \Re\left[A\, e^{i(\omega t+\phi)}\right] \\ &= A\,\cos(\omega t + \phi) \end{aligned}$$

as required. In other words, the modulus of x_0 represents the amplitude and the argument of x_0 represents the phase at time $t = 0$:

$$\begin{aligned} |x_0| &= A \\ \arg(x_0) &= \phi \end{aligned}$$

We will use the term "complex displacement" for the coefficient x_0, though it is somewhat arbitrary whether we denote x_0 or $x_0\, e^{i\omega t}$ as the complex displacement; in practice this is usually clear from the context.

To find the velocity, we take the time derivative of the displacement,

$$v = \frac{d}{dt}\Re\left\{x_0 e^{i\omega t}\right\} = \Re\left\{i\omega x_0 e^{i\omega t}\right\},$$

where we have made use of the fact that we can interchange (i.e. "commute") the order of taking a derivative with taking the real part. Thus we can write

$$v = \Re\left\{v_0 e^{i\omega t}\right\}$$

where the "complex velocity" v_0 is given by

$$v_0 = i\omega x_0. \tag{1.5}$$

Using the same procedure, the complex acceleration a_0 is given by

$$a_0 = i\omega v_0 = -\omega^2 x_0. \tag{1.6}$$

1.1.2 Response to sinusoidal driving forces

Writing the driving force as $F = \Re\left[F_0 e^{i\omega t}\right]$ (where F_0 is in general complex so as to account for the fact that the driving sinusoid may not be at phase 0 at time $t = 0$) we will look for steady-state oscillating solutions of the form

$$x = \Re\left[x_0 e^{i\omega t}\right]. \tag{1.7}$$

Substituting equation 1.7 into equation 1.2 we get

$$\Re\left\{[-\omega^2 + i\gamma\omega + \omega_0^2]x_0 e^{i\omega t}\right\} = \Re\left\{\frac{F_0}{m}e^{i\omega t}\right\}. \tag{1.8}$$

Since this is true for all t then we can remove the $\Re\{\}$ from both sides of the equation: this is because the effect of multiplying a complex number by $e^{i\omega t}$ is to rotate the imaginary part of the number onto the real axis when $\omega t = \pi/2, 3\pi/2 \ldots$; thus both the real and imaginary parts of the coefficients of $e^{i\omega t}$ are equal. Rearranging the equation then gives the solution for x_0 in terms of F_0

$$x_0 = \frac{F_0}{m[(\omega_0^2 - \omega^2) + i\gamma\omega]}. \tag{1.9}$$

We can rewrite this in terms of a *response function* $R(\omega)$, where

$$x_0 = R(\omega)F_0 \tag{1.10}$$

and

$$R(\omega) = \frac{1}{m[(\omega_0^2 - \omega^2) + i\gamma\omega]}. \tag{1.11}$$

$R(\omega)$ is a complex function that describes the displacement relative to the driving force, as a function of the driving frequency ω. The modulus of $R(\omega)$ gives the amplitude of the sinusoidal displacement for a given force and the argument of $R(\omega)$ gives the phase shift of the displacement with respect to the force sinusoid.

Several important features can be seen in the modulus and phase of the response function (Figure 1.3). At low frequencies ($\omega \Rightarrow 0$) the response is dominated by the spring constant, and is in phase with the force. At high frequencies ($\omega \Rightarrow \infty$) the response is dominated by the inertia of the mass. The amplitude of the response tends to zero, and the response is π out of phase with the

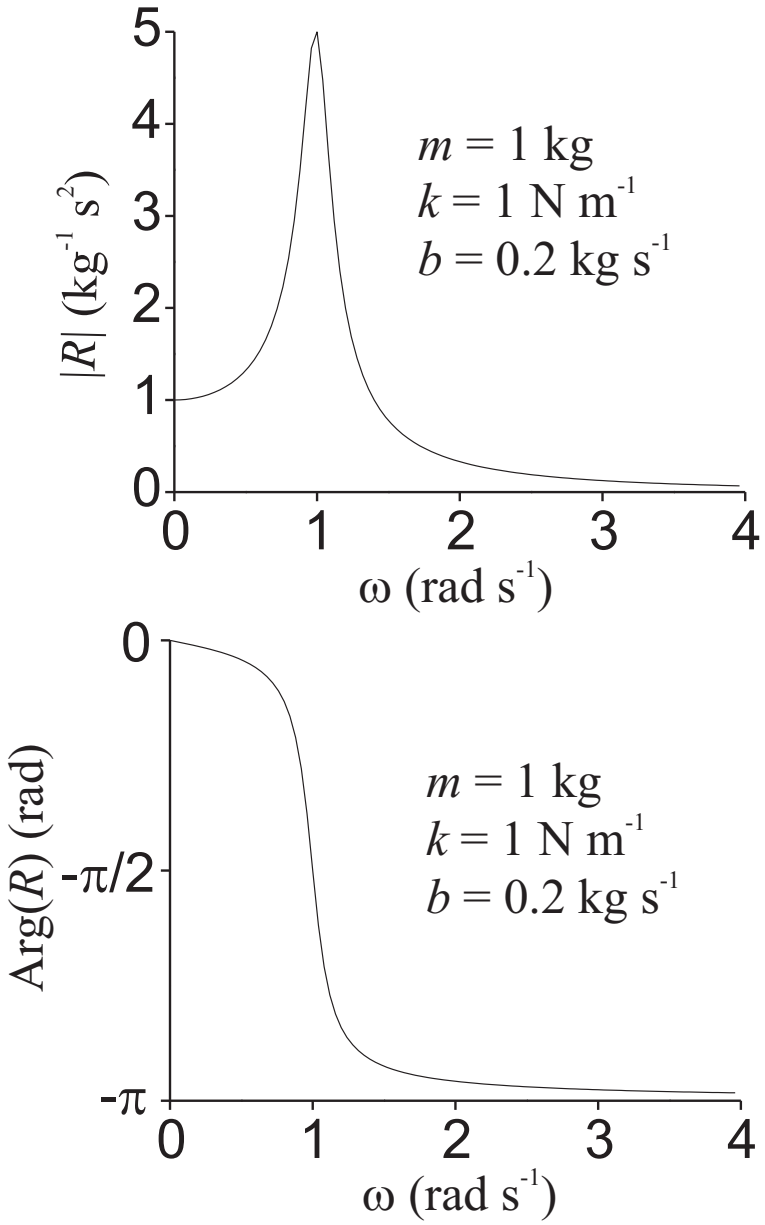


Figure 1.3: The amplitude and phase of the response function $R(\omega)$ for a damped oscillator.

driving force. At an intermediate frequency, ω_a, amplitude resonance occurs, corresponding to a maximum in the amplitude of response.

For light damping (i.e. $\gamma \ll \omega_0$), the magnitude of the response function is approximately maximised when the real part of the denominator is zero, which occurs at $\omega_a^2 - \omega_0^2 = 0$ i.e. $\omega_a \approx \omega_0$. This is the same as the frequency of free oscillations of the undamped oscillator. The response function at resonance is given by

$$R(\omega_a) \approx \frac{1}{i\omega_0\gamma}, \tag{1.12}$$

hence the amplitude of oscillations is inversely proportional to the damping, γ, and the response is $\pi/2$ behind the driving force.

1.1.3 Q factor

To understand the effects of different amounts of damping, it is helpful to recast the damping in a dimensionless form by writing

$$Q = \frac{\omega_0}{\gamma} \tag{1.13}$$

This "quality-factor" will be seen multiple times in your physics course and is a key parameter of a damped oscillator: $Q \gg 1$ when the damping is small, i.e. when the system is close to being a simple harmonic oscillator. The Q factor controls the shape of the response curve: as Q decreases, the resonance becomes broader, the amplitude at resonance is decreased, and the resonant frequency shifts to lower values (Figure 1.4). Substituting equation 1.13 into equation 1.11, we have

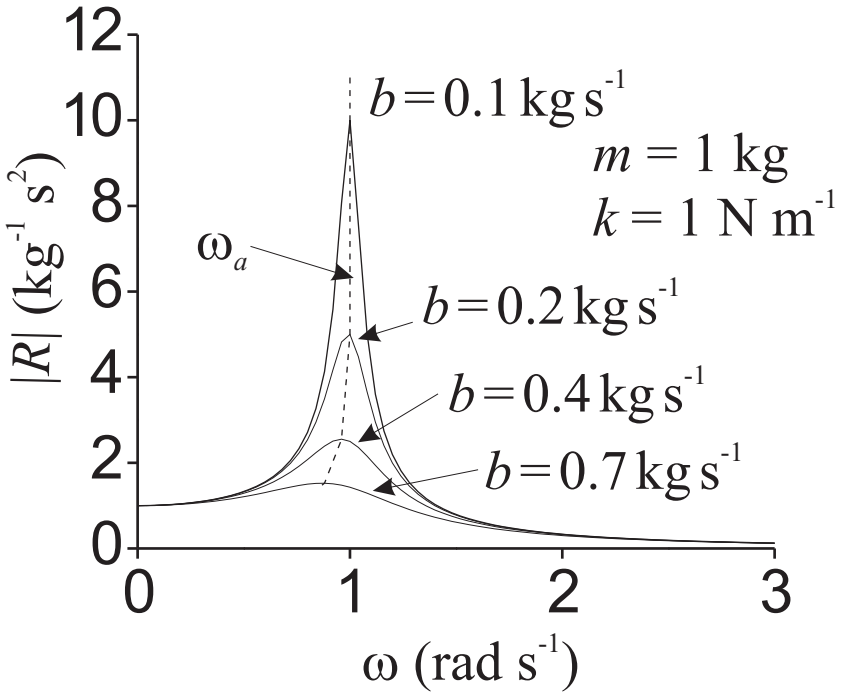


Figure 1.4: Response function for various damping parameters.

$$R(\omega) = \frac{1}{m[(\omega_0^2 - \omega^2) + i\omega_0\omega/Q]} \tag{1.14}$$

To find the resonant frequency exactly, we need to maximise $|R(\omega)|$ which is the same as minimising the modulus of the denominator in the equation above i.e. $\left|(\omega_0^2 - \omega^2) + i\omega_0\omega/Q\right|^2$. This

is most easily done by minimising the *square* of this quantity with respect to ω^2.

$$\begin{aligned}\frac{\partial}{\partial(\omega^2)}\left[(\omega_0^2-\omega^2)^2+(\omega_0\omega/Q)^2\right] &= 0\\ \Rightarrow -2(\omega_0^2-\omega^2)+(\omega_0/Q)^2 &= 0\\ \Rightarrow \omega^2 &= \omega_0^2[1-1/(2Q^2)]\\ \Rightarrow \omega &= \omega_0\sqrt{1-1/(2Q^2)}\end{aligned}$$

As mentioned above, for very light damping ($Q \gg 1$), the resonant frequency tends to ω_0: for $Q = 10$ the resonant frequency is only 0.25% different from ω_0.

1.1.4 Velocity response

We can also examine the velocity response as a function of frequency. This will be useful when we look at the power response. From Eqs. 1.5 and 1.11, the complex velocity is given by

$$\begin{aligned}v_0 &= i\omega x_0 = i\omega F_0 R(\omega)\\ &= \frac{i\omega F_0}{m[(\omega_0^2-\omega^2)+i\gamma\omega]}\\ &= \frac{F_0}{m[(\omega_0^2-\omega^2)/(i\omega)+\gamma]}\end{aligned} \tag{1.15}$$

The maximum velocity amplitude (**velocity resonance**) occurs at the angular frequency where $|v_0|$ is at a maximum. Given that

$$|v_0| = \frac{|F_0|}{m\sqrt{(\omega_0^2-\omega^2)^2/\omega^2+\gamma^2}}, \tag{1.16}$$

it can be seen by inspection of the denominator that the velocity resonance (angular) frequency ω_v is given by $\omega_v = \omega_0$ independent of the damping, and from Eq. 1.15 we can see that, at this frequency, the velocity is exactly in phase with the driving force.

1.2 Power

1.2.1 Multiplying Quantities using Complex Notation

It is frequently necessary to calculate product of two oscillating quantities, for example to work out the mean power absorbed by an oscillating system by taking the product of force and velocity. Where the force and velocity are described in terms of complex notation, $F(t) = \Re\left[F_0 e^{i\omega t}\right]$ and $v(t) = \Re\left[v_0 e^{i\omega t}\right]$ then we need to be careful about how we take the product of these two quantities.

In general the product of the real parts of any two complex numbers $\mathbf{A}$ and $\mathbf{B}$ is

$$\begin{aligned}\Re\{\mathbf{A}\}\Re\{\mathbf{B}\} &= \tfrac{1}{2}(\mathbf{A}+\mathbf{A}^*)\tfrac{1}{2}(\mathbf{B}+\mathbf{B}^*)\\ &= \tfrac{1}{4}(\mathbf{AB}+\mathbf{A}^*\mathbf{B}^*+\mathbf{AB}^*+\mathbf{A}^*\mathbf{B})\\ &= \tfrac{1}{2}\Re\{\mathbf{AB}+\mathbf{AB}^*\}\end{aligned} \tag{1.17}$$

which is *not* the same as the real part of the products. Applying this to the power calculation we get

$$P = Fv = \tfrac{1}{2}\Re\left[F_0 e^{i\omega t}\right]\Re\left[v_0 e^{i\omega t}\right] \tag{1.18}$$

$$= \tfrac{1}{2}\Re\left[F_0 v_0 e^{2i\omega t} + F_0 v_0^*\right]. \tag{1.19}$$

When we take the average over time, the term oscillating at 2ω averages to zero giving

$$\langle P\rangle = \tfrac{1}{2}\Re\left[F_0 v_0^*\right], \tag{1.20}$$

where $\langle\rangle$ denotes taking the mean value.

This result is a general way to calculate the time average of the product of two sinusoidally varying quantities represented using complex notation. The order of operation is not important ($\Re[F_0 v_0^*] = \Re[v_0 F_0^*]$).

1.2.2 Phase differences and power factor

Equation 1.20 shows that the mean power input depends on not only the amplitudes of the force and velocity, as one might expect, but also the *phase difference* between the force and the velocity, since if $F_0 = F_0 e^{i\phi_F}$ and $v_0 = v_0 e^{i\phi_v}$ then

$$\begin{aligned}\tfrac{1}{2}\Re[F_0 v_0^*] &= \tfrac{1}{2}\Re\left[F_0 e^{i\phi_F} v_0 e^{-i\phi_v}\right]\\ &= \tfrac{1}{2}\Re\left[F_0 v_0 e^{i(\phi_F-\phi_v)}\right]\\ &= \tfrac{1}{2}F_0 v_0\cos(\phi_F-\phi_v).\end{aligned} \tag{1.21}$$

The term $\cos(\phi_F - \phi_v)$ is known as the **power factor** in electrical circuits (after replacing force and velocity by voltage and current). If the force and velocity are are in phase ($\phi_F = \phi_v$), $\langle P\rangle = \frac{1}{2}F_0 v_0$ while if they are $\pi/2$ out of phase then $\langle P\rangle = 0$.

1.2.3 Power in oscillators

To calculate the mean power required to drive a damped, driven oscillator we can substitute the value of F_0 from equation 1.15 into equation 1.20

$$\begin{aligned}\langle P\rangle &= \tfrac{1}{2}\Re\left\{F_0 v_0^*\right\}\\ &= \tfrac{1}{2}\Re\left\{v_0 v_0^* m[(\omega_0^2-\omega^2)/(i\omega)+\gamma]\right\}\\ &= \tfrac{1}{2}m\gamma|v_0|^2 = \tfrac{1}{2}b|v_0|^2\end{aligned} \tag{1.22}$$

Clearly the average rate of work done on the system by the driving force must equal the average power dissipated within the system. We can check this by noting that the energy is dissipated in the damping term $F_r = bv_0$ which is always in phase with the velocity, so

$$\langle P_{\text{dissipated}}\rangle = \tfrac{1}{2}|F_r||v_0| = \tfrac{1}{2}b|v_0|^2, \tag{1.23}$$

as required.

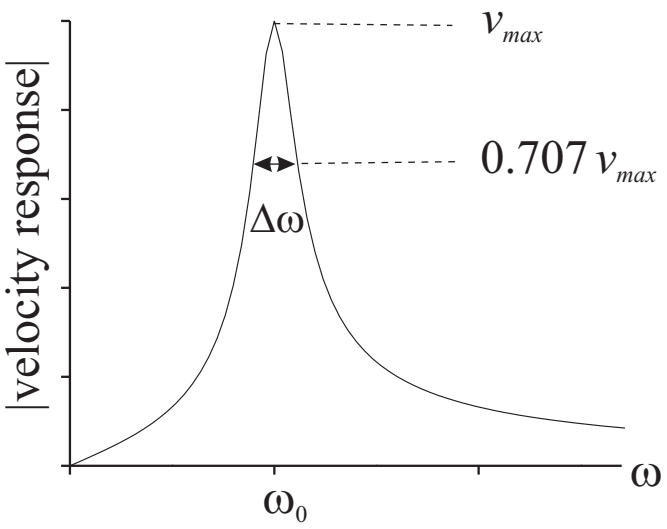


Figure 1.5: Bandwidth of a driven damped oscillator.

1.2.4 Bandwidth

It is clear from Eq. 1.22 that the maximum mean power required for a given force F_0 (i.e. **power resonance**) occurs at the velocity resonance frequency, i.e. $\omega_P = \omega_v = \omega_0$. To measure the width of the resonance, we define the *half power points* as the points where the power absorbed is half the value at velocity resonance. From Eq. 1.22, this corresponds to the two frequencies ω_+ and ω_- at which $|v_0| = |v_{\max}|/\sqrt{2}$. From equation 1.16 we have

$$|v_0|^2 = \frac{|F_0|^2}{m^2[(\omega_0^2 - \omega^2)^2/\omega^2 + \gamma^2]}. \tag{1.24}$$

which has a value of $|F_0|^2/(m\gamma)^2$ at resonance, and so the half power points occur when

$$\begin{aligned} \gamma^2\omega_\pm^2 &= (\omega_0^2 - \omega_\pm^2)^2 \\ \Rightarrow \gamma\omega_\pm &= \mp(\omega_0^2 - \omega_\pm^2) \end{aligned} \tag{1.25}$$

The frequency difference between the two half power points, $\Delta\omega = \omega_+ - \omega_-$, is a measure of the *bandwidth* of the resonance. We can arrive at this by using Eq. 1.25 to compute an expression for the *sum* of the frequencies:

$$\begin{aligned} \gamma\omega_+ + \gamma\omega_- &= -(\omega_0^2 - \omega_+^2) + (\omega_0^2 - \omega_-^2) \\ &= \omega_+^2 - \omega_-^2 \\ &= (\omega_+ - \omega_-)(\omega_+ - \omega_-) \\ \Rightarrow \omega_+ - \omega_- &= \gamma \end{aligned} \tag{1.26}$$

Thus γ, which we remember has units of s^{-1}, in fact directly gives the bandwidth $\Delta\omega$ of the resonance. As expected, the peak gets wider as the damping goes up.

The dimensionless ratio Q gives the ratio of the resonant frequency to the bandwidth

$$Q = \frac{\omega_0}{\gamma} = \frac{\omega_0}{\Delta\omega}. \tag{1.27}$$

Thus by finding the resonant frequency and the half-power points of the frequency response, we can directly measure the Q value of an oscillator (see later for a different way of measuring Q).

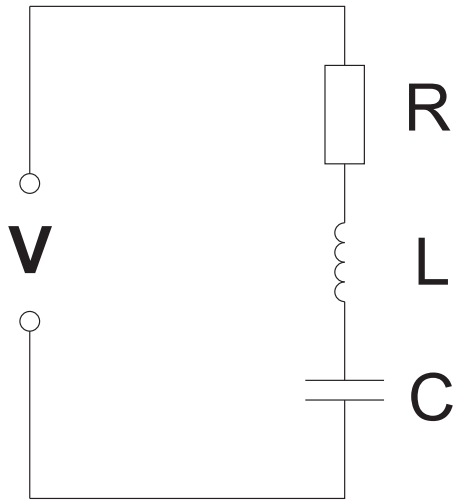


Figure 1.6: LCR circuit.

1.3 Electrical Circuits

The ideas above apply equally well to electrical resonant circuits. In particular, the series LCR circuit shown in Figure 1.6 is exactly analogous to the damped, driven mechanical oscillator since its equation of motion is of the form

$$L\ddot{q} + R\dot{q} + \frac{1}{C}q = V(t) \tag{1.28}$$

where q is the charge on the capacitor.

The equivalent quantities in the mechanical and electrical cases are shown below.

displacement, x	charge, q
velocity, v	current, I
force, F	voltage, V
mass, m	inductance, L
damping, b	resistance, R
spring constant, k	$(\text{capacitance})^{-1}$, $1/C$

We can apply all the results derived from solving the canonical equation (Eq. 1.2) by substituting for the values of ω_0, γ, and Q

$$\begin{aligned} \omega_0 &= 1/\sqrt{LC} \\ \gamma &= R/L \\ Q &= \frac{\omega_0}{\gamma} = \frac{1}{R}\sqrt{\frac{L}{C}} \end{aligned}$$

so that we have

$$\ddot{q} + \gamma\dot{q} + \omega_0^2 q = V/L \tag{1.29}$$

1.3.1 Impedance

Electrical impedance is a particularly useful concept to describe the relationship between voltage and current, as you saw last year. If a voltage of $V = \Re\{V_0 e^{i\omega t}\}$ gives rise to a current of $I =$

$\Re\{I_0 e^{i\omega t}\}$, where V_0 and I_0 are complex, then the impedance Z is given by

$$Z = V_0/I_0. \tag{1.30}$$

The electrical impedance of the circuit above is

$$Z = i\omega L + \frac{1}{i\omega C} + R = Z_L + Z_C + Z_R. \tag{1.31}$$

The power dissipated is given by

$$\begin{aligned} \langle P \rangle &= \tfrac{1}{2}\Re\{V_0 I_0^*\} \\ &= \tfrac{1}{2}|V_0|^2\Re\{1/Z\} = \tfrac{1}{2}|I_0|^2\Re\{Z\} \\ &= \tfrac{1}{2}|I_0|^2 R. \end{aligned} \tag{1.32}$$

The analogy in mechanical oscillators is also called the impedance and is the ratio of the complex force to the complex velocity

$$Z = F_0/v_0 \tag{1.33}$$

which from Eq. 1.15 is given by

$$Z = m[(\omega_0^2 - \omega^2)/(i\omega) + \gamma] \tag{1.34}$$

We can write down expressions for the mean power dissipated in terms of the impedance as

$$\begin{aligned} \langle P \rangle &= \tfrac{1}{2}\Re\{F_0 v_0^*\} \\ &= \tfrac{1}{2}|F_0|^2\Re\{1/Z\} = \tfrac{1}{2}|v_0|^2\Re\{Z\}, \end{aligned} \tag{1.35}$$

which can be compared with the electrical equivalents.

1.4 Transient Response

The previous sections looked at the response of oscillators to sinusoidal driving forces. The sinusoids examined had no start time and no end time so we have implicitly derived only the steady-state response. We can also examine the effects of initial conditions (boundary conditions) on the response of an oscillator. In doing so, we change the independent variable from angular frequency ω to time t. Consider first the un-driven harmonic oscillator

$$\ddot{x}(t) + \gamma\dot{x}(t) + \omega_0^2 x(t) = 0. \tag{1.36}$$

where, as usual, $\omega_0 = \sqrt{k/m}$ and $\gamma = b/m$ in the case of a mechanical oscillator. We try solutions of the general form Ae^{pt}, where p can in general be complex, which gives

$$p^2 + \gamma p + \omega_0^2 = 0. \tag{1.37}$$

This quadratic equation has solutions

$$\begin{aligned} p_{1,2} &= \frac{-\gamma \pm \sqrt{\gamma^2 - 4\omega_0^2}}{2} \\ &= -\frac{\gamma}{2}\left(1 \pm \sqrt{1 - 4Q^2}\right) \end{aligned} \tag{1.38}$$

Each value of p gives an independent solution to Eq. 1.36, and these solutions can be added linearly to give the general solution

$$x = A_1 e^{p_1 t} + A_2 e^{p_2 t}. \tag{1.39}$$

Constants A_1 and A_2 are chosen to satisfy the boundary conditions, which are often specified as a displacement and velocity at $t = 0$.

The form of solution depends on Q, i.e. on the level of damping in the system. Three cases can be distinguished, depending on whether $Q < 0.5$, $Q > 0.5$ or $Q = 0.5$.

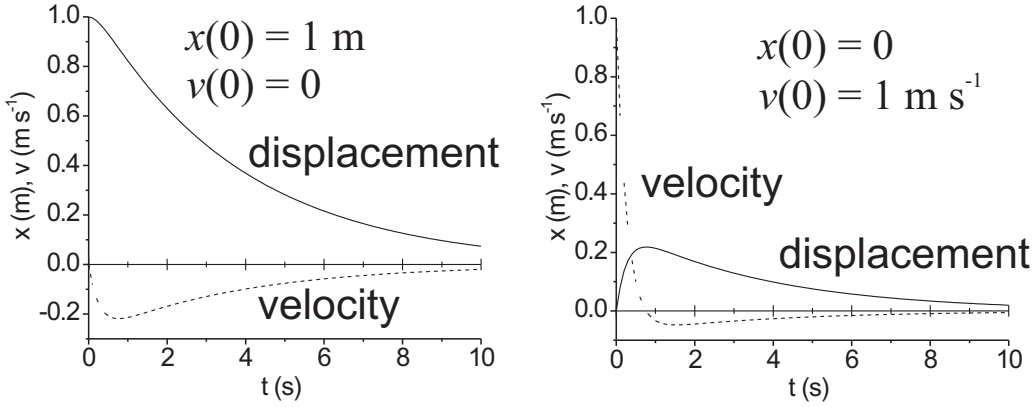


Figure 1.7: Transient response of a heavily damped oscillator ($m = 1\,\mathrm{kg}$, $k = 1\,\mathrm{N\,m^{-1}}$, $b = 4\,\mathrm{kg\,s^{-1}}$).

1.4.1 Overdamping or heavy damping ($Q < 0.5$)

Here both values of p are real and negative. The solution therefore consists of two decaying exponentials.

$$x(t) = C_1 e^{-\mu_1 t} + C_2 e^{-\mu_2 t}, \tag{1.40}$$

where

$$\mu_{1,2} = \tfrac{1}{2}\gamma\left(1 \pm \sqrt{1 - 4Q^2}\right) \tag{1.41}$$

The system returns slowly to equilibrium. Figure 1.7 shows overdamped behaviour for various initial conditions.

1.4.2 Underdamping or light damping ($Q > 0.5$)

Here p is complex, with a negative real part.

$$p_{1,2} = -\tfrac{1}{2}\gamma \pm i\omega_f, \tag{1.42}$$

where ω_f is the "free oscillation frequency" given by

$$\omega_f = \omega_0\sqrt{1 - 1/(4Q^2)}. \tag{1.43}$$

It can be seen that for large Q the free oscillation frequency is close to ω_0. The solution now contains an oscillating part, with an exponentially decaying envelope.

$$x = e^{-\frac{\gamma t}{2}}\left(A_1 e^{i\omega_f t} + A_2 e^{-i\omega_f t}\right). \tag{1.44}$$

Constants A_1 and A_2 can be complex, but in order to ensure that x is real, we require that $A_2 = A_1^*$. The solution can then be written in a number of different, but equivalent, forms, each with two constants determined by the boundary conditions

$$x = e^{-\frac{\gamma t}{2}}\left(Ae^{i\omega_f t} + A^* e^{-i\omega_f t}\right) \tag{1.45}$$

$$x = e^{-\frac{\gamma t}{2}}\Re\{Ce^{i\omega_f t}\} \tag{1.46}$$

$$x = e^{-\frac{\gamma t}{2}} B\cos\left(\omega_f t + \phi\right) \tag{1.47}$$

$$x = e^{-\frac{\gamma t}{2}}\left(B_1\cos\omega_f t + B_2\sin\omega_f t\right). \tag{1.48}$$

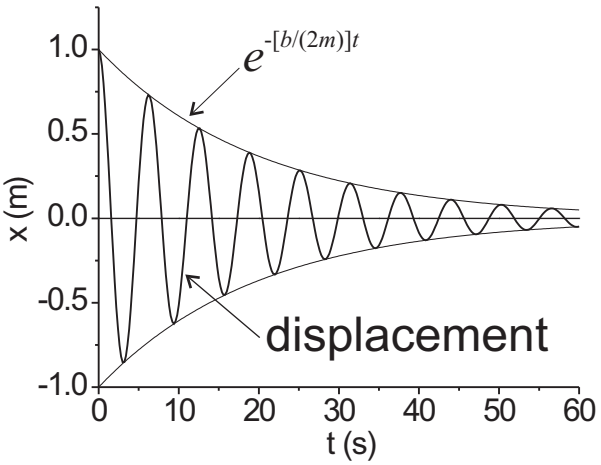


Figure 1.8: Transient response of a lightly damped oscillator ($m = 1\,\mathrm{kg}$, $k = 1\,\mathrm{N\,m^{-1}}$, $b = 0.1\,\mathrm{kg\,s^{-1}}$).

Measuring Q from transient oscillations

For a system with very light damping ($Q \gg 0.5$), the oscillations occur at a frequency close to ω_0. The amplitude decays as $e^{(-\gamma/2)t}$, and will therefore have decayed by a factor of e at time $t_1 = 2/\gamma$. During this time, the phase of the oscillation will have advanced by $2\omega_0/\gamma$ radians, corresponding to $\omega_0/(\pi\gamma)$ oscillations. Comparing with our definition of Q-factor (Eq. 1.13)

$$Q = \frac{\omega_0}{\gamma} \tag{1.49}$$

we can see that

$$Q = \pi \times \left(\text{ number of oscillations for amplitude to decay by factor } e \right) \tag{1.50}$$

We can also consider the time for the intensity of the oscillations to decay by a factor of e, corresponding to the amplitude decaying by a factor of $\sqrt{e}$. Since the decay is exponential, this will take half the time it took for the amplitude to decay by a factor of e. Hence we can see that

$$Q = 2\pi \times \left(\begin{array}{c} \text{number of oscillations for intensity to decay by} \\ \text{factor } e \end{array} \right) \tag{1.51}$$

$$= \left(\text{ number of radians for intensity to decay by factor } e \right) \tag{1.52}$$

1.4.3 Critical damping ($Q = 0.5$)

This is a special case, where only one value of p exists. Since we always need two arbitrary constants to satisfy our boundary conditions, we must now look for a different form of solution, which turns out to be

$$x = (A_1 + A_2 t)e^{pt} = (A_1 + A_2 t)e^{\frac{-\gamma t}{2}}. \tag{1.53}$$

This form of solution can be checked by substituting it back into the equation of motion. Examples of critical damping are shown in Figure 1.9 for the cases of zero initial velocity and zero initial displacement.

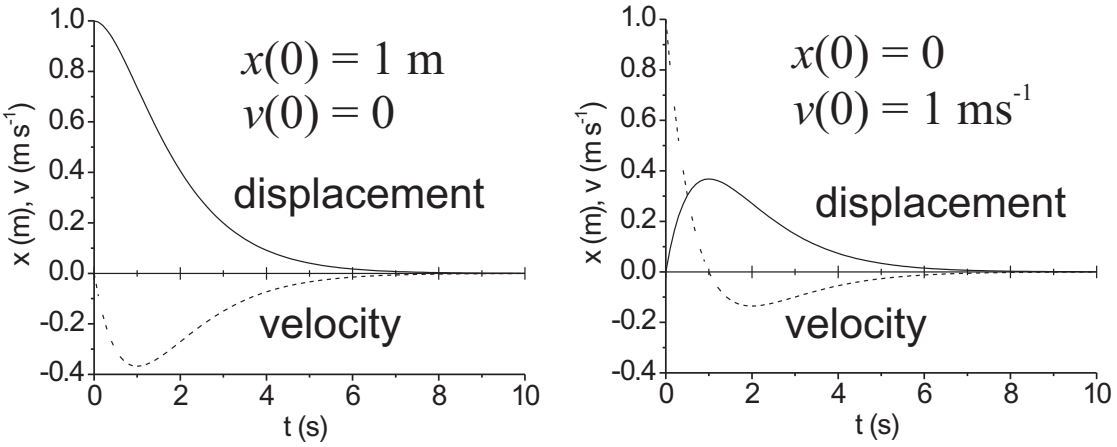


Figure 1.9: Transient response of a critically damped oscillator ($m = 1\,\mathrm{kg}$, $k = 1\,\mathrm{N\,m^{-1}}$, $b = 2\,\mathrm{kg\,s^{-1}}$).

Critical damping represents the behaviour which avoids oscillations, but where the system returns towards the origin as quickly as possible (i.e. the decay is more rapid than for heavy damping). Critical damping is used for example in the suspension of a car to give the smoothest possible ride.

1.5 Combining driven and transient oscillations

The general equation for a driven, damped oscillator is a second-order differential equation of the form

$$\ddot{x}(t) + \gamma\dot{x}(t) + \omega_0^2 x(t) = F(t)/m. \tag{1.54}$$

The general solution to this equation can be written

$$x(t) = \text{Complementary function} + \text{Particular integral} \tag{1.55}$$

where the complementary function is the solution to

$$\ddot{x}(t) + \gamma\dot{x}(t) + \omega_0^2 x(t) = 0. \tag{1.56}$$

and the particular integral is *any* solution of Eq. 1.54 which is linearly independent of the complementary function.

Imagine now that at $t = 0$ we apply a sinusoidally oscillating force $F\cos(\omega t)$ to a damped oscillator initially undisplaced and at rest. The complementary function is just the transient response we have found above. The particular integral is the steady-state oscillation at frequency ω with a phase and amplitude which we found in Section 1.1. The total solution will have two constants from the complementary function, which we can adjust to satisfy the boundary conditions at $t = 0$. The transient response will eventually die away, leaving just the steady-state solution (Figure 1.10).

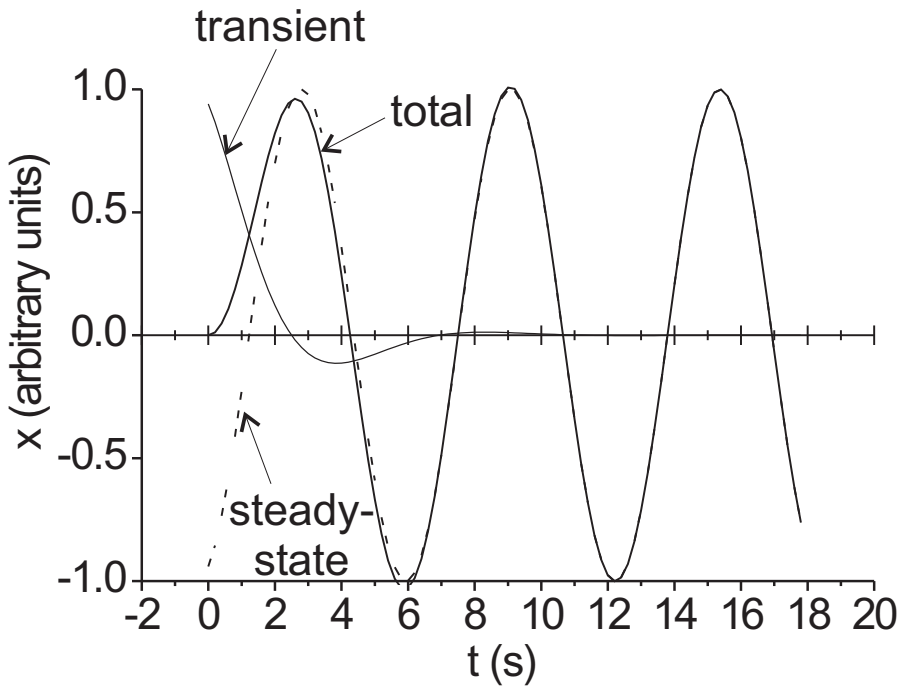


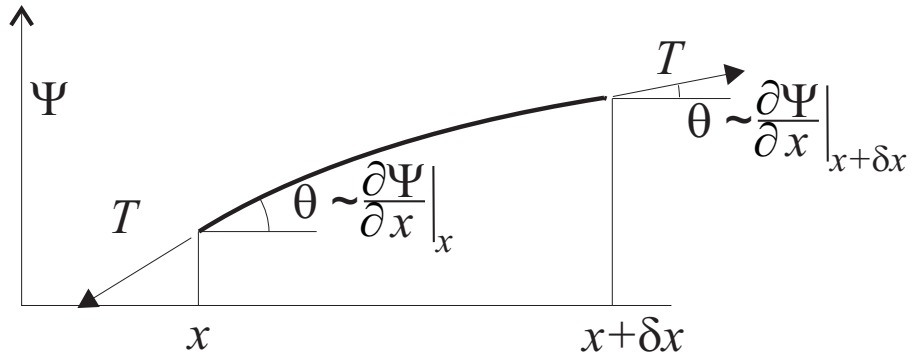
Figure 1.10: Initial response of a damped oscillator to a force $F\cos(\omega t)$ ($m = 1\,\mathrm{kg}$, $k = 1\,\mathrm{m\,s^{-1}}$, $b = 1\,\mathrm{kg\,s^{-1}}$).

2 Waves

2.1 Waves on a string

A wave is the means by which information about a disturbance at one place is carried to another without bulk translation of the intervening medium. You have already experienced many examples of wave motion last year, and have seen how the ideas of wave theory can be used to treat the problem of diffraction at simple obstructions. Here, we want to extend and generalise these concepts in order to give a more complete picture of wave motion. Many of the concepts introduced here will appear again in other courses this year, for example in quantum mechanics and in electromagnetism.

Waves travelling on a stretched string provide a useful simple example for wave motion.



The transverse displacement is denoted $\Psi(x,t)$. Consider a segment of string of length δx and mass $\rho\,\delta x$ extending from x to $x+\delta x$. The restoring force acting on one end is $T\sin\theta$, which in the limit of small angles is equivalent to $T\left.\frac{\partial\Psi}{\partial x}\right|_x$. Similarly at the other end of the element, the restoring force is $-T\left.\frac{\partial\Psi}{\partial x}\right|_{x+\delta x}$. The total restoring force is

$$T\left(\left.\frac{\partial\Psi}{\partial x}\right|_x - \left.\frac{\partial\Psi}{\partial x}\right|_{x+\delta x}\right)$$

which, using the definition of the second derivative becomes

$$T\left(\left.\frac{\partial\Psi}{\partial x}\right|_x - \left\{\left.\frac{\partial\Psi}{\partial x}\right|_x + \left.\frac{\partial^2\Psi}{\partial x^2}\right|_x \delta x\right\}\right) = -T\left.\frac{\partial^2\Psi}{\partial x^2}\right|_x \delta x$$

Finally, using Newton's second law, the equation of motion is

$$-\rho\frac{\partial^2\Psi}{\partial t^2}\,\delta x = -T\frac{\partial^2\Psi}{\partial x^2}\,\delta x$$

or

$$\frac{\partial^2\Psi}{\partial t^2} = v^2\frac{\partial^2\Psi}{\partial x^2} \tag{2.1}$$

where $v=\sqrt{T/\rho}$. We will now show that Eq. 2.1 is a *wave equation*, whose solutions consists of functions propagating at speed v.

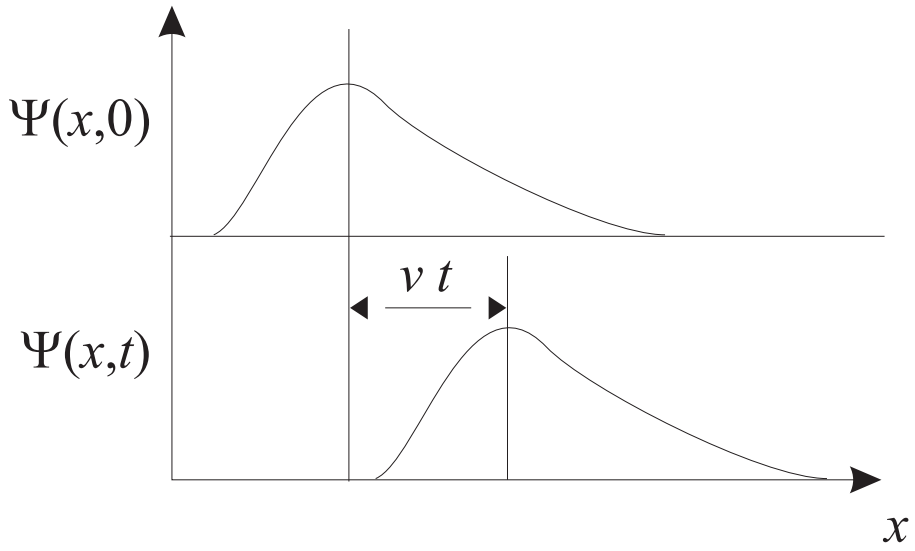


Figure 2.1: Wave motion.

2.2 Wave motion

To arrive at solutions to the 1-d wave equation, we consider a general disturbance $\Psi(x,t)$. For example, Figure 2.1 shows the disturbance that might be produced on a stretched string by moving one end sideways and then back again. As time progresses the disturbance will propagate along the string away from its source - this is described as wave motion. If the disturbance propagates unchanged at speed v, then a snapshot taken at time t will look as shown in Figure 2.1.

At time $t = 0$ let the function $f(x)$ describe the shape of the wave, so that $\Psi(x,0) = f(x)$. Since we assume the disturbance travels at speed v without changing its shape, then at time t later it will have moved a distance vt to the right. Thus

$$\begin{aligned} \Psi(x,t) &= \Psi(x-vt,0) \\ &= f(x-vt). \end{aligned}$$

(For a wave travelling to the left, $\Psi(x,t) = f(x+vt)$.))

We have just looked at the disturbance in space at some instant of time. We can also look at the disturbance in time at some position, as shown in Figure 2.2.

These two pictures are connected by the wave equation which relates the two second order partial derivatives of Ψ, $\partial^2\Psi/\partial x^2$ and $\partial^2\Psi/\partial t^2$. Let $u = x - vt$, then $\Psi(x,t) = f(x-vt) = f(u)$. Using the chain rule,

$$\frac{\partial\Psi}{\partial x} = \frac{df}{du}\frac{\partial u}{\partial x} = \frac{df}{du} \tag{2.2}$$

and

$$\frac{\partial^2\Psi}{\partial x^2} = \frac{d^2f}{du^2}\frac{\partial u}{\partial x} = \frac{d^2f}{du^2}. \tag{2.3}$$

Similarly

$$\frac{\partial\Psi}{\partial t} = \frac{df}{du}\frac{\partial u}{\partial t} = -v\frac{df}{du} \tag{2.4}$$

and

$$\frac{\partial^2\Psi}{\partial t^2} = -v\frac{d^2f}{du^2}\frac{\partial u}{\partial t} = v^2\frac{d^2f}{du^2}. \tag{2.5}$$

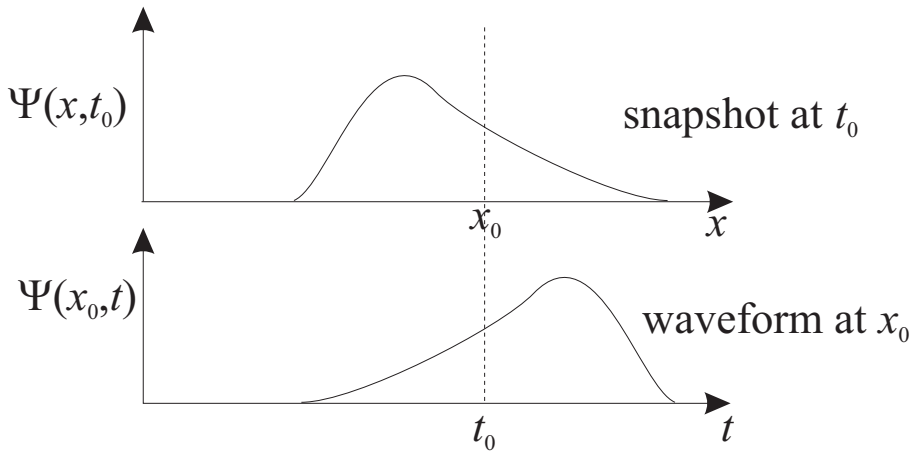


Figure 2.2: Waveforms and snapshots.

Combining Eqs. 2.3 and 2.5 we get the 1-d wave equation

$$\frac{\partial^2 \Psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \Psi}{\partial t^2}. \tag{2.6}$$

The wave (or phase) speed v is a fundamental quantity which characterises the wave and is determined by the physical properties of the wave medium. Clearly the same equation will result if the wave is travelling to the left.

There are two important features of this equation.

(a) It is general. Nothing has been specified about the type of wave motion or the medium through which the wave is propagating. It is relevant to a waveform of any shape.

(b) It is linear in Ψ, i.e. all the terms involving Ψ are raised to the first power only. Just as we have seen previously for simple oscillating systems, this leads to the principle of superposition. By substitution into the wave equation, it is easy to see that if Ψ_1 and Ψ_2 are both solutions to the wave equation, then $\Psi = \Psi_1 + \Psi_2$ (or any other linear combination of Ψ_1 and Ψ_2) is also a solution. The linearity of the wave equation means we can use Fourier analysis to examine the properties of waves, as we will see later.

2.2.1 Harmonic Waves

A useful special case of wave motion is that of a continuous sinusoidal wave, or *harmonic wave*. More general wave patterns can be made up as a combination of harmonic waves using Fourier analysis. In a harmonic wave, the displacement $\Psi(x,t)$ varies sinusoidally with time at any point x. The displacement at $x = 0$ will be given by

$$\Psi(0,t) = \Re\left(Ae^{i\omega t}\right)$$

where A is a complex number. The wave (travelling in the positive x direction) will have the general form $f(x - vt)$, hence elsewhere it must be given by

$$\Psi(x,t) = \Re\left(Ae^{i\omega(t-x/v)}\right) \tag{2.7}$$

$$= \Re\left(Ae^{i(\omega t - kx)}\right) \tag{2.8}$$

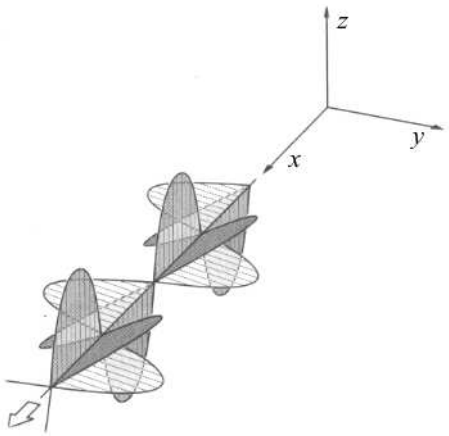


Figure 2.3: Linearly polarised wave (from Hecht, Optics).

where k is the wavenumber and is such that $\omega/k = v$, the wave or phase velocity. The quantity $(\omega t - kx)$ is the phase of the wave. The phase changes by 2π when x changes by λ, hence the wavenumber $k = 2\pi/\lambda$.

2.3 Polarisation

Waves on a stretched string are one example of *transverse* waves, where the displacement of the medium is perpendicular to the direction of motion of the wave. Clearly there are two orthogonal directions (y and z), along which the displacement can take place. The amplitude and relative phase of the displacement along these two directions define the *polarisation* of the wave. It turns out that displacements along orthogonal directions are independent, so any transverse wave can be described by the amplitude and phase of its components along these directions.

The simplest case is that of *linear* polarisation. This is achieved by oscillating the end of the string along a straight line, and each point on the string will oscillate along a parallel line as shown. Representing the y and z components of the displacement as

$$\begin{aligned} \Psi_y &= A_y \cos(\omega t - kx) \\ \Psi_z &= A_z \cos(\omega t - kx + \phi) \end{aligned}$$

we find that linear polarisation arises where $\phi = 0$ (or integer multiples of π). Clearly a wave polarised along the y (z) axis can be described by $A_z = 0$ ($A_y = 0$). Where both A_y and A_z are non-zero, the wave will be polarised along an intermediate direction (Figure 2.3), with an amplitude $A = \sqrt{A_y^2 + A_z^2}$ and an angle of polarisation $\theta = \tan^{-1} A_z/A_y$ to the y axis (Figure 2.4). Thus any linearly polarised wave can be resolved into two orthogonal linearly polarised components with the same phase.

A more interesting case occurs when the amplitudes of the two components are equal, but there is a phase difference of $\pi/2$ between them. This corresponds to *circular* polarisation, where the displacement at given value of x follows a circular path in the $y-z$ plane. The wave is said to be either right-circularly or left-circularly polarised, depending on the sense of rotation of the displacement vector in the $y-z$ plane. Looking *towards* the source of the wave, right-circular polarisation corresponds to clockwise motion of the electric field vector.

In the most general case, the amplitudes and relative phase of the two components are arbitrary. Here the displacement at any position follows an ellipse, with a size, shape and orientation that

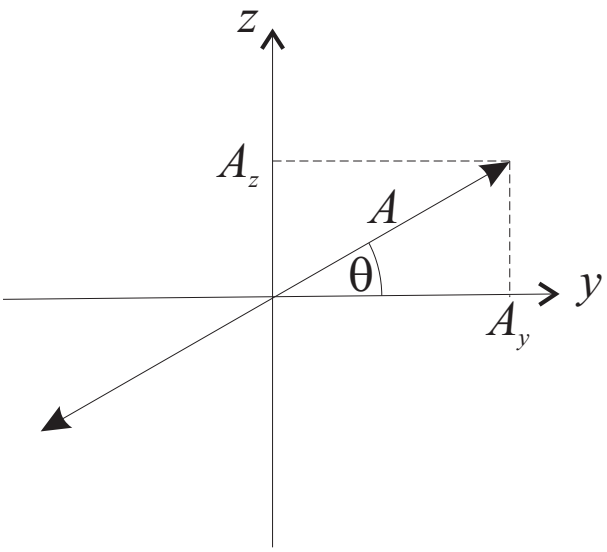


Figure 2.4: Displacement vector at fixed position for a linearly polarised wave.

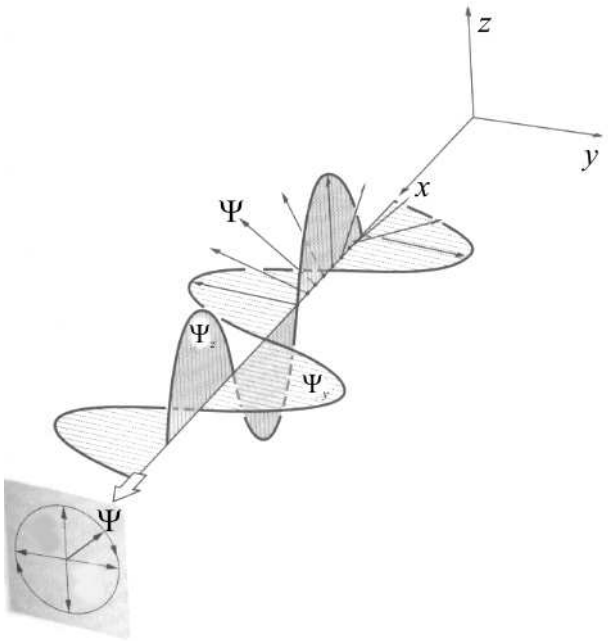


Figure 2.5: Circularly polarised wave (from Hecht, Optics).

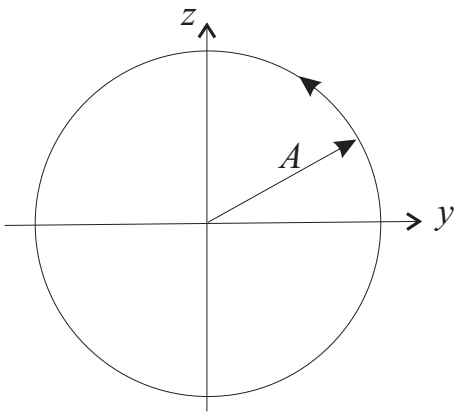


Figure 2.6: Displacement vector at fixed position for a circularly polarised wave.

depend on A_y, A_z and ϕ. Writing

$$\begin{aligned}\Psi_y &= A_y \cos(\omega t - kx)\\ \Psi_z &= A_z \cos(\omega t - kx + \phi)\\ &= A_z\left(\cos(\omega t - kx)\cos\phi - \sin(\omega t - kx)\sin\phi\right)\end{aligned}$$

and eliminating $\omega t - kx$ gives

$$\Psi_z = A_z\left(\frac{\Psi_y}{A_y}\cos\phi - \sqrt{1-\frac{\Psi_y^2}{A_y^2}}\sin\phi\right).$$

We can rearrange this into the standard equation for an ellipse

$$\frac{\Psi_y^2}{A_y^2} + \frac{\Psi_z^2}{A_z^2} - 2\frac{\Psi_y\Psi_z}{A_yA_z}\cos\phi = \sin^2\phi,$$

as shown in Figure 2.7, where

$$\tan 2\alpha = \frac{2A_yA_z\cos\phi}{A_y^2 - A_z^2}$$

Circular and linear polarisations are special cases of elliptical polarisation.

Polarisation is particularly important for electromagnetic waves, where the electric field direction defines the polarisation state. For example, when light is incident on an interface at an angle, the polarisation state of the light will affect the amplitude and phase of the reflection and transmission coefficients. When light passes through some *anisotropic* materials, the wave speed depends on the direction of polarisation with respect to the crystal axes; these materials can be used to manipulate the polarisation state of a light beam by changing the phase difference between the polarisation components.

2.4 Wave impedance

Waves transport energy away from the source of a disturbance; for example in the plucking of a violin string, a sonic boom from an aircraft, light from the sun, or radio waves from a transmitter. Clearly the way energy is transported and the amplitude of the wave resulting from the force creating the disturbance will be a function of the properties of the medium in which the wave is

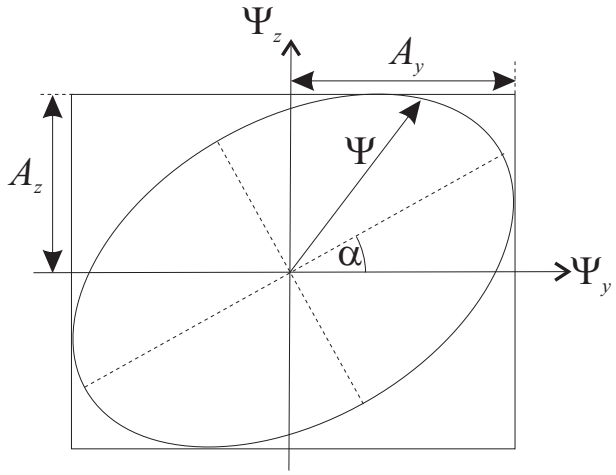


Figure 2.7: Elliptical polarisation.

Figure 2.8: Transverse force in a wave.

propagating. The concept of *wave impedance* is used to define the relationship between the force and the wave response.

$$\text{Impedance} = \frac{\text{driving force}}{\text{velocity response}}$$

The definition is an exact analogy of the idea of impedance for simple oscillating systems. It is very important to note that the velocity response is the rate of change of the displacement of the medium, *not* the speed of propagation of the wave through the medium. In the case of a wave on a string, for example, the velocity response is the transverse velocity of the elements of the string.

The power fed into a wave is given by the product of force and velocity response. If the impedance is real, the force and velocity are in phase, and the power input is maximised. If the impedance is purely imaginary (e.g. electromagnetic waves below a cut-off frequency in plasmas and waveguides), then no energy can be fed into and propagated through the medium.

We can calculate the impedance for waves on a stretched string by considering the free end of the string. As before, the transverse driving force F is given by

$$F = -T\sin\theta \approx -T\frac{\partial\Psi}{\partial x}.$$

The impedance Z is given by

$$Z = \frac{\text{transverse driving force}}{\text{transverse velocity}} = -\frac{T\partial\Psi/\partial x}{\partial\Psi/\partial t}.$$

We showed earlier in our derivation of the wave equation that for a wave travelling in the positive x direction

$$\frac{\partial\Psi}{\partial x} = \frac{df}{du} \text{ and } \frac{\partial\Psi}{\partial t} = -v\frac{df}{du}$$

so

$$\frac{\partial \Psi}{\partial x} = -\frac{1}{v}\frac{\partial \Psi}{\partial t}.$$

Thus, for a wave travelling the positive x direction

$$Z = \frac{T}{v}. \tag{2.9}$$

Since we know that the wave speed v is given by $v = \sqrt{T/\rho}$, we can rewrite this as

$$Z = \frac{T}{v} = \sqrt{T\rho} = \rho v \tag{2.10}$$

For a wave travelling in the negative x direction, the impedance is given by $-T/v$.

As you know from Part IA, energy is present in a wave in the form of kinetic and potential energy. We can use the ideas of impedance to determine the mean power transmitted by a wave. The power input into a wave is given by

$$\begin{aligned} \text{power input} &= \text{transverse force} \times \text{transverse velocity} \\ &= F\,u. \end{aligned}$$

(We now use u to represent the transverse velocity.) As we showed in Section 1.2 for an oscillator, the mean power input is given by $\langle P \rangle = \frac{1}{2}\Re[\mathbf{F}\mathbf{u}^*]$. Since $\mathbf{F} = Z\mathbf{u}$, we can rewrite this as

$$\begin{aligned} \langle P \rangle = \tfrac{1}{2}\Re[\mathbf{F}\mathbf{u}^*] &= \tfrac{1}{2}\Re[Z\mathbf{u}\mathbf{u}^*] \\ &= \tfrac{1}{2}\Re[Z]\,|\mathbf{u}|^2\,. \end{aligned} \tag{2.11}$$

Assuming Z is real and that

$$\mathbf{u} = \frac{\partial \mathbf{\Psi}}{\partial t} = i\omega A_0 e^{i(\omega t - kx)}$$

then

$$\text{mean power} = \tfrac{1}{2} Z\omega^2 A_0^2. \tag{2.12}$$

2.5 Reflection and transmission

Whenever a wave encounters a change in impedance in the medium through which it is travelling, some of its energy will be reflected, producing a backwards-travelling wave in addition to the transmitted wave. This happens for example when light is incident on a pane of glass, or when water waves suddenly encounter a shallow region. In this section we will show how to calculate the amplitudes of the reflected and transmitted waves for normal incidence onto a boundary. We will start with the example of waves on a string, but the results will be general to all waves.

Consider a wave incident on a boundary between two materials with impedances Z_1 and Z_2 (Figure 2.9). One boundary condition is obvious: the displacement in both materials must be identical at the interface. The second boundary condition is slightly less obvious: the transverse force must be continuous at the interface. If we consider an element of mass at the interface, the net force on it produces an acceleration, but in the limit as the mass tends to zero, the force must also tend to zero (otherwise the acceleration would be infinite). This means that the force must vary continuously across the interface. In the case of the stretched string, we have shown that the force is given by $T\partial\Psi/\partial x$. Hence $\partial\Psi/\partial x$ must be the same in both materials at the interface.

Let us write the incident, transmitted and reflected waves as
$A_1 \exp i(\omega t - k_1 x)$, $A_2 \exp i(\omega t - k_2 x)$ and $B_1 \exp i(\omega t + k_1 x)$.

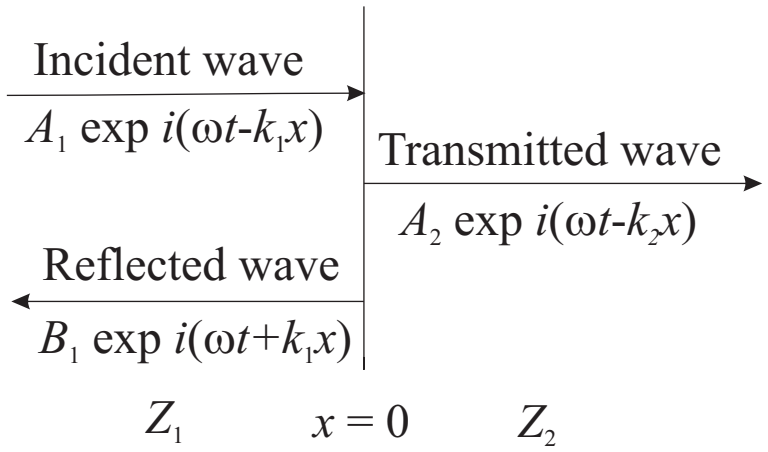


Figure 2.9: Incident, reflected and transmitted waves at an interface.

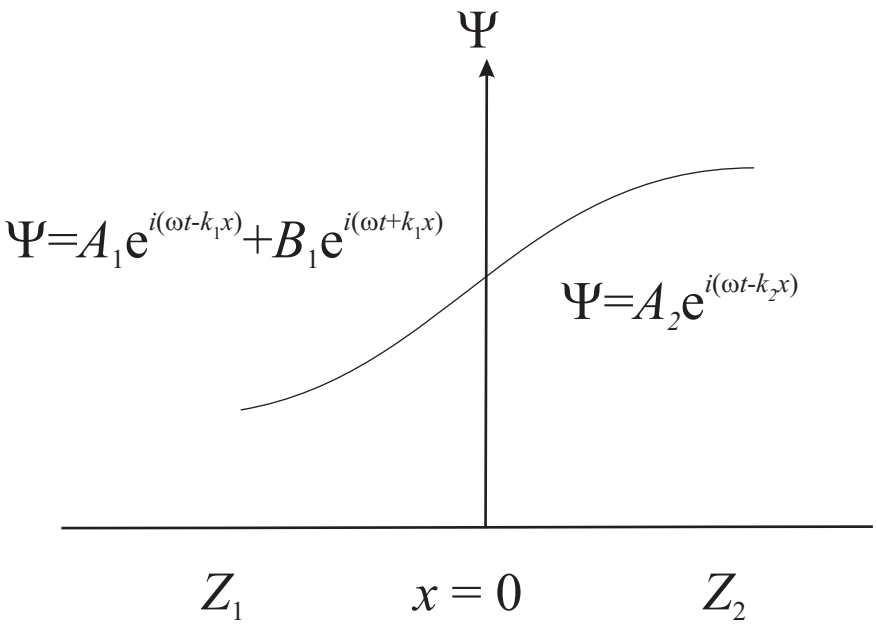


Figure 2.10: Boundary conditions at an interface.

Continuity of displacement, Ψ, at $x = 0$ gives

$$A_1 e^{i\omega t} + B_1 e^{i\omega t} = A_2 e^{i\omega t}.$$

Since this equation must be true at all times, the waves must all have the same frequency, which we have assumed anyway when defining the waves. Hence

$$A_1 + B_1 = A_2. \tag{2.13}$$

Continuity of transverse force gives

$$T(-ik_1)A_1 + T(ik_1)B_1 = T(-ik_2)A_2.$$

Now

$$k_1 = \omega/v_1 \text{ and } k_2 = \omega/v_2$$

so

$$-\frac{T}{v_1}\omega A_1 + \frac{T}{v_1}\omega B_1 = -\frac{T}{v_2}\omega A_2.$$

But

$$T/v_1 = Z_1 \text{ and } T/v_2 = Z_2$$

so

$$-Z_1 A_1 + Z_1 B_1 = -Z_2 A_2. \tag{2.14}$$

Equations 2.13 and 2.14 can be solved to give the ratios of A_1, B_1 and A_2. Multiply 2.13 by Z_2 and add to 2.14 to give

$$(Z_2 - Z_1)A_1 + (Z_2 + Z_1)B_1 = 0$$

$$\frac{B_1}{A_1} = \frac{Z_1 - Z_2}{Z_1 + Z_2} = \frac{\text{reflected amplitude}}{\text{incident amplitude}} = r. \tag{2.15}$$

This ratio, r, is known as the amplitude reflection coefficient.

Multiply 2.13 by Z_1 and subtract from 2.14 to give

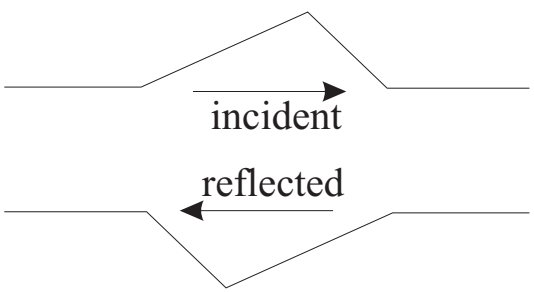
$$2Z_1 A_1 = (Z_1 + Z_2)A_2$$

$$\frac{A_2}{A_1} = \frac{2Z_1}{Z_1 + Z_2} = \frac{\text{transmitted amplitude}}{\text{incident amplitude}} = t. \tag{2.16}$$

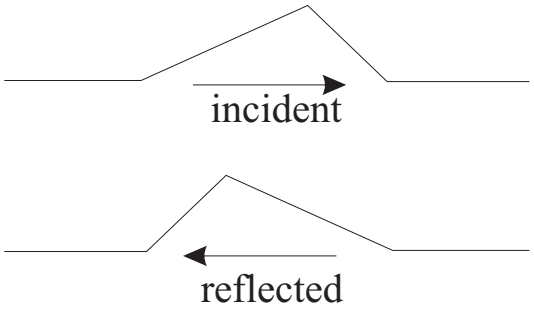
This ratio, t, is known as the amplitude transmission coefficient.

We can consider three special cases:

(a) $Z_2 = \infty$, i.e. the string is clamped at $x = 0$. In this case $r = -1$, giving inversion of the incident wave, i.e. a phase change of π on reflection. This is true whatever the form of the incident wave. There is no transmitted wave, i.e. $t = 0$.



(b) $Z_2 = 0$, i.e. the string has a free end at $x = 0$. (In fact, to maintain the tension T, it must be attached to a string with $\rho = 0$.) In this case $r = 1$, so the reflected wave is the same as the incident wave. Eq. 2.16 gives $t = 2$ (which explains the "flick" at the end of the string), but no energy is transmitted to the second string since $Z_2 = 0$, as we shall see later.



(c) $Z_2 = Z_1$, i.e. no impedance mismatch. In this case $r = 0$ and $t = 1$, so there is no reflection and the wave is totally transmitted. The simplest way of achieving this is to have identical strings, i.e. no boundary.

These results are quite general and apply to a wide variety of wave systems. At any impedance mismatch the basic procedure is to consider (i) continuity of "displacement", Ψ and (ii) continuity of "force" $Z\partial\Psi/\partial t$. For harmonic waves at normal incidence, condition (i) gives equation 2.13

$$A_1 + B_1 = A_2$$

and condition (ii) gives the equivalent of equation 2.14

$$Z_1 i\omega A_1 + (-Z_1) i\omega B_1 = Z_2 i\omega A_2$$

remembering that the impedance of a wave propagating in the negative x direction is negative.

Note that the amplitude reflection and transmission coefficients we have derived are for the "displacement" term. In some cases, we need expressions for the reflection and transmission coefficients for the "force" term (e.g. pressure in a sound wave, voltage in a transmission line or the electric field in an electromagnetic wave). In this case Z_1 and Z_2 are replaced by $1/Z_1$ and $1/Z_2$ respectively.

Note also that if Z_1 and/or Z_2 are complex then there are phase differences between the incident, reflected and transmitted waves.

2.5.1 Reflection and transmission of energy

We showed earlier (Eq. 2.12) that the rate at which energy is transmitted in a harmonic wave is given by

$$\tfrac{1}{2} Z \omega^2 A^2.$$

We can use this to find the energy reflected and transmitted at the boundary considered in the previous section

$$\begin{aligned} \text{rate of energy input} &= \tfrac{1}{2} Z_1 \omega^2 A_1^2 \\ \text{rate of energy reflection} &= \tfrac{1}{2} Z_1 \omega^2 B_1^2 \\ \text{rate of energy transmission} &= \tfrac{1}{2} Z_2 \omega^2 A_2^2 \end{aligned}$$

Thus

$$\frac{\text{reflected energy}}{\text{incident energy}} = \frac{Z_1 B_1^2}{Z_1 A_1^2} = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 = R \tag{2.17}$$

and

$$\frac{\text{transmitted energy}}{\text{incident energy}} = \frac{Z_2 A_2^2}{Z_1 A_1^2} = \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2} = T. \tag{2.18}$$

These ratios are known respectively as the power reflection and transmission coefficients.

Obviously these equations must be consistent with the conservation of energy, i.e.

$$\text{reflected power} + \text{transmitted power} = \text{incident power}$$

or equivalently

$$R + T = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 + \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2} = \left(\frac{Z_1 + Z_2}{Z_1 + Z_2}\right)^2 = 1.$$

For the special cases dealt with in Section 2.5 the results, as expected, are as follows:

(a) $Z_2 = \infty$, $R = 1$, $T = 0$, i.e. all energy reflected

(b) $Z_2 = 0$, $R = 1$, $T = 0$, i.e. all energy reflected

(c) $Z_2 = Z_1$, $R = 0$, $T = 1$, i.e. all energy transmitted.

To determine the power reflection coefficient when Z_1 and/or Z_2 are complex, we recall that the average power input is given by $\frac{1}{2}\Re(Z)\omega^2 A^2$, giving

$$\begin{aligned} R &= \frac{\frac{1}{2}\Re(Z_1)\omega^2 |B_1|^2}{\frac{1}{2}\Re(Z_1)\omega^2 |A_1|^2} \\ &= \frac{B_1}{A_1}\frac{B_1^*}{A_1^*} = rr^* = \left|\frac{Z_1 - Z_2}{Z_1 + Z_2}\right|^2. \end{aligned}$$

2.5.2 Impedance Matching

Very often it is necessary to transmit waves from one medium to another, or to extract energy at the boundary of the medium in which the wave is travelling. To obtain as high a level of energy transmission as possible it is essential to minimise reflection at the boundary, i.e. as far as possible in designing the system the impedances should be matched. This is particularly important when designing electrical transmission lines; this will be discussed in more detail in the Electromagnetism course. Another example arises in optics, where we wish to minimise the reflection when light is goes from one medium to another, for example from air to glass. In that case, we cannot match the impedances exactly, but there are tricks we can play by using an intervening layer. This is known as a $\lambda/4$ coupler, and is shown in Figure 2.11.

The wave travelling to the left in medium 1 is the sum of Ψ_r, the wave reflected at the 12 interface, and Ψ_{trt}, that part of the wave transmitted at 12 which is then reflected at 23 and transmitted at 21. The wave travelling to the left in medium 1 can be made to have zero amplitude if Ψ_r and Ψ_{trt} interfere destructively. This occurs when they have equal amplitudes and are π out of phase. It is easy to see that the phase condition is satisfied when $l = \lambda/4$, where λ is the wavelength in medium 2. The condition on the magnitude of Z_2 is harder to derive, but turns out to be $Z_2 = \sqrt{Z_1 Z_3}$. (Note that in order to derive this condition, we cannot simply use the expressions for r and t at the first interface which we derived above. The reason for this is that 4, not 3 waves are

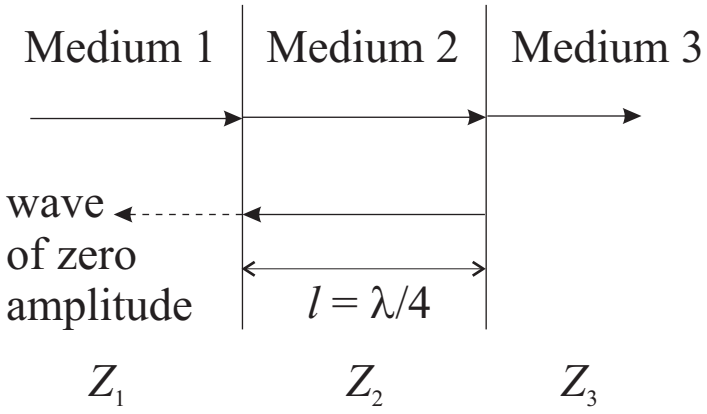


Figure 2.11: Impedance matching.

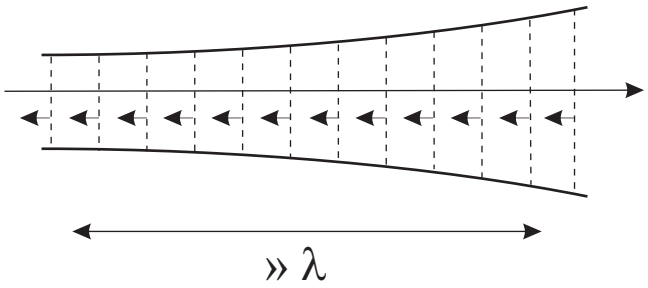


Figure 2.12: Gradual impedance change.

involved, and we must go back to the continuity equations for force and displacement to solve the boundary condition in the presence of all 4 waves.)

This method of impedance matching is used on camera lenses, which are coated with a $\lambda/4$ layer to avoid reflections, in a technique known as "blooming". Note that total removal of reflection can only work for one wavelength, and others will not be completely out of phase. This gives camera lenses their characteristic purple colour because the wavelength for minimum reflection is chosen to be green, so we see reflection of some red and blue light.

Another technique for matching impedances involves a gradual impedance change from one medium to another (Figure 2.12). At a sharp boundary (with a transition width $\ll \lambda$), there will be substantial reflection, with a magnitude that depends on $Z_1 - Z_2$, i.e. the difference in impedance. If the transition takes place gradually (over a distance $\gg \lambda$), infinitesimal reflections occur at each infinitesimal change. The total reflected wave Ψ_r is the sum of all these infinitesimal reflected waves. Since the reflected waves are generated over a distance $\gg \lambda$, these waves will have a wide range of phases. Adding up these waves will produce a resultant wave with a very small amplitude, consequently most of the energy is transmitted. Examples of this technique for impedance matching can be found in acoustics (sound waves), such as

(i) The design of the horns in old-fashioned gramophones, intended to provide impedance matching between the needle and the free air

(ii) The bell of a clarinet which, as with the horn in (i) has a gradual change in shape, increasing in width.

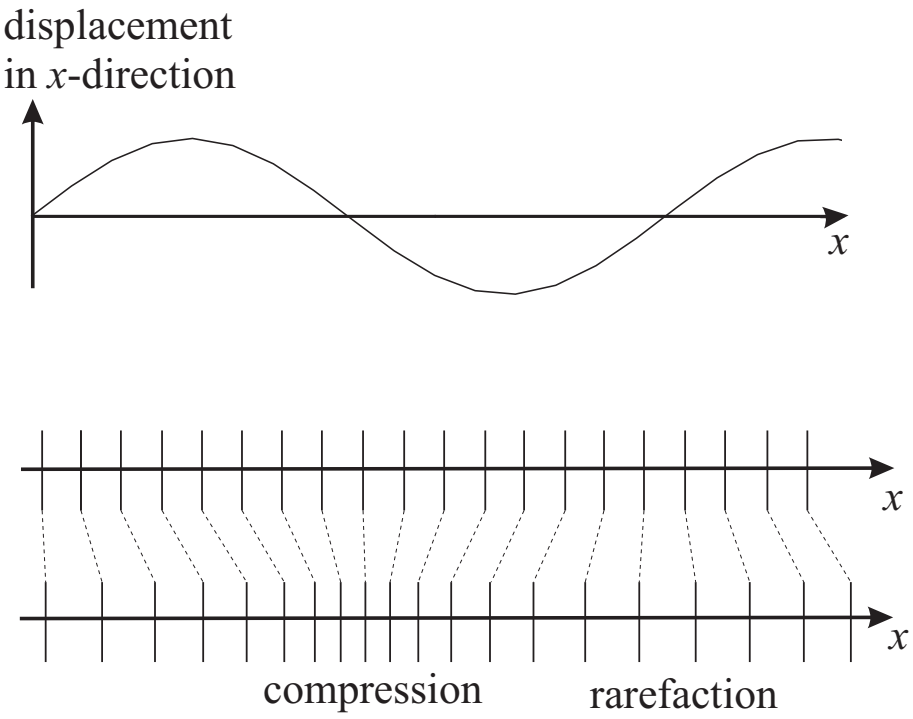


Figure 2.13: Displacement of gas in a longitudinal wave.

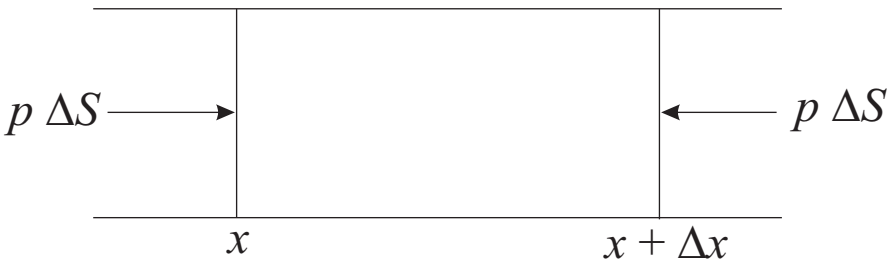
2.6 Longitudinal Waves

Sound waves are one important example of *longitudinal waves*, where the displacement of the medium is in the same direction as the direction of wave propagation. This contrasts with *transverse waves*, where the displacement is perpendicular to the direction of propagation. Unlike transverse waves, the direction of displacement is uniquely defined, and hence the concept of polarisation does not arise.

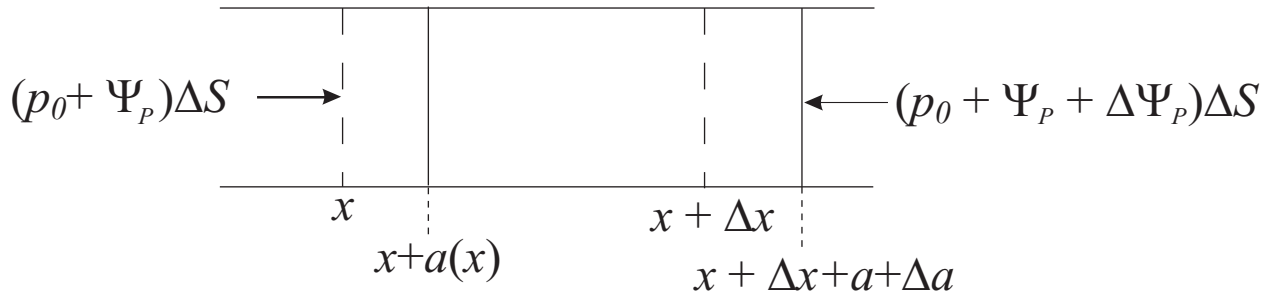
Transmission of a longitudinal wave occurs by compression and rarefaction of the medium, as the particles move together and apart. This corresponds to a change in pressure, with the maximum pressure corresponding to the minimum displacement, as shown in Figure 2.13.

2.6.1 Sound waves in a gas

To derive the wave equation and speed of propagation for sound waves in a gas, we take the same general approach as in Section 2.1, again assuming small displacements. We consider the element of a column of gas of cross-sectional area ΔS which, when the gas is in equilibrium at pressure p and no wave is propagating, lies between x and $x + \Delta x$.



The displacement of the wave is described by $a(x,t)$ and the additional pressure caused by the presence of the wave is denoted $\Psi_p(x,t)$. In the presence of the wave, the plane originally at x is displaced by $a(x)$ to $x + a(x)$.



The plane originally at $x + \Delta x$ is displaced to $x + \Delta x + a + \Delta a = x + \Delta x + a(x) + \frac{\partial a}{\partial x}\Delta x$. The volume change of the element is given by

$$\Delta S \Delta a = \Delta S \frac{\partial a}{\partial x}\Delta x.$$

The fractional change in volume is

$$\begin{aligned} \frac{\Delta V}{V} &= \frac{\Delta S \frac{\partial a}{\partial x}\Delta x}{\Delta S \Delta x} \\ &= \frac{\partial a}{\partial x}. \end{aligned} \tag{2.19}$$

We now treat the pressure in the same way: the net force on the element in the x direction due to the pressure imbalance between the two ends is given by

$$\begin{aligned} F_{\text{net}} &= (p+\Psi_p)\Delta S - (p+\Psi_P+\Delta\Psi_p)\Delta S \\ &= (p+\Psi_p)\Delta S - \left(p+\Psi_p+\frac{\partial \Psi_p}{\partial x}\Delta x\right)\Delta S \\ &= -\frac{\partial \Psi_p}{\partial x}\Delta x \Delta S \end{aligned} \tag{2.20}$$

i.e. it depends only on the pressure gradient.

Now we need to work out the relationship between changes in pressure and changes in volume for the element. In a sound wave the pressure changes occur very rapidly and there is essentially no time for any heat exchange with the surroundings, i.e. all the work done in compression goes into heating up the gas. The process thus occurs adiabatically and the relevant equation of state for the gas is

$$pV^\gamma = \text{constant}. \tag{2.21}$$

You will remember from last year that $\gamma = C_p/C_v$ and is 1.4 for air.

Differentiation gives

$$dp\, V^\gamma + p\gamma V^{\gamma-1} dV = 0$$

so

$$dp = -\gamma p \frac{\Delta V}{V}.$$

We can see that dp is simply the pressure change associated with the passage of the wave, i.e. Ψ_p. Combining with Eq. 2.19 gives

$$\Psi_p = -\gamma p \frac{\Delta V}{V} = -\gamma p \frac{\partial a}{\partial x}. \tag{2.22}$$

Differentiating with respect to x gives

$$\frac{\partial \Psi_p}{\partial x} = -\gamma p \frac{\partial^2 a}{\partial x^2} - \gamma \frac{\partial p}{\partial x}\frac{\partial a}{\partial x}.$$

The second term on the right-hand side is negligible in most practical situations i.e. when $a \ll \lambda$. From Eq. 2.20, the force on the element is thus

$$\gamma p \frac{\partial^2 a}{\partial x^2} \Delta x \Delta S.$$

Applying Newton's second law to the element, which has mass $\rho \Delta x \Delta S$ where ρ is the equilibrium density of the gas, gives

$$\gamma p \frac{\partial^2 a}{\partial x^2} \Delta x \Delta S = \rho \Delta x \Delta S \frac{\partial^2 a}{\partial t^2}.$$

Hence

$$\frac{\partial^2 a}{\partial x^2} = \frac{\rho}{\gamma p}\frac{\partial^2 a}{\partial t^2}$$

i.e. the wave equation, with wave speed v given by

$$v = \sqrt{\frac{\gamma p}{\rho}} = \sqrt{\frac{\gamma RT}{m}} \tag{2.23}$$

where m is the molar mass.

This expression implies that the speed of sound at a given temperature depends on the molar mass, and hence will be larger for helium, say, than for air. Different gases have different value of γ as well but these are likely to have a much smaller effect on v. The speed of sound in air at standard temperature and pressure is $340\,\mathrm{m\,s^{-1}}$.

You will remember that the r.m.s. speed of molecules in a gas is given by

$$\sqrt{< v^2 >} = \sqrt{\frac{3RT}{m}}.$$

Hence the r.m.s. molecular speeds will be close to, but slightly larger than the wave speed. Remember, though, that the passage of a sound wave does not require the net movement of gas molecules from one point to another.

We have derived the equation for the displacement of the gas when a sound wave passes through it. However, this is a difficult quantity to measure directly, and it is usually the pressure variations which are of greater importance. We know that the excess pressure is given by

$$\Psi_p = -\gamma p \frac{\partial a}{\partial x}.$$

If we assume a harmonic wave of the form

$$a = a_0 \exp i(\omega t - kx)$$

then

$$\Psi_p = i\gamma p k a \tag{2.24}$$

i.e. the pressure leads the displacement by $\pi/2$ and has amplitude $\gamma p k a_0$.

2.6.2 Sound waves in solids and liquids

Longitudinal waves can also be supported in liquids and solids, and their analysis is very similar to the derivation of the wave equation for a gas. The properties of the medium are described a generalisation of Eq. 2.22, the relationship between the pressure due to the wave and the strain in the medium

$$\Psi_p = -K\frac{\partial a}{\partial x}$$

where K is the relevant modulus. This gives

$$v = \sqrt{\frac{K}{\rho}} \tag{2.25}$$

where ρ is the density. For gases and liquids, the pressure is isotropic, but the expansion and rarefaction takes place only in the direction of passage of the wave. Hence the fractional volume change is directly proportional to the fractional change in length of the element ($\Delta V/V = \Delta a/\Delta x = \partial a/\partial x$). The relevant modulus is the *bulk modulus*, B, where

$$dp = -B\frac{dV}{V}$$

Solids are rather more complicated, since they can support shear stresses. For passage of a wave through a thin solid bar, the relevant modulus is *Young's modulus*, Y, the ratio of longitudinal stress (pressure) in the bar to its longitudinal strain ($\partial a/\partial x$).

$$Y = \frac{\text{stress}}{\text{strain}} = \frac{dp}{\partial a/\partial x}.$$

When the bar is compressed longitudinally it can expand in the transverse direction (the ratio of the two strains is known as *Poisson's ratio*). However, in a bulk solid, the medium cannot expand sideways, hence more longitudinal pressure is required to produce a longitudinal strain, and the modulus in therefore larger.

2.7 Waves in 3 Dimensions

Electromagnetic waves are an example of waves which can propagate in any direction through 3-dimensional space. To define a wave in three dimensions, we need to specify the shape and orientation of the wavefronts. The simplest example is the *plane wave*, where the wavefronts are planar, and in principle infinite in extent (Figure 2.14). A harmonic plane wave can be written

$$\Psi(\mathbf{r}, t) = Ae^{i(\omega t - \mathbf{k}.\mathbf{r})}$$

where $\mathbf{k}$ is the wavevector which defines the direction of motion of the wave, and is perpendicular to the wavefronts. The modulus of $\mathbf{k}$ is related to the wavelength via $|\mathbf{k}| = 2\pi/\lambda$. We can split up $\mathbf{k}$ into its Cartesian components $\mathbf{k} = (k_x, k_y, k_z)$, giving

$$\Psi(x, y, z, t) = Ae^{i(\omega t - (k_x x + k_y y + k_z z))}.$$

In 3 dimensions, the wave equation becomes

$$\frac{\partial^2\Psi}{dx^2} + \frac{\partial^2\Psi}{dy^2} + \frac{\partial^2\Psi}{dz^2} = \nabla^2\Psi = \frac{1}{v^2}\frac{\partial^2\Psi}{dt^2}. \tag{2.26}$$

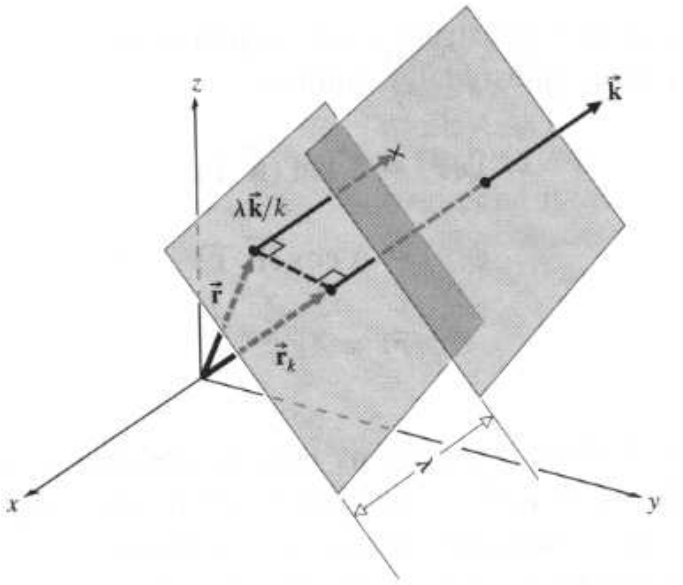


Figure 2.14: A plane wave (from Hecht, Optics)

Another special case of a three-dimensional wave is the spherical wave made by a point disturbance at the origin ($\mathbf{r} = 0$).

$$\Psi(\mathbf{r}, t) = \frac{A}{r} e^{i(\omega t - k|\mathbf{r}|)}.$$

The wavefronts are now spheres surrounding the origin. The intensity (proportional to amplitude squared) decays as r^{-2}, thus conserving energy as the wave propagates away from the source.

2.8 Standing Waves

A standing wave is a wave confined within a region of space by applying boundary conditions, e.g. a wave on a violin string, light in a laser cavity, sound waves in a crystal, or electrons in a solid. The analysis of standing waves was covered in detail last year; here we will give a summary of the main results. A travelling wave within the medium will undergo reflections at the boundaries to generate waves with different k-vectors, and the standing wave can be found as a superposition of these waves. The simplest case to analyse is where perfect reflection occurs at the boundary.

In one dimension, we can consider a wave confined on a string of length b. The reflection coefficient at the fixed ends is $r = -1$, hence the boundary conditions are $\Psi(0, t) = \Psi(b, t) = 0$. Superimposing counter-propagating waves gives

$$\begin{aligned} \Psi &= A\cos(\omega t - kx) - A\cos(\omega t + kx) \\ &= 2A \sin \omega t \sin kx. \end{aligned}$$

The boundary conditions are satisfied if $kb = n\pi$, where $n = 1, 2, 3 \ldots$, i.e.

$$k = \frac{n\pi}{b}.$$

Only a finite number of values for k are allowed, and the number of nodes in the standing wave pattern increases with frequency. For a wave speed v, the allowed frequencies are given by

$$\omega = \frac{nv\pi}{b}.$$

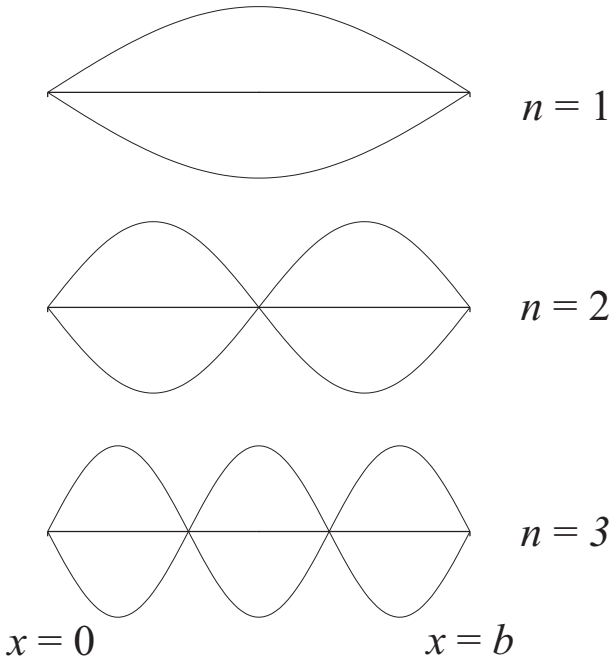


Figure 2.15: Standing waves on a string.

As you saw last year, it is easy to extend these ideas into two or three dimensions. This can be done either by superimposing reflected waves as above, or by looking directly for solutions of Eq. 2.26 which satisfy the boundary conditions. For waves confined in a 3-d box, for example, the wavefunctions are given by

$$\Psi = A \sin k_x x \sin k_y y \sin k_z z \cos \omega t.$$

For a box of dimensions a, b, c, the boundary conditions are

$$\Psi(0, y, z, t) = \Psi(a, y, z, t) = 0,$$

$$\Psi(x, 0, z, t) = \Psi(x, b, z, t) = 0,$$

$$\Psi(x, y, 0, t) = \Psi(x, y, c, t) = 0.$$

The wavevector is therefore quantised according to

$$k_x = \frac{l\pi}{a},\ k_y = \frac{m\pi}{b},\ k_z = \frac{n\pi}{c}.$$

where $l,\ m,\ n = 1, 2, 3 \ldots$.

2.9 Dispersive Waves

So far, we have assumed that the speed of waves is independent of their frequency. This is true for electromagnetic waves in a vacuum, and for small-amplitude waves on a stretched string. In many important cases, though, the wave speed (phase velocity $= \omega/k$) depends on the frequency of the wave. These waves are known as *dispersive waves*. Examples include surface waves in deep water and optical waves in a medium such as glass. The relationship between ω and k is known as the dispersion relation, as shown in Figure 2.16

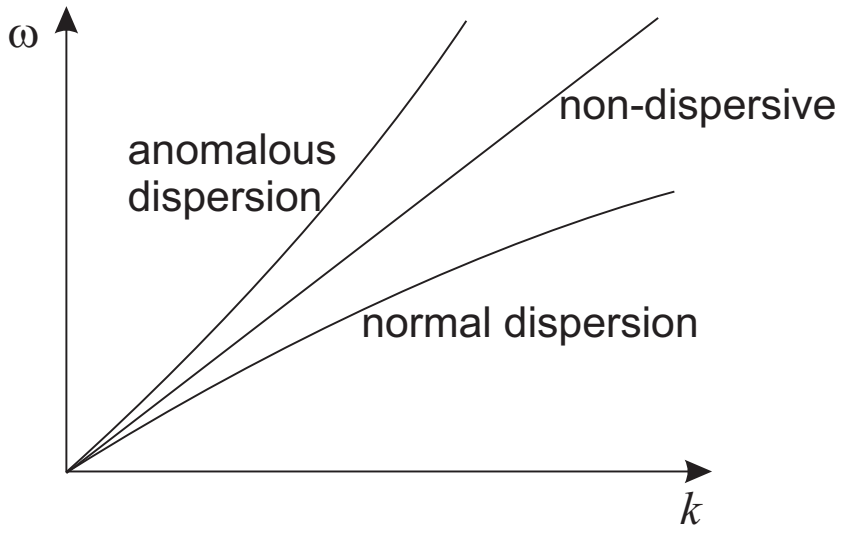


Figure 2.16: Dispersion relation for "normal" and "anomalous" dispersion.

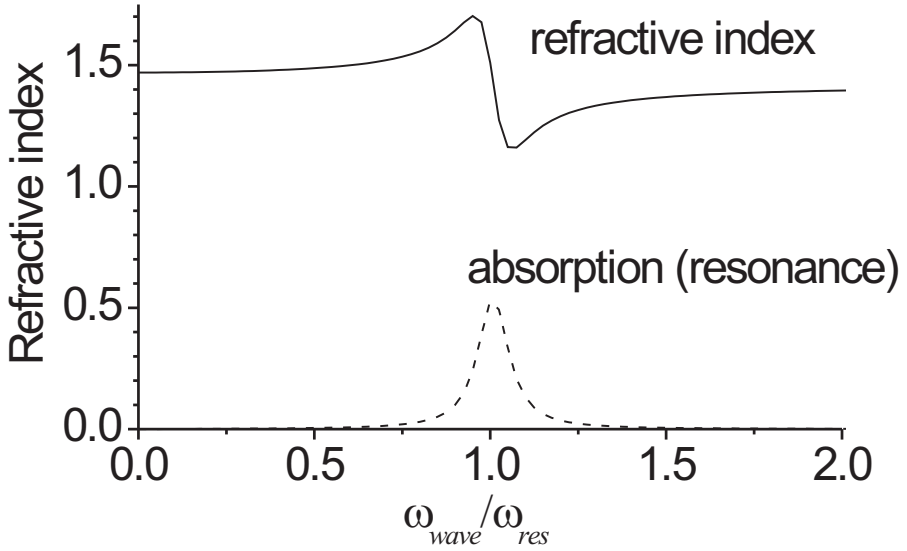


Figure 2.17: Refractive index variation with frequency (ω_{wave}) in the region of a resonance at ω_{res}.

One obvious example where the wave speed depends on frequency is in the passage of light through glass. The wave speed is measured by the refractive index, n, where $c_{medium} = c/n_{medium}$, and n is larger for blue light than for red light. When light passes through a prism it is deviated through an angle that depends on n (recall Snell's law $n_1 \sin\theta_1 = n_2 \sin\theta_2$). Blue light will therefore be deviated though a larger angle than red light, thus splitting a white light beam up into its characteristic spectral components.

The physical origin of dispersion depends on the details of the wave motion, but in many cases (such as for light in a medium), dispersion arises because the medium has a resonance at a particular frequency which can be excited by the passage of the wave. The wave speed depends on the strength of response of the medium to a displacement, and hence will depend on how close the wave frequency is to the resonance frequency. The change of refractive index with frequency in the region of a resonance is shown in Figure 2.17.

Although waves on a string are non-dispersive, waves in systems of connected masses do show dispersion. A very important example of this is in the passage of displacement waves (*phonons*)

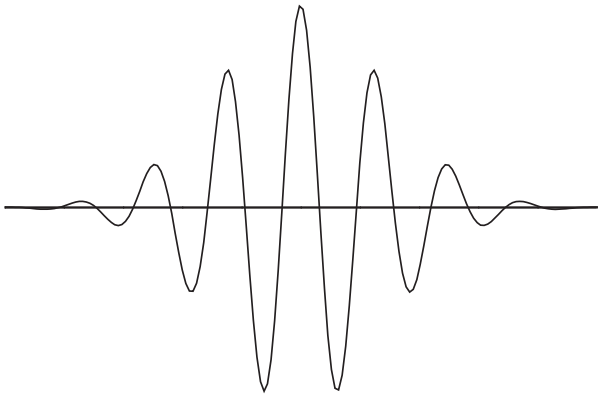


Figure 2.18: A wave group.

through solids, where the atoms act as individual masses connected by springs (the chemical bonds). You might think that spacing between atoms is so small that the solid will always act as a continuum, but for high frequencies (corresponding to the frequencies of infrared light), dispersion is very noticeable. This is an important topic in solid state physics, and will be covered in more detail in Statistical Physics.

2.9.1 Wave Groups

A harmonic wave which is infinite in extent and constant in magnitude will contain purely one frequency, and its propagation will be determined by a single phase velocity. Although this wave can carry energy, it is inherently featureless, and can carry no information. All real waves have a beginning and an end, i.e. their amplitude varies with time. For example, information is transmitted through optical fibres as a series of pulses, each representing a *wave packet*. To make up a wave packet, we need to superimpose waves with a range of frequencies (and hence a range of wavenumbers (Δk)), a *wave group*. Fourier analysis shows that any waveform can be made by superimposing a range of harmonic waves. For the pulse shown below, the shorter the pulse, the larger the range of individual waves that must be superimposed to make it up.

In a dispersive medium, the individual components in a wave group will not travel at the same speed. They will therefore spread out in space and time, hence the origin of the term *dispersion*. This will cause the shape of the wavepacket to change with propagation, and in general the wavepacket will spread out and eventually disappear. Dispersion also means that the wave group (to be specific, the point of maximum amplitude in the wave group) will move through the medium at a speed which is different from the phase velocity. Since this *group velocity* is the speed at which energy moves through the medium it is an important parameter to understand and derive.

2.9.2 Group velocity

Consider a wave packet such as that shown in Fig. 2.18, which can be described as a sinusoidal *carrier wave* with wavenumber k_0 multiplied by an envelope $f(x)$ so that the wave at time $t = 0$ is described by

$$\Psi(x,0) = \Re\left[f(x)e^{ik_0x}\right].$$

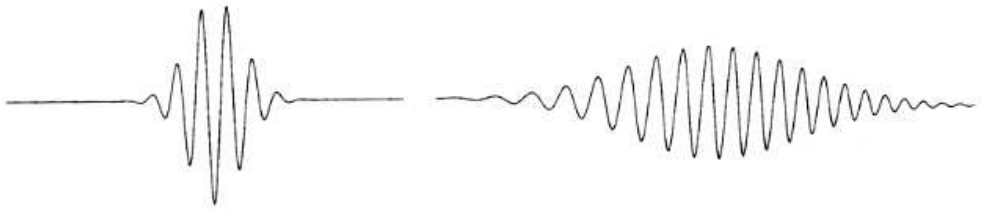


Figure 2.19: Wave group before and after passage through a dispersive medium.

For example, in Fig. 2.18 $f(x)$ would be a Gaussian i.e. purely real, but $f(x)$ can in general be complex, so that it modulates the amplitude and/or phase of the carrier wave. We write the envelope in terms of its component sinusoids

$$\begin{aligned}\Psi(x,0) &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{ik_1x}dk_1e^{ik_0x}\right] \\ &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{i(k_0+k_1)x}dk_1\right]\end{aligned}$$

where $F(k_1)$ is the Fourier transform of $f(x)$ (we have ignored factors of $\sqrt{2\pi}$ here to simplify the algebra, as these factors cancel out at the end). Each sinusoid with wavenumber k_0+k_1 will propagate at its own phase velocity, so we can write the waveform at some later time t as

$$\begin{aligned}\Psi(x,t) &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{i[(k_0+k_1)x-(\omega_0+\omega_1)t]}dk_1\right] \\ &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{i(k_1x-\omega_1t)}dk_1e^{i(k_0x-\omega_0t)}\right]\end{aligned} \tag{2.27}$$

where $\omega_0=\omega(k_0)$ and $\omega_0+\omega_1=\omega(k_0+k_1)$ for a dispersion relation given by $\omega=\omega(k)$.

We now make the assumption that $F(k_1)$ has significant amplitude only over a small wavenumber range $\pm\Delta k$ where $\Delta k \ll k_0$, so that we only need to consider the group of propagating waves in the band $k_0\pm\Delta k$. In other words, we have assumed that $f(x)$ is many carrier wavelengths across. With this assumption, we can Taylor-expand the dispersion relation about k_0 to give

$$\omega=\omega_0+\left.\frac{\partial\omega}{\partial k}\right|_{k_0}k_1+\frac{1}{2}\left.\frac{\partial^2\omega}{\partial k^2}\right|_{k_0}k_1^2+\cdots \tag{2.28}$$

For small k_1 we can drop the second order and higher terms and write

$$\omega_1\approx v_gk_1 \tag{2.29}$$

where we have defined the *group velocity* v_g as

$$v_g=\left.\frac{\partial\omega}{\partial k}\right|_{k_0} \tag{2.30}$$

which can be compared with the phase velocity

$$v_p=\frac{\omega}{k}. \tag{2.31}$$

Substituting Eq. 2.29 into Eq. 2.27 we have

$$\begin{aligned}\Psi(x,t) &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{ik_1(x-v_g t)}dk_1 e^{i(k_0 x-\omega_0 t)}\right] \\ &= \Re\left[\int_{-\infty}^{\infty} F(k_1)e^{-ik_1 v_g t}e^{ik_1 x}dk_1 e^{i(k_0 x-\omega_0 t)}\right]\end{aligned}$$

Recognising the k_1 integral as an inverse Fourier transform, we can make use of the convolution theorem and the fact that the Fourier transform of a delta-function is a complex exponential to yield

$$\begin{aligned}\Psi(x,t) &= \Re\left[f(x) * \delta(x-v_g t)e^{i(k_0 x-\omega_0 t)}\right] \\ &= \Re\left[f(x-v_g t)e^{ik_0(x-v_p t)}\right] \qquad (2.32)\end{aligned}$$

where $v_p = \omega_0/k_0$. Equation 2.32 says that after a time t the carrier wave $e^{ik_0 x}$ has propagated a distance $v_p t$ while the modulating envelope $f(x)$ has propagated a distance $v_g t$. For a non-dispersive wave $\omega/k = \frac{\partial\omega}{\partial k}$ for all frequencies, and so the envelope will propagate at the same speed as the carrier wave, but for a dispersive wave the two velocities can be different, and so the wave crests will move relative to the envelope. Typically we use the envelope as a tracer of the information carried by the wave packet, so the group velocity will be the speed of interest when looking at the propagation of information.

Note that the above result is restricted to wave groups consisting of sinusoidal components with a narrow spread of frequencies. A broad-band signal can be thought of as consisting of a number of narrow-band wave groups, and if the group velocity is different between these groups (i.e. the second-order term in the expansion in Eq. 2.28 is non-negligible) then the different wave groups will begin to spread apart as time progresses and the envelope will become distorted.

Another way of deriving the group velocity is to consider the speed at which the point of maximum amplitude in the envelope moves. At time t, this corresponds to the point x where all the components making up the wave have the same phase, and hence interfere constructively. The phase of any one component is given by $(\omega t - kx + \phi)$. Thus, at the point of maximum amplitude, $(\omega t - kx + \phi)$ is constant for all the wave components, i.e. it is independent of ω.

Hence

$$\frac{d}{d\omega}(\omega t - kx + \phi) = 0$$

so

$$t - \left(\frac{dk}{d\omega}\right)_{\omega_0} x = 0.$$

This is satisfied if

$$\frac{x}{t} = \left(\frac{d\omega}{dk}\right)_{\omega_0} = v_g(\omega_0) \qquad (2.33)$$

where ω_0 is the average or dominant frequency of the group.

2.10 Guided Waves

Very often we want to transmit energy from one localised point to another using a wave. Point sources typically emit over a range of angles, and hence the amplitude in any one direction will fall off with distance, making the transfer of energy inefficient. To send energy more directly from

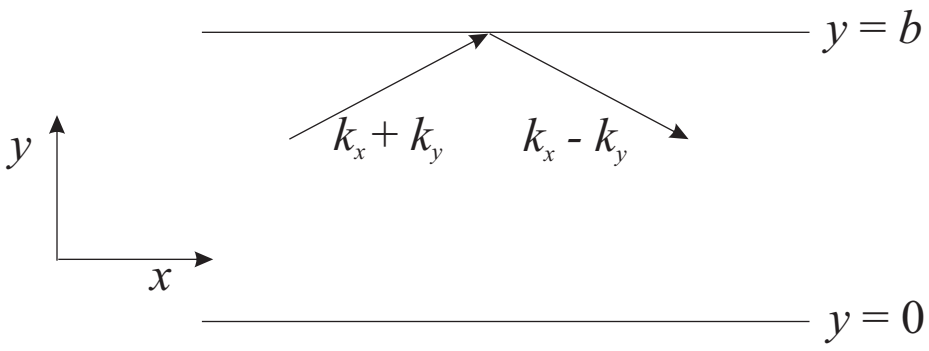


Figure 2.20: Two-dimensional waveguide.

one point to another it is necessary to confine the wave in a *waveguide*. In a waveguide, the wave is guided through some container or enclosure through which it can propagate with little or no loss in amplitude. Waveguides formed from rectangular metal tubes can be used to guide microwave radiation, where the dimension of the tube is comparable with the wavelength of the radiation (centimetres for microwaves). Other examples of waveguides include optical fibres (for optical waves), which we will discuss in more detail later, and stethoscopes (for acoustic waves).

To study the physics of waveguides, we will consider a simple example of guided waves on a two-dimensional elastic membrane under tension T (per unit length) which is clamped along two opposite edges a distance b apart. A wave propagating at an arbitrary angle to the x direction will undergo multiple reflections at the edges with a phase change of π at each, to give two waves

$$\begin{aligned} \Psi_A &= A \exp i(\omega t - k_x x - k_y y) \\ \Psi_B &= -A \exp i(\omega t - k_x x + k_y y). \end{aligned}$$

The resultant is given by

$$\begin{aligned} \Psi &= \Psi_A + \Psi_B \\ &= A \exp i(\omega t - k_x x)(\exp(-ik_y y) - \exp(ik_y y)) \\ &= -2iA \sin(k_y y) \exp i(\omega t - k_x x). \end{aligned}$$

This resultant is a wave travelling in the x direction whose amplitude is modulated by a standing wave in the y direction.

The boundary conditions require that $\Psi = 0$ at $y = 0$ and $y = b$, and this gives

$$\sin k_y b = 0$$

i.e.

$$k_y = \frac{n\pi}{b}$$

where n is an integer. Thus only discrete values of k_y are allowed.

To determine the dispersion relation we need to find the value of k_x (the wavevector for propagation) and hence derive the wave speed ω/k_x for waves travelling along the guide. Now

$$k^2 = k_x^2 + k_y^2 = \frac{\omega^2}{v^2}$$

where k is the wavevector and v is the wave speed for unguided waves on the membrane, i.e. $v^2 = \omega^2/k^2 = T/\rho$, ρ being the mass per unit area. Thus

$$\begin{aligned} k_x^2 &= k^2 - \frac{n^2\pi^2}{b^2} \\ &= \frac{\omega^2}{v^2} - \frac{n^2\pi^2}{b^2} \end{aligned} \tag{2.34}$$

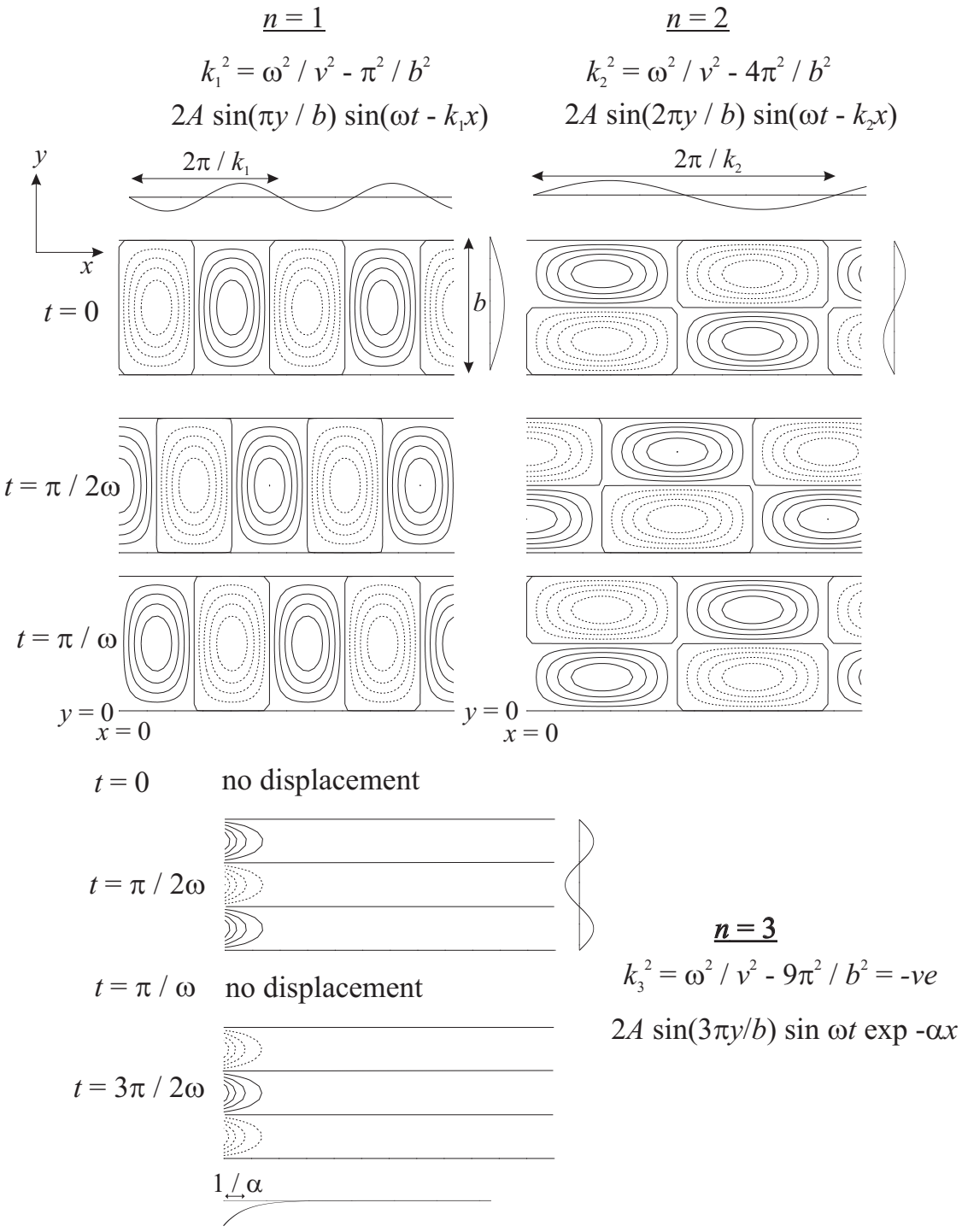


Figure 2.21: Modes in a waveguide.

$$\omega^2 = v^2\left(k_x^2 + \frac{n^2\pi^2}{b^2}\right). \tag{2.35}$$

The speed (phase velocity) of the waves travelling along the guide is given by

$$v_p = \frac{\omega}{k_x} = \frac{\omega}{\sqrt{\left(\frac{\omega^2}{v^2} - \frac{n^2\pi^2}{b^2}\right)}} \tag{2.36}$$

Thus v_p is a function of ω and the waves travelling along the guide are therefore dispersive. (Note that the unguided waves are non-dispersive.) The group velocity $d\omega/dk_x$ for these waves can be obtained by differentiating Eq. 2.35 as follows:

$$2\omega\frac{d\omega}{dk_x} = 2v^2k_x$$

so

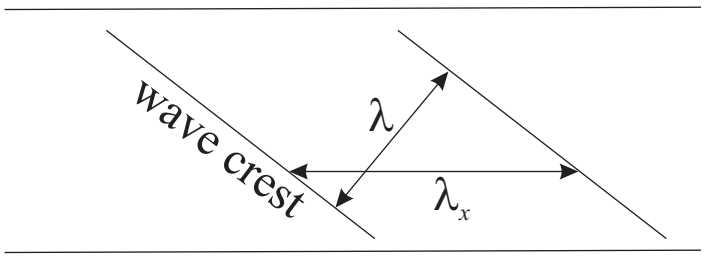
$$\begin{aligned} v_g &= \frac{d\omega}{dk_x} \\ &= v^2\frac{k_x}{\omega} \\ &= \frac{v^2}{\omega}\sqrt{\left(\frac{\omega^2}{v^2} - \frac{n^2\pi^2}{b^2}\right)}. \end{aligned} \tag{2.37}$$

2.10.1 Properties of the guided waves

There are a number of important features of these guided waves.

(i) There is a whole series of permitted modes in the guide corresponding to different value of n, with dispersion relations as shown in Figure 2.22.

(ii) Obviously $k_x < k$, so the wavelength of the guided wave, λ_x, is longer than that of the unguided wave, λ.



(iii) Consequently, the phase velocity, v_p is greater than v, the wave speed of the unguided waves (see Eq. 2.36). As $v_g v_p = v^2$ the group velocity v_g is less than v.

(iv) As $k_x \to 0$, $v_p \to \infty$ and $v_g \to 0$ (i.e. $d\omega/dk_x = 0$ at $k_x = 0$). The fact that $v_p \to \infty$ does not violate special relativity since energy and information are transmitted at v_g.

(v) In the limit, as k_x and ω get very large, the behaviour approaches that of an unguided wave, with v_p and $v_g \to v$.

(vi) From Eq. 2.34, it is clear that $k_x = 0$ (i.e. the wavelength of the guided wave becomes infinite) when

$$\omega = \frac{n\pi v}{b}.$$

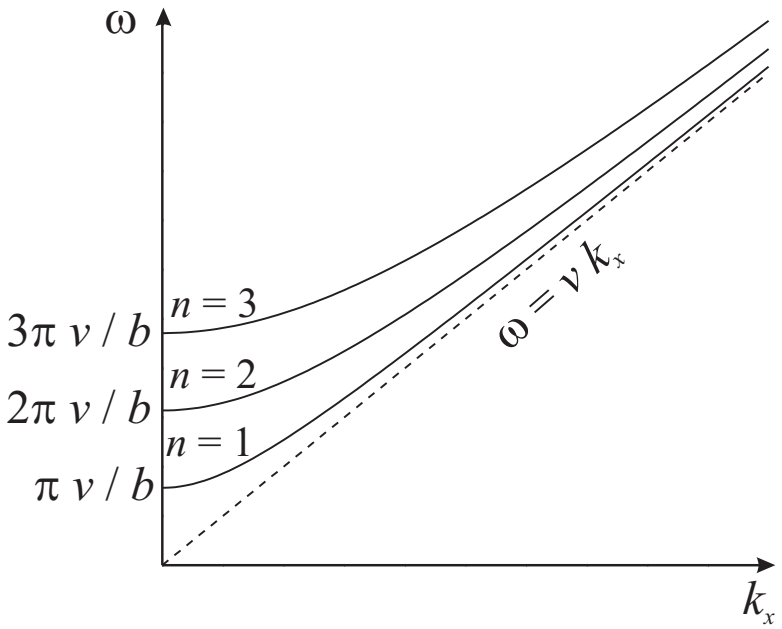


Figure 2.22: Dispersion relation for a 2-d guided wave.

Below this frequency, k_x^2 becomes negative and propagation in the normal sense is not possible. This frequency is therefore called the cut-off frequency, ω_c, for the particular mode. The lowest cut-off frequency occurs for $n = 1$, and is sometimes called the cut-off frequency of the guide.

(vii) Usually we are interested in propagating waves covering a small range of frequencies centred on a particular frequency ω_0. From the dispersion relation and the plot of ω against k_x it is clear that, in general, several different modes (different values of n) are possible. Unless the spatial profile of the input signal is chosen very carefully, the input signal will excite a number of different modes, each of which travels at a different speed. The output signal will therefore be distorted with respect to the input. To avoid this problem the guide can be designed so that only the $n = 1$ mode can propagate, i.e.

$$\frac{\pi v}{b} < \omega_0 < \frac{2\pi v}{b}$$

so

$$\frac{\pi v}{\omega_0} < b < \frac{2\pi v}{\omega_0}.$$

Optical fibres are an extremely important example of a waveguide, since they are used extensively in telecommunications systems where they transmit a large proportion of the digital information on the telephone and internet networks. An optical fibre comprises a cylindrical silica core, surrounded by a cladding of lower refractive index (Figure 2.23). Optical waves can be guided down the core of the fibre. The data is sent as a series of pulses of light, and the data rate which can be achieved over a given length of fibre is determined by the spreading out of these pulses as they are transmitted along the fibre, i.e. by the dispersion. The aim is to choose a wavelength where the dispersion is a minimum, but the fibre must be designed so that this wavelength is in a region of the spectrum where the losses in the fibre (due to absorption, for example) are very small.

Just as in the membrane waveguide, different modes will exist, and will have different group velocities. Propagation of light in more than one mode would therefore lead to major pulse spreading, hence most fibres are designed with a core which is sufficiently small that only one propagating

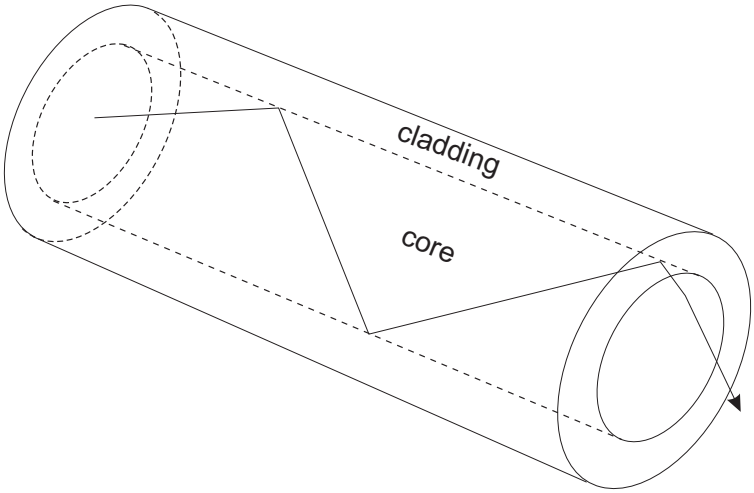


Figure 2.23: Schematic representation of an optical fibre.

mode exists. With just this mode, there is still dispersion since the wave is confined, and there is more dispersion since the refractive index itself depends on wavelength. The trick to designing a fibre is to choose the geometry and operating wavelength such that these two contributions to the dispersion have opposite signs and therefore cancel. Current optical fibre systems can carry data at rates of 10 Gbits/s per wavelength channel by modulating a single-wavelength source operating at 1550 nm. By combining hundreds of such channels with slightly different wavelengths (wavelength-division multiplexing), transmission rates of many Tbits/s can be achieved.

2.10.2 Evanescent waves

Now consider what happens to a wave at frequencies below the cut-off frequency. A wave propagating in a waveguide has the form

$$\Psi = A \sin k_y y \exp i(\omega t - k_x x).$$

Now, below the cut-off frequency ω_c, where

$$\omega_c = \frac{n\pi v}{b}$$

k_x^2 becomes negative, i.e.

$$k_x = \pm i\alpha$$

where α is real. Hence

$$\Psi = A \sin k_y y \exp i\omega t \exp \pm \alpha x \tag{2.38}$$

i.e. the wave no longer shows an oscillatory behaviour in the x direction, but the amplitude decays exponentially with distance along the wave. A wave of this sort is called an *evanescent wave*.

If the membrane extends to distances $x \gg 1/\alpha$, the entire surface would move up and down in phase, with an amplitude that decays away in the x direction. (The positive exponential solution with a positive exponent would not conserve energy, and hence does not contribute.) Since the phase is now independent of position (as in standing waves), there is no net power flow along the guide. As a consequence, a travelling wave with $\omega < \omega_c$ which is incident on the guide will undergo perfect reflection (with a phase shift) and a standing wave will be set up outside the guide, as shown in Figure 2.24.

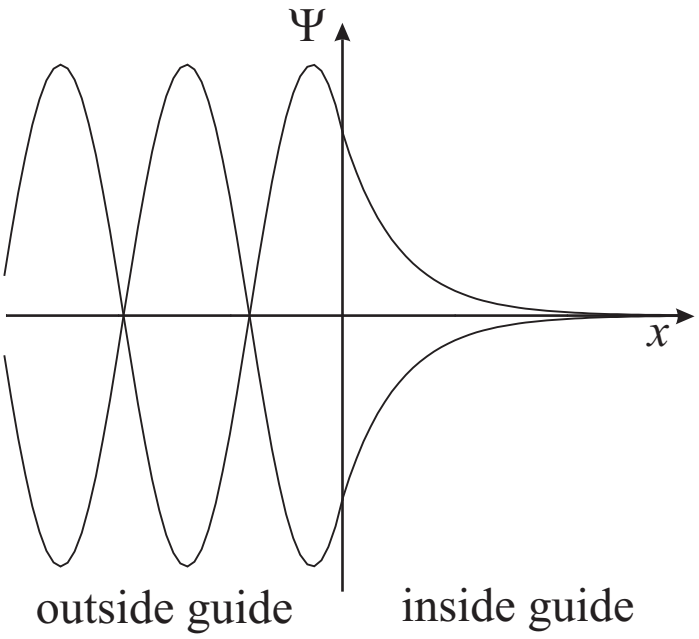


Figure 2.24: Travelling wave with $\omega < \omega_c$ incident on a waveguide.

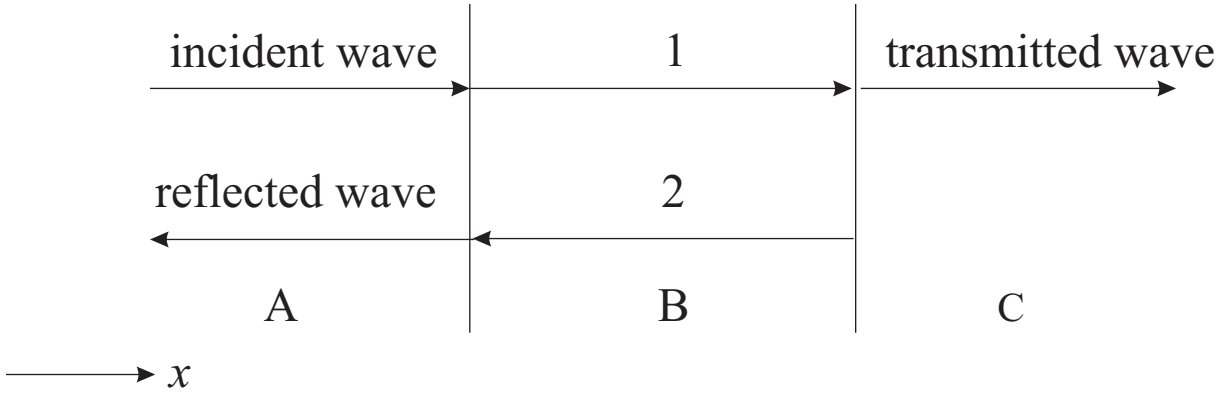


Figure 2.25: Evanescent wave between two discontinuities.

If, however, there were a second discontinuity beyond which it is again possible to have a travelling wave, some of the incident wave would be transmitted into this region, (C), and there will only be partial reflection at the boundary AB. There are two evanescent waves in region B; wave 1 decays as $\exp -\alpha x$, and wave 2 increases as $\exp \alpha x$. The combination of these two gives a position dependence of the phase angle for the total disturbance in region B and, as a result, a net flow of energy from region A to region C is now possible.

You will see more examples of this phenomenon in quantum mechanics, where the presence of an evanescent wavefunction for a particle in a classically forbidden region leads to the phenomenon of *tunnelling*.

3 Fourier transforms

3.1 Superposition

Recall the equation of motion for the damped driven oscillator

$$F(t) = m\ddot{x}(t) + b\dot{x}(t) + kx(t). \tag{3.1}$$

We include the time dependence of F and x explicitly to remind us that in general they are not necessarily sinusoidal. This equation is *linear*, since the force depends linearly on the displacement and its derivatives. A consequence of this linearity is that solutions can be superimposed, i.e.

$$\begin{array}{ccccc} \text{if} & F_1(t) & \text{gives displacement} & x_1(t) \\ \text{and} & F_2(t) & \text{gives displacement} & x_2(t) \\ \text{then} & c_1F_1(t) + c_2F_2(t) & \text{gives displacement} & c_1x_1(t) + c_2x_2(t). \end{array}$$

Consider a force which is the sum of two sinusoidal driving forces at different frequencies. We can calculate the individual responses to the two forces (using the response function $R(\omega)$), and add the results.

$$\begin{array}{ccc} \mathbf{F}_1 e^{i\omega_1 t} & \rightarrow & R(\omega_1)\mathbf{F}_1 e^{i\omega_1 t} \\ + & & + \\ \mathbf{F}_2 e^{i\omega_2 t} & \rightarrow & R(\omega_2)\mathbf{F}_2 e^{i\omega_2 t} \\ \| & & \| \\ \mathbf{F}_1 e^{i\omega_1 t} + \mathbf{F}_2 e^{i\omega_2 t} & \rightarrow & R(\omega_1)\mathbf{F}_1 e^{i\omega_1 t} + R(\omega_2)\mathbf{F}_2 e^{i\omega_2 t} \end{array} \tag{3.2}$$

This leads to the idea that if we can decompose a force into sinusoids, we can determine the response to that force in terms of the sum of the responses to each of the sinusoids.

3.2 Fourier Series

So far, we have dealt with sinusoidal oscillations. In general, however, we may need to know the response of an oscillator to any driving force. In some cases, we can solve the equation of motion explicitly as a function of time (see Section 1.4), however it is usually more convenient to work in terms of frequency. For example, suppose we wish to find the response of an oscillator to a square wave force. You will have seen in IA Maths that any periodic function can be made up as a sum of sine and cosine functions, using a Fourier series. Once we know the Fourier components that make up the driving force, we can calculate the response to each of these components using the response function $R(\omega)$. By the principle of superposition, just as in the previous section, we can then add up the responses to each Fourier component to obtain the total response.

Before we extend these ideas to the general case of non-periodic functions, let us recall the mathematics of Fourier series.

Consider a periodic function $f(t)$, with a period T corresponding to an angular frequency $\omega_0 = 2\pi/T$. The function can be written as a Fourier series

$$f(t) = \frac{1}{2}A_0 + \sum_{n=1}^{\infty}\left(A_n \cos\left(\frac{2\pi nt}{T}\right) + B_n \sin\left(\frac{2\pi nt}{T}\right)\right) \tag{3.3}$$

i.e. it is a superposition of sinusoidal waves with frequencies $\omega_0, 2\omega_0, 3\omega_0 \ldots$. To find the value of a particular coefficient A_m or B_m, we multiply the above equation by $\cos\left(\frac{2\pi mt}{T}\right)$ or $\sin\left(\frac{2\pi mt}{T}\right)$ and

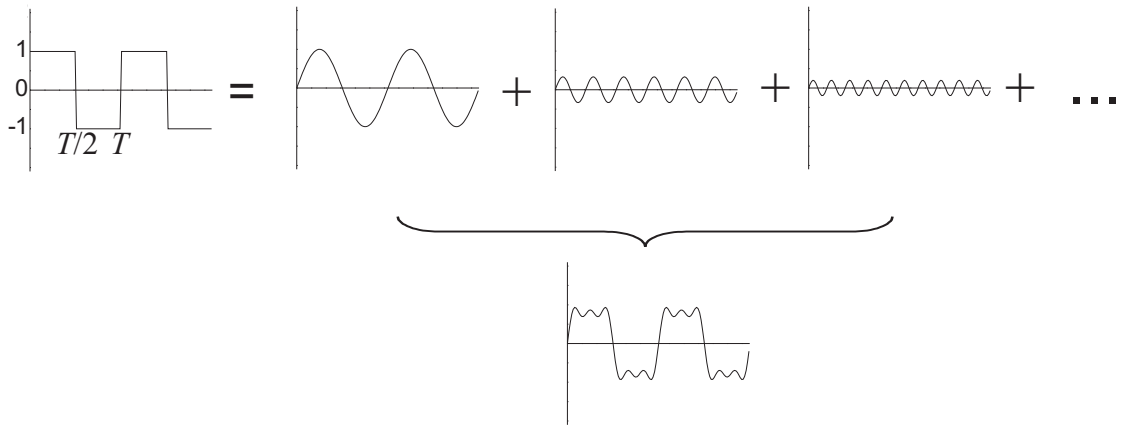


Figure 3.1: Fourier analysis of a square wave.

integrate from $-T/2$ to $T/2$ to give

$$A_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\cos\left(\frac{2\pi nt}{T}\right)dt \tag{3.4}$$

$$B_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin\left(\frac{2\pi nt}{T}\right)dt. \tag{3.5}$$

These results rely on the fact that

$$\begin{aligned}
\int_{-T/2}^{T/2}\cos\left(\tfrac{2\pi mt}{T}\right)\cos\left(\tfrac{2\pi nt}{T}\right)dt &= 0 \quad \text{for } m \neq n \\
&= T/2 \quad \text{for } m = n \\
\int_{-T/2}^{T/2}\cos\left(\tfrac{2\pi mt}{T}\right)\sin\left(\tfrac{2\pi nt}{T}\right)dt &= 0 \quad \text{for all } m, n \\
\int_{-T/2}^{T/2}\sin\left(\tfrac{2\pi mt}{T}\right)\sin\left(\tfrac{2\pi nt}{T}\right)dt &= 0 \quad \text{for } m \neq n \\
&= T/2 \quad \text{for } m = n.
\end{aligned}$$

Symmetry considerations are often helpful in evaluating the Fourier coefficients. For example, the square wave shown in Figure 3.1 is an *odd* function ($f(x) = -f(-x)$), hence $A_n = 0$ for all n. The other coefficients can be shown to be

$$B_n = \frac{4}{\pi n} \text{ for } n \text{ odd} \tag{3.6}$$

$$B_n = 0 \text{ for } n \text{ even}, \tag{3.7}$$

giving

$$f(t) = \frac{4}{\pi}\left(\sin(\omega_0 t) + \frac{1}{3}\sin(3\omega_0 t) + \frac{1}{5}\sin(5\omega_0 t)\ldots.\right) \tag{3.8}$$

Figure 3.2 shows the effect of driving a damped oscillator with a square wave, calculated from its response to the Fourier components. Since the response of the oscillator tends to zero at high frequencies, it is often only necessary to consider the first few Fourier components to obtain a good approximation to the response.

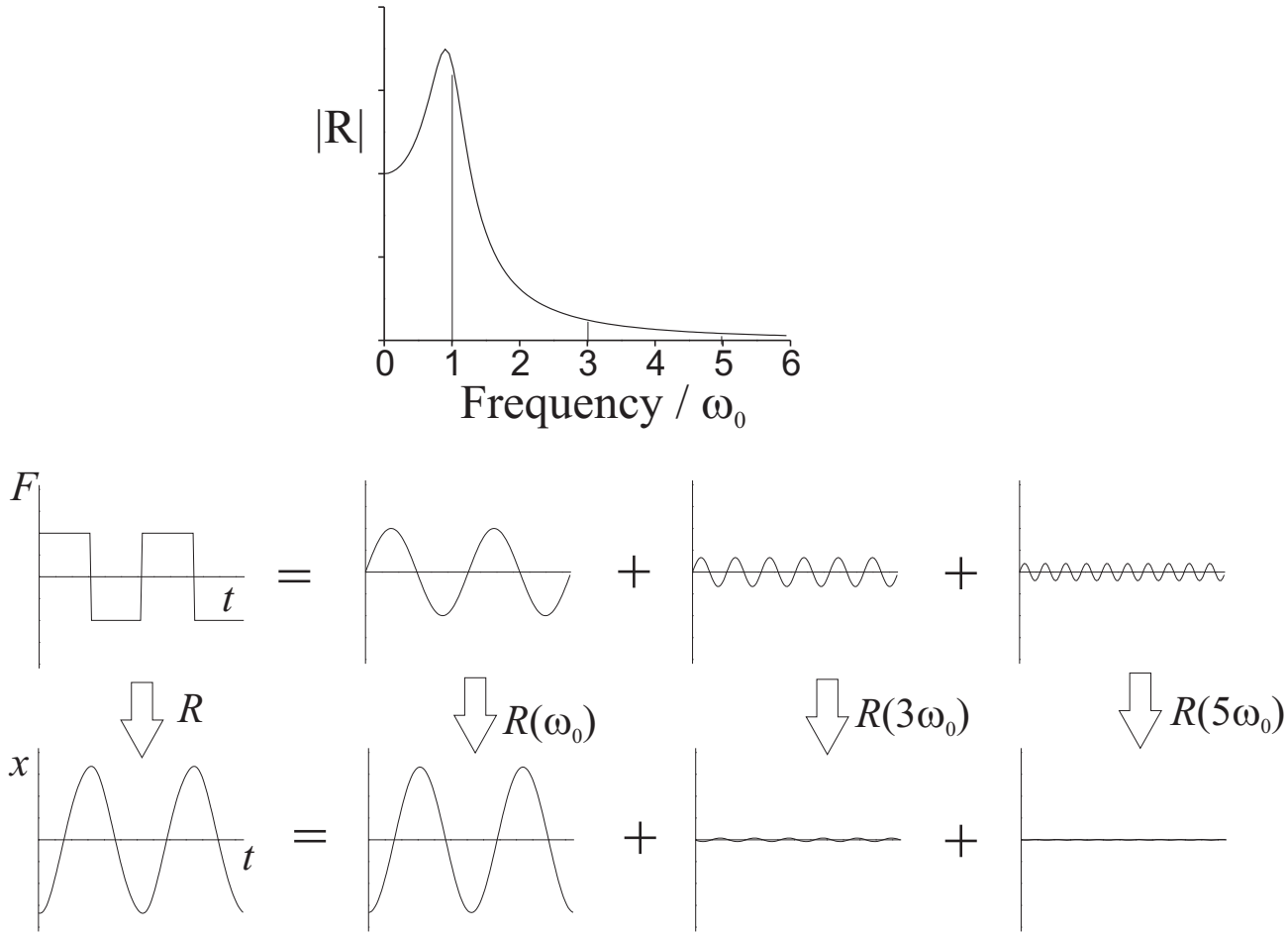


Figure 3.2: Response function for a damped oscillator with $Q = 1.7$, and the response of this oscillator to a square wave at frequency ω_0

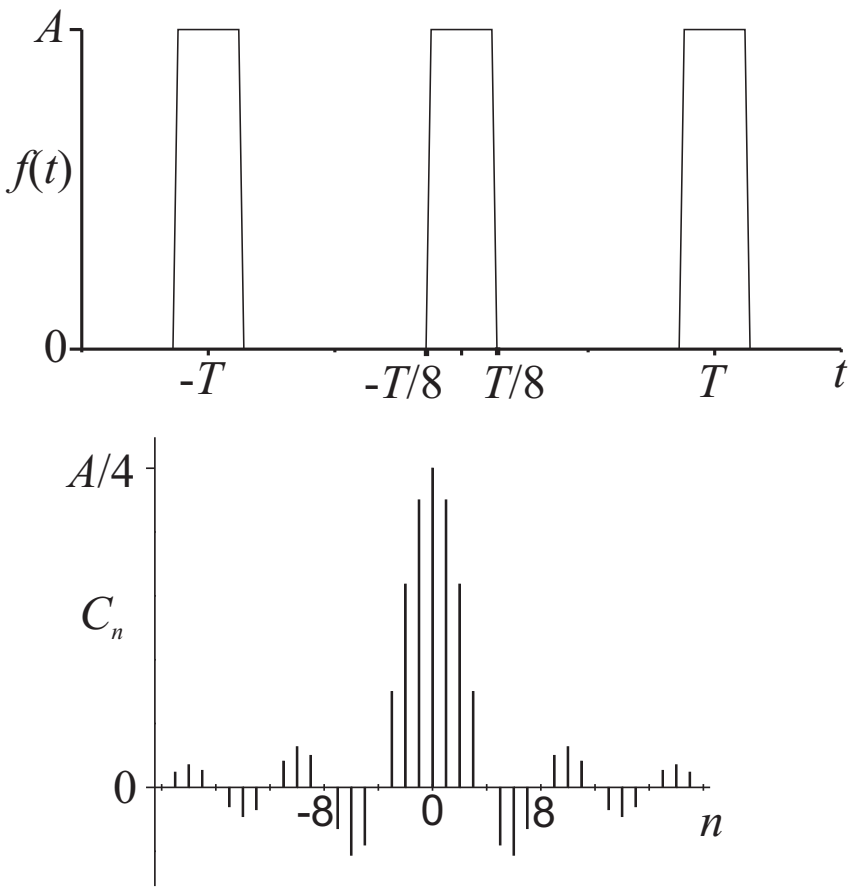


Figure 3.3: Fourier coefficients for a periodic function.

3.2.1 Complex coefficients

Rather than write the series as a sum of sine and cosine terms, it often convenient to describe the Fourier components with complex coefficients which represent the amplitude and phase of each component.

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{i2\pi nt/T} = \sum_{n=-\infty}^{\infty} C_n e^{in\omega_0 t}. \tag{3.9}$$

The coefficients C_n are given by multiplying by $e^{-im\omega_0 t}$ and integrating from $-T/2$ to $T/2$,

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-in\omega_0 t} dt. \tag{3.10}$$

This results relies on the fact that

$$\begin{aligned} \int_{-T/2}^{T/2} e^{in\omega_0 t} e^{-im\omega_0 t} dt &= 0 \quad \text{for } m \neq n \\ &= T \quad \text{for } m = n. \end{aligned} \tag{3.11}$$

We are interested in functions $f(t)$ which are real, and in this case we find that $C_{-m} = C_m^*$. This ensures that the imaginary parts of the terms in Eq. 3.9 cancel to zero when summing over positive and negative n.

As an example of calculating Fourier coefficients using this method, let us consider the function

shown in Figure 3.3, which repeats with period $T = 2\pi/\omega_0$. In the range $-T/2 < t < T/2$,

$$f(t) = A \quad -T/8 < t < T/8 \tag{3.12}$$
$$= 0 \quad T/8 < |t| < T/2 \tag{3.13}$$

The Fourier coefficients are given by

$$C_n = \frac{1}{T}\int_{-T/8}^{T/8} Ae^{-in\omega_0 t}\,dt \tag{3.14}$$
$$= \frac{A}{T}\left[\frac{e^{-in\omega_0 t}}{-in\omega_0}\right]_{-T/8}^{T/8} \tag{3.15}$$
$$= \frac{A}{in\omega_0 T}\left(e^{in\omega_0 T/8} - e^{-in\omega_0 T/8}\right) \tag{3.16}$$
$$= \frac{A}{2in\pi}\left(e^{in\pi/4} - e^{-in\pi/4}\right) \tag{3.17}$$
$$= \frac{A}{\pi n}\sin(n\pi/4) \tag{3.18}$$
$$= \frac{A}{4}\,\mathrm{sinc}\,(n\pi/4). \tag{3.19}$$

The coefficient is zero whenever n is a multiple of 4.

3.3 Fourier Transforms

We now need to extend the idea of Fourier series to non-periodic functions. Conceptually, this can be seen as taking the limit of a Fourier series as $T \to \infty$. The fundamental frequency tends to zero, and the frequency components making up the series become infinitesimally close together. The sum over discrete frequency components then becomes an integral over a continuous spectrum of frequency components. Mathematically, we write our function as

$$f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(\omega)e^{i\omega t}d\omega \tag{3.20}$$

where

$$g(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt. \tag{3.21}$$

These equations define the Fourier transform; $g(\omega)$ is the Fourier transform of $f(t)$ and $f(t)$ is the inverse Fourier transform of $g(\omega)$. There is some arbitrariness in the choice of constants before the integrals: their product must be $(2\pi)^{-1}$, but you may see various conventions for how this is split up between the two integrals.

We denote the Fourier transform with the *operator* $\mathcal{F}$ (or sometimes "F.T.") which transforms the function f into the function g (i.e separating a function into its component sinusoids)

$$\mathcal{F}[f(t)] = g(\omega), \tag{3.22}$$

and its inverse $\mathcal{F}^{-1}$ which derives f given g (adding the sinusoids back together)

$$\mathcal{F}^{-1}[g(\omega)] = f(t). \tag{3.23}$$

Fourier transforms are not just applied to functions of time: they can equally be applied to functions of any variable. For example we can Fourier transform a function of a spatial variable x to derive a function of wavenumber k, i.e.

$$\mathcal{F}[f(x)] = g(k) \tag{3.24}$$

The spaces represented by ω and k are "reciprocal spaces" complementary to t and x respectively, since they have units of s^{-1} and m^{-1} respectively. The idea of a reciprocal space will be familiar from crystallography, where Fourier transforms abound.

Fourier theory is a very powerful way of analysing linear systems like oscillators, because the Fourier transform $g(\omega)$ represents the splitting up of our arbitrary function $f(t)$ into its component sinusoids. For each angular frequency ω, then $|g(\omega)|$ tells us the amplitude of the component sinusoid at that frequency and $\arg[g(\omega)]$ tells us the phase. For a system like an oscillator, we know the response to each of these sinusoids is another sinusoid at the same frequency, related to the input sinusoid via a response function $R(\omega)$. Thus we can calculate the total response by adding back together the output sinusoids, using the inverse Fourier transform. Thus we can write down the solution for the response of an oscillator to an arbitrary input force $F(t)$ in terms of Fourier transforms:

$$x(t) = \mathcal{F}^{-1}\left[R(\omega)\mathcal{F}\{F(t)\}\right]. \tag{3.25}$$

This kind of approach is very general, and we will use Fourier theory many times in this course, and in other courses this year.

3.3.1 The power spectrum

The spectrum of a signal, often called the power spectrum, is simply the modulus squared of its Fourier transform, i.e. $|g(\omega)|^2$. This particular form occurs in many experimental situations where the phase of $g(\omega)$ cannot be measured or is not meaningful. For example, a prism splits light into sinusoids of different frequencies i.e. a Fourier transform. The light wave is oscillating at very high frequencies, so we typically cannot measure the phase of the light in each spectral channel and instead we record the intensity of that light as a function of frequency, i.e. we take a value proportional to the modulus squared of the light amplitude, either by looking at the spectrum with our eye or electronically in a *spectrometer*.

3.3.2 Example Fourier transforms

The Fourier transform can be evaluated analytically in simple cases. Consider the top hat-function

$$f(t) = A \quad -\Delta t/2 < t < \Delta t/2 \tag{3.26}$$

$$= 0 \quad |t| > \Delta t/2. \tag{3.27}$$

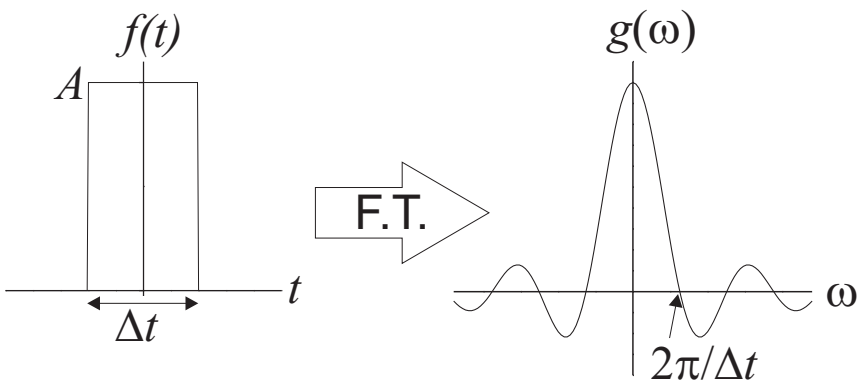


Figure 3.4: Fourier transform of a top-hat function.

The Fourier transform, $g(\omega)$ is given by

$$g(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\Delta t/2}^{\Delta t/2} A e^{-i\omega t} dt \tag{3.28}$$

$$= \frac{A}{\sqrt{2\pi}} \left[\frac{e^{-i\omega t}}{-i\omega} \right]_{-\Delta t/2}^{\Delta t/2} \tag{3.29}$$

$$= \frac{A}{\sqrt{2\pi}} \left(\frac{e^{i\omega \Delta t/2} - e^{-i\omega \Delta t/2}}{i\omega} \right) \tag{3.30}$$

$$= \frac{2A}{\sqrt{2\pi}\,\omega} \sin(\omega \Delta t/2) \tag{3.31}$$

$$= \frac{A\Delta t}{\sqrt{2\pi}} \operatorname{sinc}(\omega \Delta t/2). \tag{3.32}$$

This demonstrates the more general result that the width in the frequency domain is inversely proportional to the width in the time domain.

The Fourier transform of an arbitrary function can be calculated rapidly by computer using a "Fast Fourier Transform" (FFT). Figure 3.5 shows the time-domain waveform of the spoken syllable "ah", together with its frequency spectrum computed using an FFT. The fundamental frequency is easily visible in the time domain - this is the frequency of oscillation of the vocal chords which gives the perceived "pitch" of the vowel. Other frequency components are hard to distinguish in the waveform, but are easy to see in the frequency spectrum. These "formants" are determined by resonances in the vocal tract, and give the characteristic timbre which allow us to distinguish the vowel as"ah".

3.4 Delta Functions

An function which is important in the theory of Fourier transforms is the *Dirac delta-function*. This function is used to denote a sharp "spike" which occurs over an infinitesimal time, but with finite area. An example of where this might be used is the idea of an "impulse" in Newtonian mechanics, where a sharp "kick" imparts finite momentum even though it occurs over an infinitesimally short time Δt. The momentum change is $F\Delta t$, so as Δt tends to zero, then F must tend to infinity during the kick. We would represent the force as a function of time in this limiting case as a Dirac delta-function.

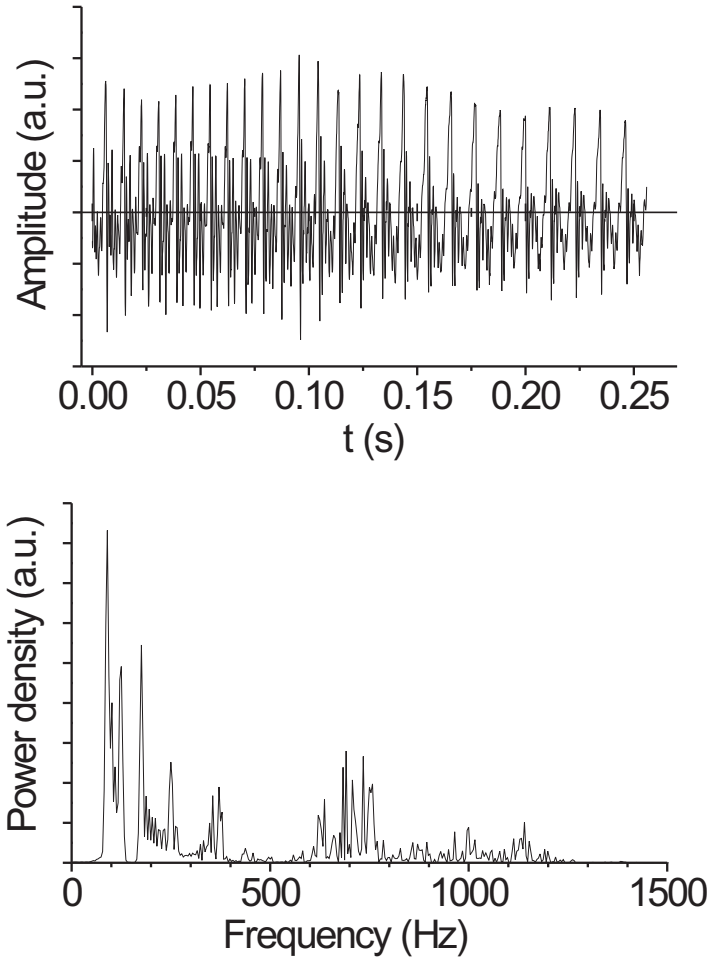


Figure 3.5: Waveform of "ah" and its Fourier transform (shown as a power spectrum).

Formally, the Dirac delta function $\delta(t)$ is defined by the property

$$\int_{-\infty}^{\infty} \delta(t) f(t)\, dt = f(0) \tag{3.33}$$

for any arbitrary function $f(t)$. This means that $\delta(t)$ is a unit-area "spike" at $t = 0$ and is zero everywhere else. The function $\delta(t - t_0)$ is the same "spike" offset from the origin by an amount t_0 so

$$\int_{-\infty}^{\infty} \delta(t - t_0) f(t)\, dt = f(t_0)\,. \tag{3.34}$$

In other words, multiplying a function by $\delta(t - t_0)$ and integrating returns a "sample" of the value of the function at time t_0.

We can use this property to show that the Fourier transform of a delta function is a complex exponential:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \delta(t - t_0) e^{-i\omega t}\, dt = \frac{e^{-i\omega t_0}}{\sqrt{2\pi}}\,. \tag{3.35}$$

Note that the Fourier transform of a δ-function at the origin is a constant (i.e. independent of ω).

3.5 Convolution

We now introduce the idea of the *convolution* of two functions $f(y)$ and $g(y)$, denoted by the operator "$*$". The definition of convolution is

$$f(y) * g(y) = \int_{-\infty}^{\infty} f(u)g(y-u)\,du. \tag{3.36}$$

The easiest example of convolution to visualise is when one of the functions is a δ-function.

$$f(y) * \delta(y-y_0) = \int_{-\infty}^{\infty} f(u)\delta(y-y_0-u)\,du = f(y-y_0). \tag{3.37}$$

In other words, the function $f(y)$ is reproduced centred around the δ-function rather than around zero. For general functions, we can think of $g(y)$ as being made up of the sum of an infinite number of δ-functions with different"heights".

$$g(y) = \int_{-\infty}^{\infty} g(u)\delta(y-u)\,du \tag{3.38}$$

Hence, when we take the convolution of f and g, each of the δ-functions making up g is replaced by a copy of f. The function g is therefore "smeared out" by the function f (and vice versa).

Convolution has an important application in optics where imaging systems such as microscopes or telescopes are concerned. These instruments are imperfect, so a point source object (a δ-function) is smeared out in the image. The image produced from a delta function object is described by the resolution function (or *point spread function*) of the instrument ($f(\mathbf{r})$). For a general object, the image produced is the convolution of the object with the resolution function. The operation of trying to remove the effects of the resolution function is known as *deconvolution*.

Convolutions are particularly useful in the context of Fourier transforms, and hence in Fraunhofer diffraction. It can be shown that the Fourier transform of the product of two functions is the convolution of the Fourier transforms of the individual functions, i.e. if

$$F(q) = \mathcal{F}[f(x)] \tag{3.39}$$

and

$$G(q) = \mathcal{F}[g(x)] \tag{3.40}$$

then

$$\mathcal{F}[f(x)g(x)] = \frac{1}{\sqrt{2\pi}}\,F(q) * G(q). \tag{3.41}$$

Similarly, the Fourier transform of the convolution of two functions is the product of the Fourier transforms of the individual functions.

$$\mathcal{F}[f(x) * g(x)] = \sqrt{2\pi}\,F(q)G(q). \tag{3.42}$$

3.6 Impulse response functions

Consider the response of an oscillator (or any linear system) to a sharp impulse at time $t = 0$. The response after time t to the impulse

$$F(t) = \delta(t) \tag{3.43}$$

defines the **impulse response function** $R'(t)$. For the damped oscillator considered above, the impulse response function is simply the transient after giving the system an initial velocity $v_0 = 1/m$ at $t = 0$. Suppose, though, that we do not know the transient response of our system, only the frequency response function $R(\omega)$ in the frequency domain. We can then use Fourier theory to find $R'(t)$. Equation 3.25 tells us to take the Fourier transform of $F(t)$, multiply by $R(\omega)$ to calculate the response in the frequency domain, then inverse Fourier transform the result to get the response in the time domain. The Fourier transform of $\delta(t)$ is a constant, hence the impulse contains all frequencies with equal amplitude

$$\delta(t) \xrightarrow{\text{F.T.}} F_\omega(\omega) = \frac{1}{\sqrt{2\pi}}. \tag{3.44}$$

The response is given by

$$x_\omega(\omega) = R(\omega)F_\omega(\omega) = \frac{1}{\sqrt{2\pi}}R(\omega). \tag{3.45}$$

To find the response in the time domain, $R'(t)$, we take the inverse Fourier transform

$$R'(t) = \frac{1}{\sqrt{2\pi}}\mathcal{F}^{-1}\left\{R(\omega)\right\}. \tag{3.46}$$

Hence we can see that $R'(t)$ and $R(\omega)$ are directly related by the Fourier transform: if we know the frequency response we can calculate the impulse response, and vice versa.

If we know $R'(t)$, we can calculate the response to any driving force $F(t)$. We can write $F(t)$ as the sum of an infinite number of δ-functions with different weights:

$$F(t) = \int_{-\infty}^{\infty} F(t')\delta(t - t')\, dt'. \tag{3.47}$$

Using the principle of superposition, we can then find the responses of the system to each of these δ-functions, and add them up:

$$x(t) = \int_{-\infty}^{\infty} F(t')R'(t - t')\, dt'. \tag{3.48}$$

We note that this is just the *convolution* of F with R': this is to be expected from the fact that Eq. 3.25 contains the product of their Fourier transforms. Hence, knowing either $R'(t)$ or $R(\omega)$ gives us all the information we need about a well-behaved linear system in order to calculate its response to any driving force.

This is of some practical importance when designing experiments to measure response. Suppose we are designing a sensitive experiment where the apparatus must be free of vibrations. We need to arrange that it does not have any strong resonances which can be excited by noise from the surroundings. To check for resonances, we would like to measure $R(\omega)$. The obvious way to do this would be to fit a transducer to the apparatus to drive it at a particular frequency, and then measure the amplitude of vibration using a sensor (a microphone), also attached to the apparatus. The amplitude and phase of the response must be measured as the driving frequency is changed - a time-consuming and tedious experiment. An alternative, and much more efficient, experiment is as follows: attach the microphone as before, and then hit the apparatus with a hammer, recording the microphone output. This directly gives $R'(t)$. We then use a computer to take the Fourier transform of the microphone output, which gives $R(\omega)$ and allows us to examine the resonance behaviour.

$f(t)$	$g(\omega)$
delta-function $\delta(t-t_0)$	complex exponential $\frac{e^{-i\omega t_0}}{\sqrt{2\pi}}$
top-hat $f(t)=\begin{cases}0 & \text{if } \lvert t\rvert > \frac{1}{2}\\ 1 & \text{if } \lvert t\rvert \le \frac{1}{2}\end{cases}$	sinc $\frac{\sin(\omega/2)}{\sqrt{2\pi}\omega/2}$
Gaussian $e^{-t^2/2}$	Gaussian $e^{-\omega^2/2}$
decaying exponential $f(t)=\begin{cases}0 & \text{if } t<0\\ e^{-t} & \text{if } t\ge 0\end{cases}$	low-pass filter $\frac{1}{\sqrt{2\pi}(1+i\omega)}$
comb function $\sum_{n=-\infty}^{\infty}\delta(t-n)$	comb function $\sqrt{2\pi}\sum_{n=-\infty}^{\infty}\delta(\omega-2\pi n)$

Table 3.1: Fourier transforms of some useful functions

3.7 Composing Fourier transforms

Evaluating the Fourier transform or inverse Fourier transform of a function by performing the integrals in Eqs. 3.20 and 3.21 can be time-consuming. An alternative in many cases is to make use of a few "building-block" Fourier transforms and combine them using known mathematical properties of the Fourier transform. Some commonly-used functions and their transforms are given in Table 3.1 and the following results can be used to extend and combine the transforms:

Reciprocity: If you know the forward transform of a function, you can obtain the inverse transform of the same function by "flipping" the result about the $t = 0$ axis: if the F.T. of $f(t)$ is $g(\omega)$, then the inverse F.T. of $f(\omega)$ is $g(-t)$. This follows from the similarity of the integrals defining forward and inverse transforms, and means that all the following results apply when you replace $\mathcal{F}$ with $\mathcal{F}^{-1}$.

Scaling law: If we "stretch" a function horizontally by an amount a, the corresponding dimension in the transform is compressed by the same factor: $\mathcal{F}[f(t/a)] = |a|g(a\omega)$. This reciprocal scaling is the origin of Heisenberg's uncertainty principle, since wavefunctions of quantum variables such as position and momentum form a Fourier transform pair. Note that the vertical dimension of the transformed function is stretched by $|a|$.

Linearity: If a function is the superposition of two other functions, its F.T. is the superposition of the respective F.T.'s: $\mathcal{F}[a_1 f_1(t) + a_2 f_2(t)] = a_1\mathcal{F}[f_1(t)] + a_2\mathcal{F}[f_2(t)]$ where a_1 and a_2 are arbitrary constants and f_1 and f_2 are arbitrary functions.

The convolution theorem: The Fourier transform of the product of two functions equals the convolution of the individual Fourier transforms:

$$\mathcal{F}[f_1(t)f_2(t)] = \frac{1}{\sqrt{2\pi}}\mathcal{F}[f_1(t)] * \mathcal{F}[f_2(t)]$$

And the Fourier transform of the convolution of two functions equals the product of the individual Fourier transforms:

$$\mathcal{F}[f_1(t) * f_2(t)] = \sqrt{2\pi}\,\mathcal{F}[f_1(t)]\mathcal{F}[f_2(t)].$$

3.7.1 Examples

Here we give a couple of simple examples of the application of the Fourier transform properties to extend the known transforms to some useful cases.

Cosine function

A cosine can be decomposed into a sum of exponentials

$$\cos(\omega_0 t) = \tfrac{1}{2}(e^{i\omega_0 t} + e^{-i\omega_0 t})$$

Using the first line of Table 3.1 and **reciprocity**, the F.T. of $e^{i\omega_0 t}$ is $\sqrt{2\pi}\delta(\omega - \omega_0)$ and the F.T. of $e^{-i\omega_0 t}$ is $\sqrt{2\pi}\delta(\omega + \omega_0)$. Using **linearity** we have

$$\mathcal{F}[\cos(\omega_0 t)] = \frac{\sqrt{2\pi}}{2}[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)] \tag{3.49}$$

So the F.T. of a cosine function is a pair of delta-functions, one on either side of the origin.

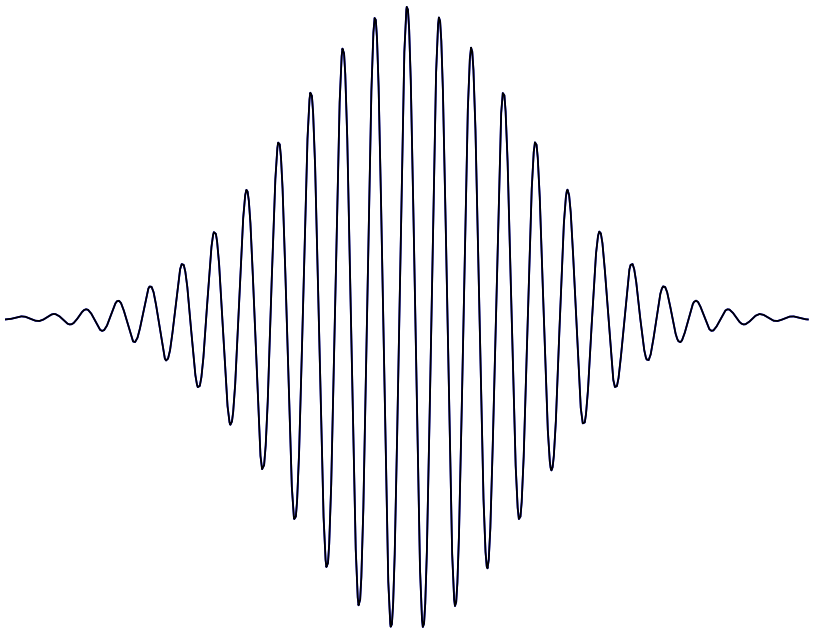


Figure 3.6: A Gaussian wavepacket

Gaussian wavepacket

A commonly encountered function in wave theory is the Gaussian wavepacket, a sinusoid multiplied by a Gaussian envelope as shown in Fig. 3.6. We can write this as

$$f(t) = Ae^{-t^2/2\sigma^2}\cos\omega_0 t$$

where σ is the width of the Gaussian, A is the peak amplitude and ω_0 is the "carrier frequency" of the packet. We decompose this function into the product of two functions we know the transforms of, namely a Gaussian and a cosine. For the Gaussian, we can use the **scaling** law and **linearity** to get the transform:

$$\begin{aligned} \mathcal{F}\left\{e^{-t^2/2}\right\} &= e^{-\omega^2/2} \\ \Rightarrow \mathcal{F}\left\{Ae^{-t^2/2\sigma^2}\right\} &= A\sigma e^{-\omega^2\sigma^2/2} \end{aligned}$$

The **convolution theorem** then allows us to combine the F.T. of the Gaussian with the F.T. of the cosine (Eq. 3.49) to give

$$\begin{aligned} g(\omega) &= \frac{1}{\sqrt{2\pi}}\mathcal{F}\left\{Ae^{-t^2/2\sigma^2}\right\} * \mathcal{F}\left\{\cos\omega_0 t\right\} \\ &= \tfrac{1}{2}A\sigma\left([\delta(\omega-\omega_0)+\delta(\omega+\omega_0)] * e^{-\omega^2\sigma^2/2}\right) \\ &= \tfrac{1}{2}A\sigma\left(e^{-(\omega-\omega_0)^2\sigma^2/2} + e^{-(\omega+\omega_0)^2\sigma^2/2}\right). \end{aligned} \tag{3.50}$$

In other words we get two Gaussians centred around $\pm\omega_0$.

The width of the Gaussians in the frequency domain is inversely proportional to the width of the Gaussian in the time domain, so we have the result that a short packet occupies more *bandwidth*

than a longer packet. If, for example, we were using such packets to transmit digital pulses over a radio link at a central frequency of ω_0, then it is clear that if we want to transmit shorter pulses we need a larger frequency allocation to avoid overlapping nearby frequency channels.

3.8 Symmetry

It is often helpful to know the symmetry properties of Fourier transforms in order to check that we have got the right results. It is straightforward to show from Eq. 3.21 that the Fourier transform of a function $f(t)$ which is purely real (something which is true of most of the physical variables we will be taking the Fourier transform of) has so-called **Hermitian** symmetry, i.e. that $g(-\omega) = g(\omega)^*$. This means that the properties of the function can be determined purely from the positive frequency components of the Fourier transform, so typically we only plot the positive half of the F.T. in these cases.

Similarly, we can show that if a function is both real and symmetric (i.e. $f(-t) = f(t)$) then its Fourier transform is both real and symmetric, while if the function is both real and antisymmetric (i.e. $f(-t) = -f(t)$), then its transform is purely imaginary and antisymmetric.

4 Optics and Diffraction

4.1 Electromagnetic waves

Optics deals with the propagation of electromagnetic waves (EM waves for short), which obey all the phenomena we have seen so far in wave theory, including reflection and transmission, and propagation of energy. In a vacuum, EM waves are non-dispersive, so that light of any wavelength propagates at the same speed. If you do the Electromagnetism course in Physics B, you will show from Maxwell's equations for electromagnetism that this speed has the value $c_0 = 1/\sqrt{\epsilon_0\mu_0} \approx 3 \times 10^8 \text{m/s}$ in a vacuum, where ϵ_0 and μ_0 are the electric permittivity and magnetic permeability of free space respectively, familiar from IA electromagnetism.

If the electromagnetic wave is not travelling in a vacuum, but is instead travelling in a medium such as a solid, liquid, gas or plasma, the medium will interact with the wave. Currents and charges induced in the medium by the passage of the wave will create additional electric and magnetic fields to those from the wave itself. In most cases of uniform media, this difference can be accounted for by introducing extra constants into Maxwell's equations. For dielectric media, where an electric field induces a proportional charge polarisation in the medium we replace the electric permittivity of free space ϵ_0 with the permittivity of the medium ϵ, which can be expressed in terms of a relative permittivity ϵ_r, as $\epsilon = \epsilon_0\epsilon_r$. Similarly, for magnetic media the magnetic field induces a proportional magnetisation of the material, and Maxwell's equations can be altered by replacing the permeability of free space μ_0, with the permeability of the medium $\mu = \mu_0\mu_r$, where the relative permeability μ_r is defined in the obvious way. In such media, the speed of light is changed to

$$c = \frac{1}{\sqrt{\epsilon\mu}} = \frac{1}{\sqrt{\epsilon_r\mu_r}}\frac{1}{\sqrt{\epsilon_0\mu_0}} = \frac{c_0}{\sqrt{\epsilon_r\mu_r}} \tag{4.1}$$

where c_0 is the speed of light in a vacuum. The constant $\sqrt{\epsilon_r\mu_r}$ is commonly given the name the refractive index of the medium, denoted by n, so that

$$c = \frac{c_0}{n} \tag{4.2}$$

We have come across the idea that we need to have not only a wave speed but also to have a wave impedance which allows us to derive the reflection and transmission coefficients at the boundaries between media of different impedance. An EM plane wave consists of an E-field transverse to the direction of propagation and a proportional B-field at right angles to the E-field and it turns out that the impedance for an EM wave is related to the ratio between the two fields

$$Z = \frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = \frac{|\boldsymbol{E}|}{|\boldsymbol{B}|/\mu} \tag{4.3}$$

where the impedance Z is given by

$$Z = \sqrt{\mu/\epsilon} \tag{4.4}$$

In a vacuum Z has the value $Z_0 \approx 377\Omega$. Since the magnitude and direction of the B field can be deduced from the E field, by convention we always analyse EM wave motion by tracking the E field. For example, the polarisation direction of an EM wave is referred to in terms of the direction of the E-field.

Knowing the impedance of an EM wave allows us to determine the power transmitted. The power flowing through unit area, or intensity I for an EM field is given by the magnitude of the *Poynting vector*

$$I(t) = |\boldsymbol{E}(t) \times \boldsymbol{B}(t)/\mu|. \tag{4.5}$$

Since the B field is perpendicular to the E field and its magnitude is given by $B(t) = E(t)\mu/Z$ then

$$I(t) = \frac{|\boldsymbol{E}(t)|^2}{Z}. \tag{4.6}$$

which can be compared to the expressions for power in a mechanical oscillator or wave, with the E-field taking the role of force.

A generalisation that can be made about media that transmit light at optical wavelengths is that they are almost all dielectric media, and the their relative permeabilities are almost all unity i.e. $\mu_r \approx 1$. In this case the wave impedance can be directly related to the refractive index:

$$c = c_0/\sqrt{\epsilon_r} = c_0/n \tag{4.7}$$
$$Z = Z_0/\sqrt{\epsilon_r} = Z_0/n \tag{4.8}$$

Thus, at optical wavelengths, reflection and transmission formulae are usually quoted in terms of the refractive indices of the media, instead of the wave impedances. For example the amplitude reflection coefficient is given by

$$r = \frac{Z_2 - Z_1}{Z_2 + Z_1} = \frac{Z_0/n_2 - Z_0/n_1}{Z_0/n_2 + Z_0/n_1} = \frac{n_1 - n_2}{n_1 + n_2} \tag{4.9}$$

4.2 Physical optics

There are many ways of treating optical phenomena, each of them emphasising different aspects and approximating other aspects so as to allow a tractable analysis. We can arrange these as a hierarchy as to the type of approximation which is involved:

1. Quantum Electrodynamics: the full theory of E.M. interaction with matter, but too complicated for everything but simple systems
2. Maxwell's Equations: Valid when the energy of individual photons is negligible in comparison to the light intensities being examined, but the boundary conditions are complicated to compute for all but the simplest cases.
3. Physical Optics: This involves so-called scalar wave theory, i.e. we ignore polarisation (mostly), and simplify the boundary conditions. We use Huygens' construction of secondary waves to derive most results.
4. Ray Optics: Here we ignore wave properties (i.e. assume λ is much smaller than any scale of interest). This gives so-called geometric optics and is equivalent to Newton's "corpuscular theory".

We will be using physical optics approximations for most of this course. This allows us to tackle a range of problems where the system may include complex shaped apertures but where the significant features are typically larger than a wavelength.

4.3 Diffraction

Diffraction concerns the passage of a wave past some obstruction. The phenomenon is general to all waves, but is particularly important in optics. You have already been introduced to the diffraction grating last year; here we will develop a more general treatment which will show the

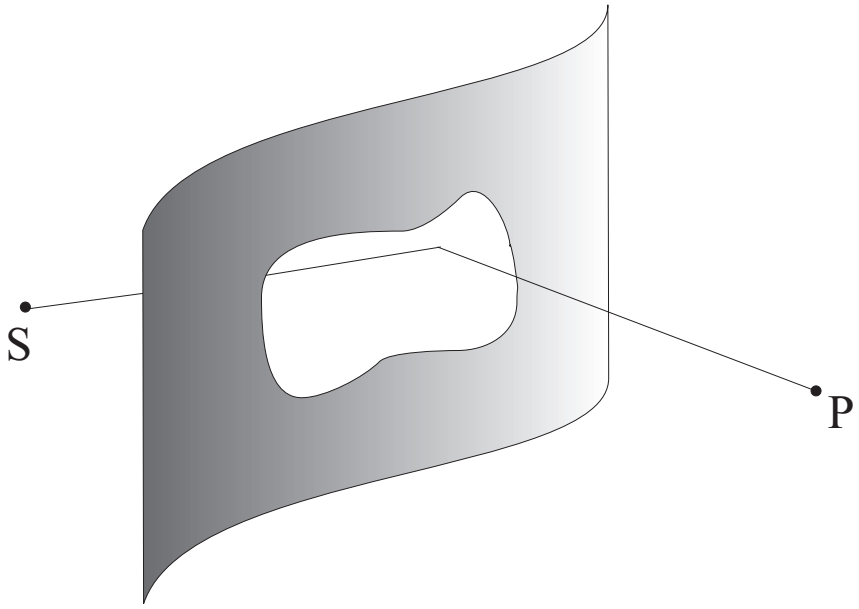


Figure 4.1: Diffraction from an aperture.

relationship to Fourier transforms, will generalise into two dimensions, and will allow us to study diffraction closer to the aperture.

The most general diffraction geometry is shown in Figure 4.1. A source of waves, S, is placed behind an aperture, and we wish to calculate the distribution of intensity as a function of position on the other side of the aperture. We will assume that the illumination produced by S is *coherent*, i.e. there is a well-defined phase relationship between the parts of the wave arriving at different points on the aperture.

4.3.1 Huygens' Principle

Huygens' principle indicates that each point on a wavefront acts as a source of secondary wavelets which propagate, overlap, interfere, and thus carry the wavefront forward (Figure 4.2). This leads to a useful construction for analysing reflection and refraction, as you saw last year. We will use the idea of wavelets to study diffraction, by considering the wavelets generated from all points on the aperture.

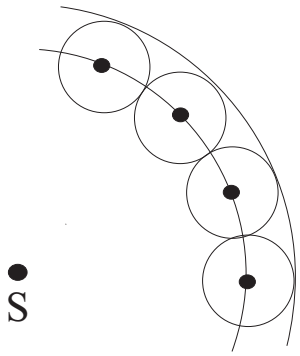
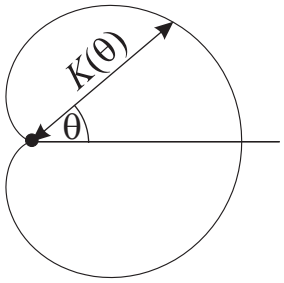


Figure 4.2: Huygens' Principle: secondary wavelets emerging from the primary wavefront originating from the source point S.

Before we go on, it is worth noting some limitations of Huygens' principle. The simple idea of spherical wavelets would lead to a backward-propagating wavefront as well as a forward-propagating one. This is not observed experimentally. To fix Huygens' principle, we have to introduce an *obliquity factor*, $K(\theta)$, which describes the fall-off in intensity of the wavelets with angle, θ, away from the forward direction. The form of $K(\theta)$ is given by

$$K(\theta) = \frac{1+\cos\theta}{2}. \tag{4.10}$$



This extension to Huygens principle is known as *Huygens-Fresnel* theory.

4.3.2 The Diffraction Integral

For simplicity we will consider a planar aperture, Σ, although the approach extends naturally to apertures of any shape. Consider an element of aperture $dx\,dy$ at position (x, y), as shown in Figure 4.3.

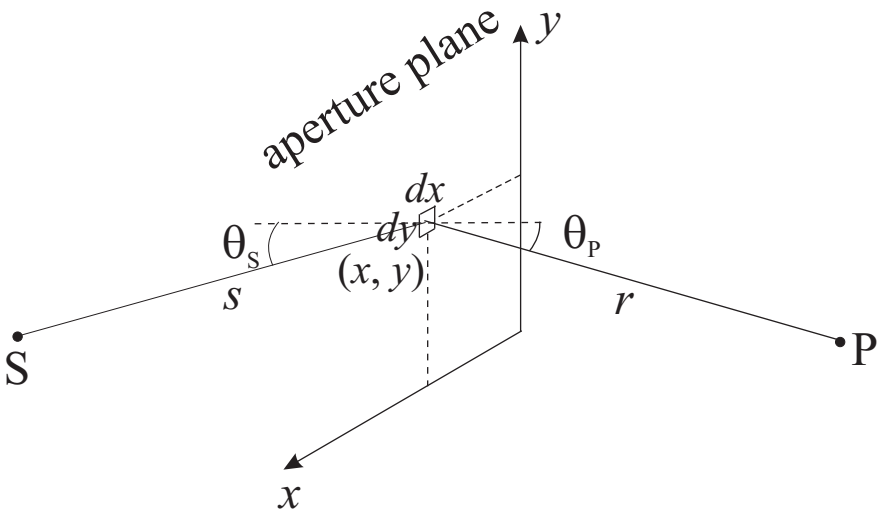


Figure 4.3: Diffraction from a planar aperture.

In most of the following we will be considering a monochromatic 3-dimensional wave. We split the wave into a spatial term $\psi(\boldsymbol{r})$ and a temporal term $e^{-i\omega t}$

$$\Psi(\boldsymbol{r}, t) = \Re\{\psi(\boldsymbol{r})e^{-i\omega t}\} \tag{4.11}$$

Since the wave is monochromatic with a known value of ω everywhere, then we can trace the propagation of the wave $\Psi(\boldsymbol{r}, t)$ through the propagation of the complex spatial phasor $\psi(\boldsymbol{r})$. This considerably simplifies the notation and the mathematics.

The source S produces spherical waves with a "strength" a_S, and is at a distance s from the element of aperture. The wave arriving at the aperture element is therefore

$$\psi_1(\boldsymbol{r}) = \frac{a_S e^{iks}}{s}. \tag{4.12}$$

The aperture (in a coordinate plane denoted by Σ) can change the amplitude or phase of the incident radiation, and for a planar aperture we can describe its transmission properties by the *aperture function*, $h(x, y)$. Usually $h = 0$ or 1 for obstructing or open areas respectively. The element of aperture can now be considered to act as a source of secondary wavelets with a strength and phase given by

$$a_\Sigma = A_0\, \psi_1(x, y)\, h(x, y)\, dx\, dy. \tag{4.13}$$

It can be shown that the constant $A_0 = -i/\lambda$ (we will not prove this here, but a derivation can be found in Lipson, Optical Physics).

Next we need to calculate the amplitude of the secondary wavelet reaching our observation point P, which is a distance r away from the aperture element. This amplitude is given by

$$\begin{aligned} d\psi_P &= -\frac{i}{\lambda}\frac{a_S\, e^{iks}}{s}\, h(x, y)\, dx\, dy\, K(\theta)\, \frac{e^{ikr}}{r} \\ &= -\frac{i}{\lambda}\, h(x, y)\, K(\theta)\, \frac{a_S\, e^{ik(s+r)}}{s\, r}\, dx\, dy. \end{aligned} \tag{4.14}$$

$K(\theta)$ here is the obliquity factor, which turns out to depend on the angles θ_S and θ_P associated with the directions of the vectors from the aperture element to points S and P, relative to the aperture normal:

$$K = \frac{\cos\theta_S + \cos\theta_P}{2}. \tag{4.15}$$

To calculate the total amplitude at point P, we finally sum over all the elements of the aperture

$$\psi_P = \iint_\Sigma -\frac{i}{\lambda}\, h(x, y)\, K(\theta_S, \theta_P)\, \frac{a_S\, e^{ik(s+r)}}{s\, r}\, dx\, dy. \tag{4.16}$$

Note that both s and r are functions of x and y.

This general result allows the amplitude of diffracted light to be calculated at any point P, although it breaks down for points very close ($< \lambda$) to the edge of an aperture, at which point the wave equation must be solved directly, taking into account the vector nature of the electromagnetic field. We will not consider these *very near-field* cases any further, and will concentrate on distances $> \lambda$.

4.4 Fraunhofer Diffraction

Although Eq. 4.16 can be evaluated numerically, we are most interested in some limiting cases where we can make approximations which allow us to solve the problem analytically, and hence to gain some physical insight into the diffraction process.

First we consider the diffraction pattern in a plane at a distance L from the aperture (Figure 4.4). We denote the coordinates of point P in this plane by (x_0, y_0). We will assume that the source S is a large distance behind the aperture (and centred on the aperture) so that $s \to \infty$, and the aperture

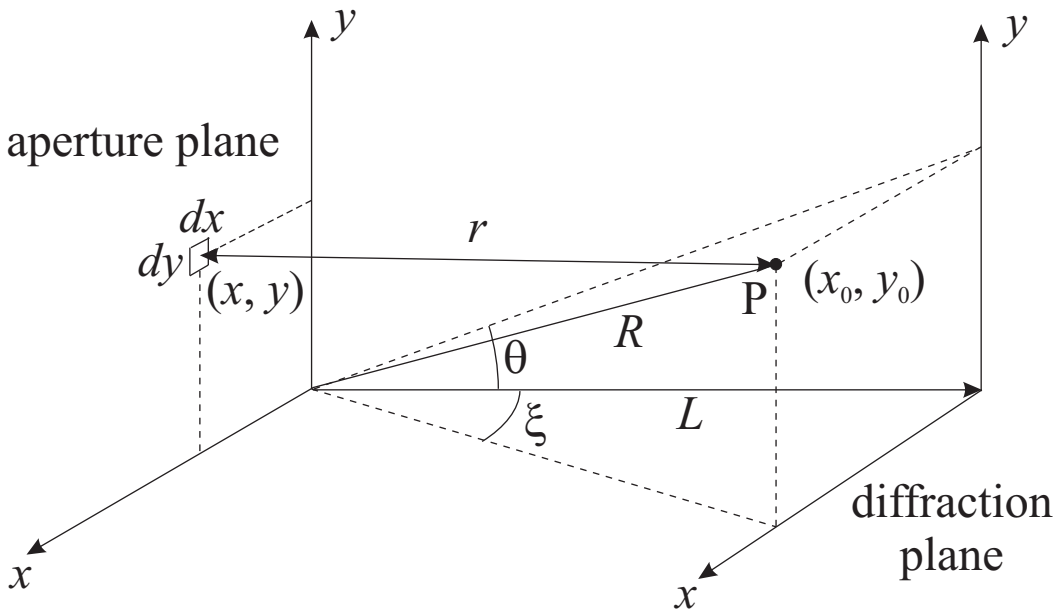


Figure 4.4: Geometry for Fraunhofer diffraction.

is illuminated with a plane wave at normal incidence. We will consider points P where x_0/L and y_0/L are sufficiently small that we can assume that the obliquity factor $K = 1$. Using the coordinate system shown in Figure 4.4, we find that the distance r from the element of aperture $dx\,dy$ to point P is given by

$$\begin{aligned} r^2 &= L^2 + (x_0 - x)^2 + (y_0 - y)^2 \\ &= L^2 + x_0^2 + y_0^2 - 2(x_0 x + y_0 y) + x^2 + y^2 \\ &= R^2 - 2(x_0 x + y_0 y) + x^2 + y^2 \\ &= R^2 \left(1 - 2\frac{x_0 x + y_0 y}{R^2} + \frac{x^2 + y^2}{R^2}\right) \end{aligned}$$

where $R^2 = L^2 + x_0^2 + y_0^2$. Using a binomial expansion, we obtain

$$r \approx R - \frac{x_0 x + y_0 y}{R} + \frac{x^2 + y^2}{2R} \tag{4.17}$$

The phase of each wavelet goes as kr, hence the phase will in general have terms which vary linearly and quadratically with the position (x, y) in the aperture. If we make L (and hence R) large enough, we can arrange that the quadratic term in the phase is negligible:

$$\frac{k(x^2 + y^2)}{2R} \ll \pi \tag{4.18}$$

for all elements in the aperture. This is the condition for *Fraunhofer diffraction*. The wave amplitude at P is then given by the *Fraunhofer integral*

$$\psi_P \propto \iint_\Sigma \psi_\Sigma\, h(x, y)\, e^{\left(-ik\left(\frac{x_0 x + y_0 y}{R}\right)\right)}\, dx\, dy \tag{4.19}$$

where ψ_Σ = constant for illumination of the aperture with a plane wave at normal incidence. Note we have dropped the prefactor in the integral since we are mostly interested in the shape of the diffraction pattern rather than its absolute intensity.

A 1-D aperture consists of patterns which are extended in the x direction, such as slits or diffraction gratings. In this case, the integral over x just gives a multiplicative constant, and the Fraunhofer

integral becomes

$$\psi_P \propto \int_\Sigma h(y)\, e^{-ik\frac{y_0 y}{R}}\, dy. \tag{4.20}$$

For large L, angles are small, hence $\sin\theta \approx \theta \approx y_0/R$, giving

$$\psi_P \propto \int h(y)\, e^{-iky\sin\theta}\, dy = \int h(y)\, e^{-iqy}\, dy \tag{4.21}$$

with $q = k\sin\theta$.

So, we can see that ψ_P (as a function of y_0, or of angle θ) is related to $h(y)$ by a Fourier transform. Specifically

$$\psi_P(q) \propto \mathrm{FT}\{h(y)\}. \tag{4.22}$$

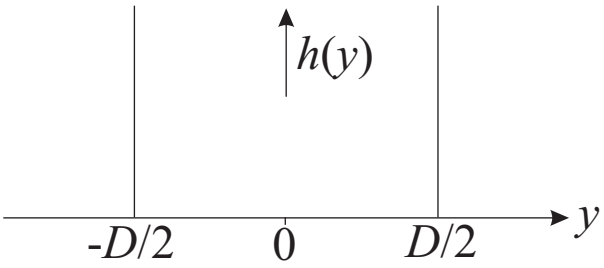
y and q are *reciprocal coordinates* which play the same role as time and frequency in the Fourier transforms you have seen previously. By analogy, you can see that the spatial frequencies present in the diffraction pattern are determined by the form of the aperture function (and vice versa).

4.4.1 Some simple examples

(a) 2 narrow slits (Young's slits)

Consider an aperture consisting of 2 narrow (δ-function) slits a distance D apart, spaced evenly about the origin

$$h(y) = \delta(y + D/2) + \delta(y - D/2). \tag{4.23}$$



Using Eq. 4.21 gives

$$\psi_P = \int \left(\delta(y + D/2) + \delta(y - D/2)\right) e^{-iqy}\, dy \tag{4.24}$$

Remembering that

$$\int \delta(x - x_0) f(x)\, dx = f(x_0) \tag{4.25}$$

we obtain

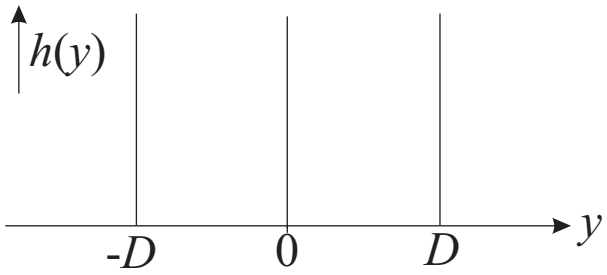
$$\psi_P(q) \propto \left(e^{-iqD/2} + e^{iqD/2}\right) = 2\cos\left(\frac{qD}{2}\right). \tag{4.26}$$

The intensity therefore varies as

$$I_P(q) = I_0 \cos^2(qD/2). \tag{4.27}$$

Note that the smaller the spacing between the slits, the larger the spacing between maxima in the diffraction pattern.

(b) 3 narrow slits

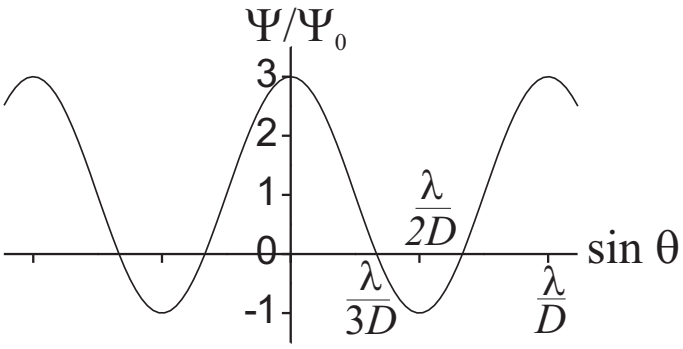


Following the same approach, if

$$h(y) = \delta(y + D) + \delta(y) + \delta(y - D) \tag{4.28}$$

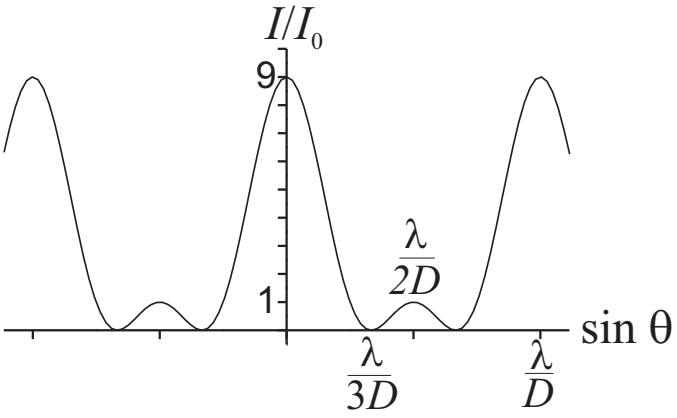
then

$$\psi_P \propto e^{iqD} + 1 + e^{-iqD} = 1 + 2\cos(qD) \tag{4.29}$$



so

$$I_P(q) = I_0(1 + 2\cos(qD))^2. \tag{4.30}$$



The diffraction intensity has a regular pattern showing one subsidiary maximum between the primary maxima.

You will recall from last year that Eq. 4.29 can be visualised using a *phasor diagram* (Figure 4.5).

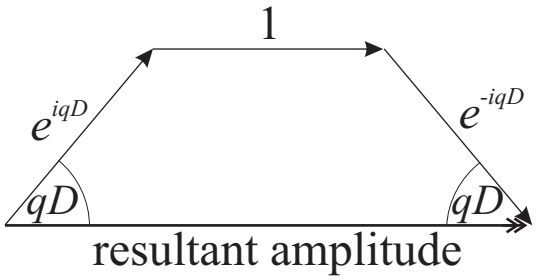


Figure 4.5: Phasor diagram for diffraction by 3 narrow slits.

(c) N narrow slits - a grating

The aperture function for N narrow slits with spacing D is

$$h(y) = \sum_{m=0}^{N-1} \delta(y - mD) \tag{4.31}$$

hence

$$\psi_P \propto 1 + e^{-iqD} + e^{-i2qD} + e^{-i3qD} + \ldots + e^{-i(N-1)qD}. \tag{4.32}$$

This is the sum of a simple geometric series, giving

$$\psi_P \propto \frac{1 - e^{-iNqD}}{1 - e^{-iqD}} = e^{i(N-1)qD/2} \frac{\sin(NqD/2)}{\sin qD/2} \tag{4.33}$$

so

$$I_p = I_0 \left[\frac{\sin(NqD/2)}{\sin qD/2} \right]^2. \tag{4.34}$$

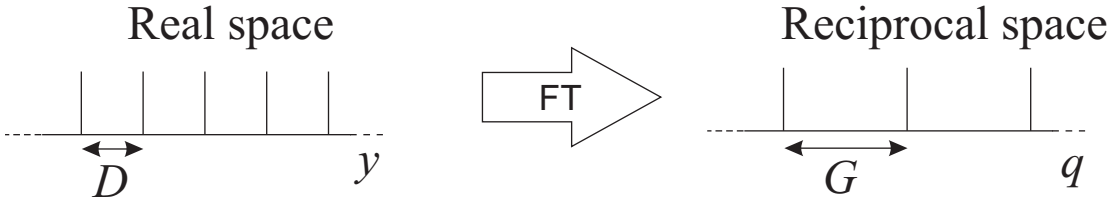
This function has primary maxima when $\sin(qD/2) = 0$, i.e. when $q = 2m\pi/D$ where m is an integer. It also has subsidiary maxima whenever $\sin(NqD/2) = \pm 1$, i.e. where $q = (2M + 1)\pi/(ND)$, where M is an integer. For a grating with N slits, there will be $N - 2$ subsidiary maxima and $N - 1$ zeros between primary maxima. The larger the value of N, the larger the ratio between primary maxima and subsidiary maxima.

When light of a single wavelength is incident on a grating with a large number of slits, diffraction will therefore occur into well-defined directions defined by

$$k \sin\theta = q = 2m\pi/D = mG. \tag{4.35}$$

where $G = 2\pi/D$.

Schematically (for $N \to \infty$)



For a grating with N slits, the width of the primary maxima is given by the position of the zero of $\sin(NqD/2)$, i.e. at

$$q = \frac{2\pi}{ND} = \frac{G}{N}. \tag{4.36}$$

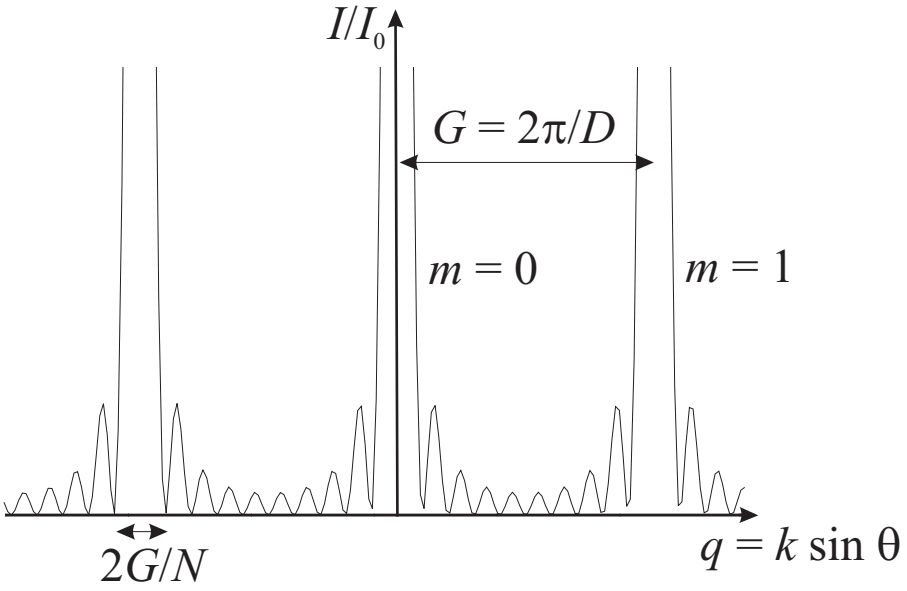


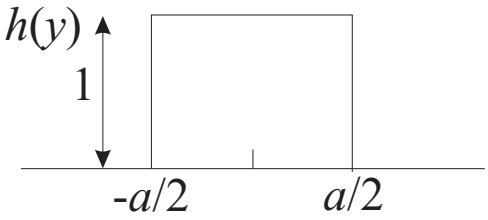
Figure 4.6: Diffracted intensity from a finite grating ($N = 10$). Note that the primary maximum (cut off in the graph) is a factor of ~ 40 larger than the first subsidiary maximum.

Increasing the width of the grating, and hence N, will therefore give narrower diffraction peaks. Due to the finite width of the diffraction peaks, illumination at a single wavelength will produce diffraction at a range of angles. This width will limit the *chromatic resolving power* of the grating - its capacity to separate waves of different wavelengths. Grating spectrometers will be discussed in more detail in Section 4.4.4.

(d) Single wide aperture

The aperture function does not necessarily need to consist of a series of δ-functions. For example, we can find the diffraction pattern of a single slit of width a.

$$\begin{aligned} h(y) &= 1 \text{ for } |y| < a/2 \\ &= 0 \text{ for } |y| > a/2. \end{aligned}$$



The diffraction amplitude is

$$\begin{aligned} \psi_P(q) \propto \int_{-a/2}^{a/2} e^{-iqy} dy &= \frac{-1}{iq}\left[e^{-iqy}\right]_{-a/2}^{a/2} \\ &= \frac{a \sin(qa/2)}{qa/2} \\ &= a \,\mathrm{sinc}\left(\frac{qa}{2}\right). \end{aligned} \tag{4.37}$$

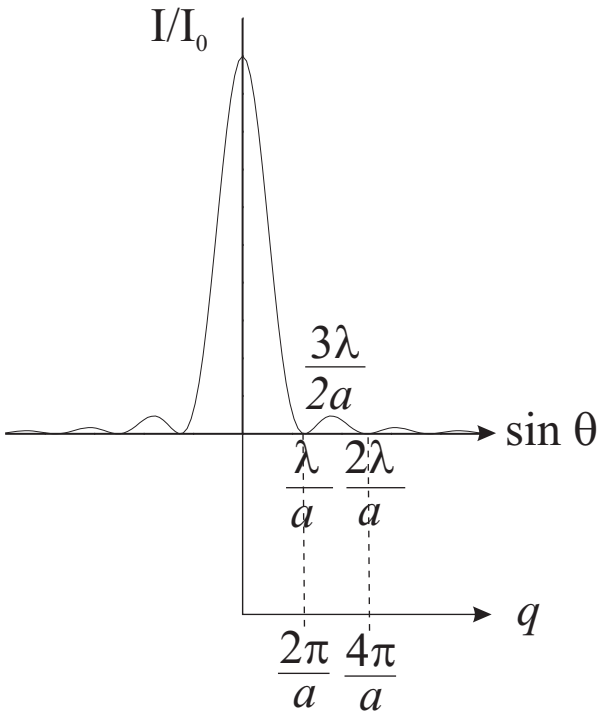


Figure 4.7: Diffracted intensity for a single slit.

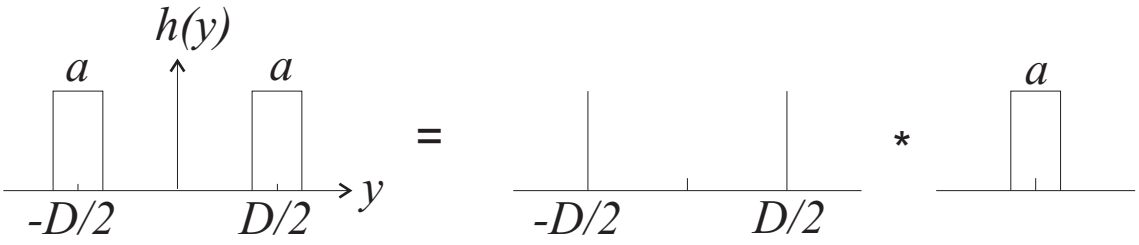
So the intensity observed on the screen is

$$I_p(q) \propto a^2 \,\mathrm{sinc}^2\left(\frac{qa}{2}\right) \tag{4.38}$$

where, as before, $q = k\sin\theta = (2\pi/\lambda)\sin\theta$.

4.4.2 Complicated apertures

We have already shown that the diffraction pattern in the Fraunhofer limit is given by the Fourier transform of the aperture function. To analyse complex apertures, we can therefore make use of the theorems of Fourier theory seen earlier. One example is the diffraction pattern of two wide slits. The aperture function is the convolution of two delta functions (two narrow slits) with a top-hat function (wide aperture).



The diffraction pattern will simply be the product of the Fourier transforms of these two functions, each of which we have evaluated previously. For two slits of width a with their centres separated by a distance D, the diffraction pattern is the product of Eq. 4.27 with Eq. 4.38.

$$\psi_P \propto \cos\left(\frac{qD}{2}\right) a\,\mathrm{sinc}\left(\frac{qa}{2}\right) \tag{4.39}$$

The $\cos^2$ fringes from the two slits are now modulated by a sinc^2 envelope function due to the finite width of the slits. In the example shown in Figure 4.8, with $D = 3a$, the third fringe is almost completely eliminated by the zero in the envelope function. In general, this leads to *missing orders* where a maximum expected in the underlying function coincides with a minimum in the envelope function.

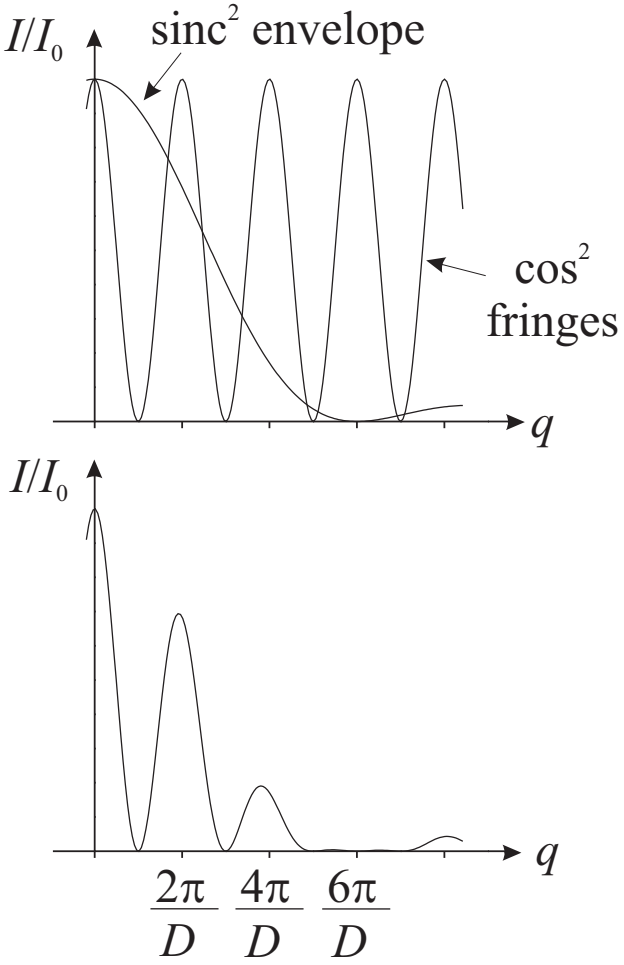


Figure 4.8: $\cos^2$ fringes modulated by a sinc^2 envelope, for $D = 3a$.

We can apply a similar approach to a more complicated aperture: a diffraction grating of finite width, with slits of non-zero width. The ideas are shown schematically in Figure 4.9

4.4.3 Conditions for observing Fraunhofer diffraction

So far we have considered situations where the source and the observation plane are a large distance from the aperture. This means that the waves arriving at the aperture and at the observation screen can be considered to be plane waves, i.e. the radius of curvature of the wavefronts is negligible. In many situations, however, it is inconvenient to have to operate at such large distances from the aperture, so alternative arrangements need to be made to reach the Fraunhofer limit. The most obvious of these is to place lenses either side of the aperture to convert the diverging beam from the source into a parallel beam, and to focus the diffracted parallel beams onto an observation screen (Figure 4.10). Clearly the source and observation screen must be placed at the focal points of the lenses.

Fraunhofer conditions can also be achieved using a single lens as shown in Figure 4.11. The source

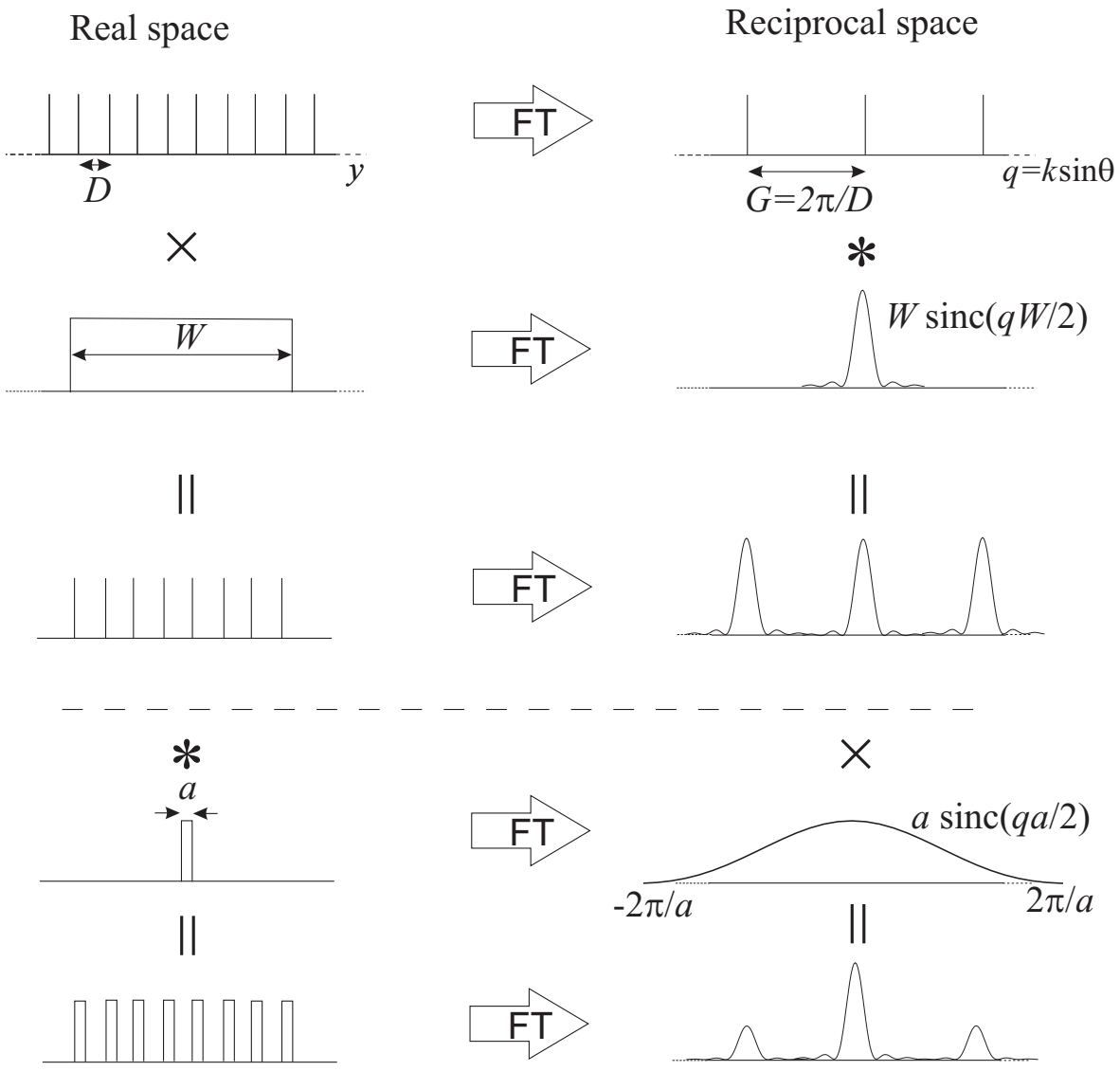


Figure 4.9: Diffraction from a grating of finite width with finite slit size.

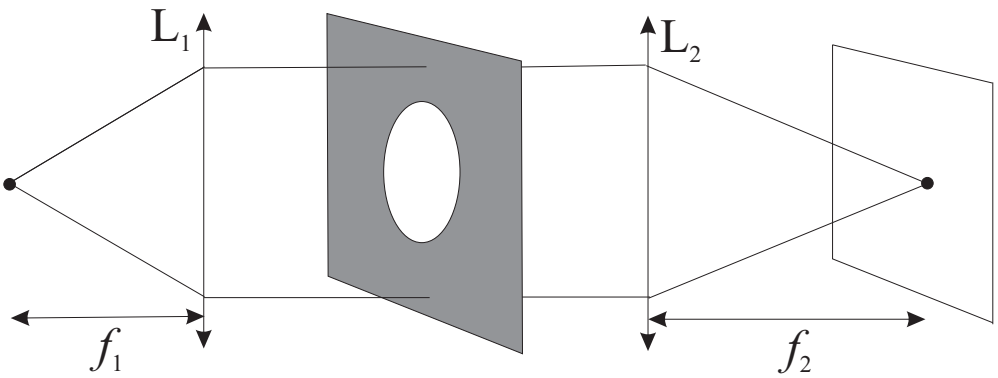


Figure 4.10: Use of lenses to produce Fraunhofer far-field conditions for diffraction.

and screen are placed in the object and image planes, i.e. S and P are *conjugate* points. An aperture placed anywhere along the optical axis will produce a Fraunhofer diffraction pattern in the image plane, since it turns out that the optical paths to point P' vary approximately linearly with position in the aperture.

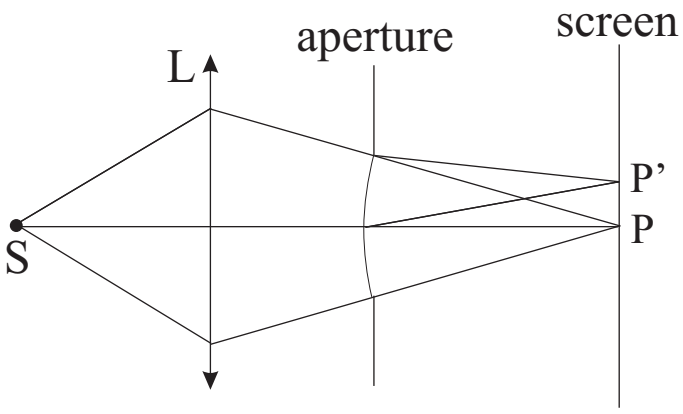


Figure 4.11: Use of a single lens to produce Fraunhofer far-field conditions.

4.4.4 Grating spectrometers

The technique shown in Figure 4.10 is employed in the *grating spectrometer*, where a diffraction grating is used to measure the spectrum of incident light (Figure 4.12). In this case, concave mirrors are often used rather than lenses (since they suffer less chromatic abberation).

Incident light is focussed onto a narrow entrance slit before encountering the mirror, thus producing a well-defined angle of incidence, θ_1 on the grating. Light is diffracted through an angle which depends on its wavelength, and on the spacing of lines in the grating, according to

$$D(\sin\theta_2 - \sin\theta_1) = m\lambda \tag{4.40}$$

The second mirror and exit slit select a narrow range of angles of diffracted light, corresponding to a narrow range of wavelengths, thus the relative intensity of light in this wavelength range can be measured. The grating is then rotated (changing both θ_1 and θ_2) in order to measure the entire spectrum. At any given angle, a number of different wavelengths may satisfy the diffraction condition, corresponding to different *orders* i.e. different values of m. This is inconvenient since the detector itself cannot distinguish between the different orders, hence it is often necessary to use filters to remove higher order diffractions in order to obtain an accurate spectrum from a broadband source. The *spectrograph* works on a similar principle, except that the exit slit is removed and

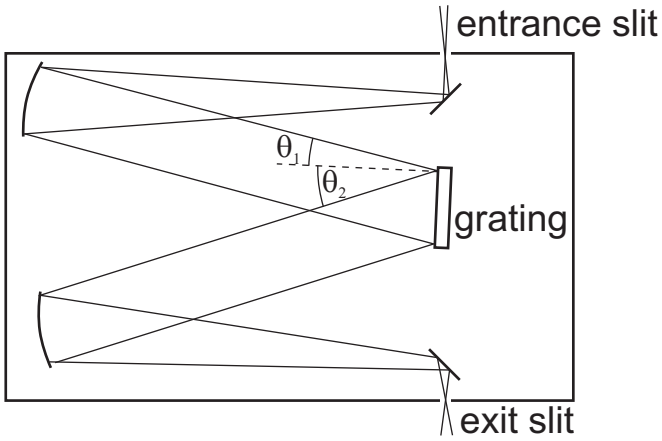


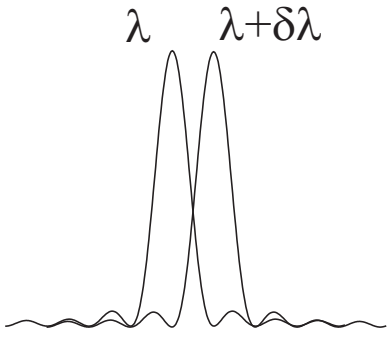
Figure 4.12: A grating spectrometer.

the grating angle is fixed. The different wavelengths are dispersed spatially in the exit plane, and nowadays are typically measured in parallel using an array detector.

The resolution that can be achieved in a spectrometer is limited in most cases by the width of the entrance and exit slits (which should be equal for best performance). The range of angles incident on and collected from the grating is proportional to the width of the slits and inversely proportional to the focal length of the mirrors. Differentiating Eq. 4.40 shows that this range of angles corresponds to a range of wavelengths which will be collected by the detector, thus limiting the resolution. Using narrower slits is an obvious way to improve the resolution, although this also reduces the amount of light that can be coupled into the spectrometer, causing problems where the source is of low intensity.

With sufficiently small slits, the resolution of the spectrometer will eventually become limited by the diffraction properties of the grating, as introduced in Section 4.4.1. Light of a particular wavelength λ will now produce a well-defined angle of incidence, but for a grating of finite width the peaks in the diffraction pattern will have a finite width. If we assume for simplicity that $\theta_1 = 0$, then we have $D \sin\theta = m\lambda$, where $\theta = \theta_2$. For illumination at two wavelengths λ and $\lambda + \delta\lambda$, diffraction peaks will be produced at angles θ_λ and $\theta_{\lambda+\delta\lambda}$, where

$$D \sin\theta_\lambda = m\lambda \quad \text{and} \quad D \sin\theta_{\lambda+\delta\lambda} = m(\lambda + \delta\lambda) \tag{4.41}$$



For λ, the first minimum in the diffracted beam profile (from Eq. 4.36) occurs at

$$\sin\theta = \frac{m\lambda}{D} + \frac{\lambda}{W} \tag{4.42}$$

where $W = (N-1)D \approx ND$ is the width of the grating.

The wavelengths can be resolved if the maximum for $\lambda + \delta\lambda$ coincides with this minimum for λ - the *Rayleigh Criterion*, i.e. if

$$m\lambda + \frac{\lambda D}{W} = m(\lambda + \delta\lambda) \tag{4.43}$$

so that

$$\frac{\lambda}{\delta\lambda} = \frac{mW}{D} = mN \tag{4.44}$$

is the *chromatic resolving power* of the grating. This is a measure of how well the grating separates light of two wavelengths.

$\delta\lambda$ is the minimum wavelength difference (near λ) which the grating can effectively distinguish.

Notes:

The greater the number N of slits/lines, the higher the resolving power - wider gratings give narrower diffraction beams which are easier to distinguish.

The order m acts as a kind of "gearing" - the widths of the diffracted beams are fixed (by W), but their angular separation increases with m, so it is easier to distinguish wavelengths at higher order.

4.4.5 Two-dimensional apertures

We have already seen how to calculate the diffraction pattern for a two-dimensional aperture.

$$\psi_P \propto \iint_\Sigma \psi_\Sigma \, h(x,y) \, e^{\left(-ik\left(\frac{x_0 x + y_0 y}{R}\right)\right)} \, dx \, dy. \tag{4.45}$$

We now define the diffraction pattern in terms of two angles, θ and ξ, both of which are small, giving $\sin\theta \approx \theta \approx y_0/R$ and $\sin\xi \approx \xi \approx x_0/R$. Writing $q = k\sin\theta$ and $p = k\sin\xi$, the Fraunhofer integral then becomes

$$\psi_P \propto \iint_\Sigma h(x,y) \, e^{-i(px+qy)} \, dx \, dy. \tag{4.46}$$

This is the two-dimensional Fourier transform of $h(x,y)$.

This is particularly easy to evaluate if $h(x,y)$ is separable, i.e. if $h(x,y) = f(x)g(y)$.

Rectangular aperture

Here the aperture function is separable:

$$\begin{aligned} h(x,y) &= 1, \ |x| < a/2 \text{ and } |y| < b/2 \\ &= 0, \ \text{otherwise.} \end{aligned}$$

The diffraction pattern therefore becomes

$$\begin{aligned} \psi_P &\approx \int_{-a/2}^{a/2} e^{-ipx} \, dx \int_{-b/2}^{b/2} e^{-iqy} \, dy \\ &= a \ \text{sinc}\left(\frac{pa}{2}\right) b \ \text{sinc}\left(\frac{qb}{2}\right). \end{aligned} \tag{4.47}$$

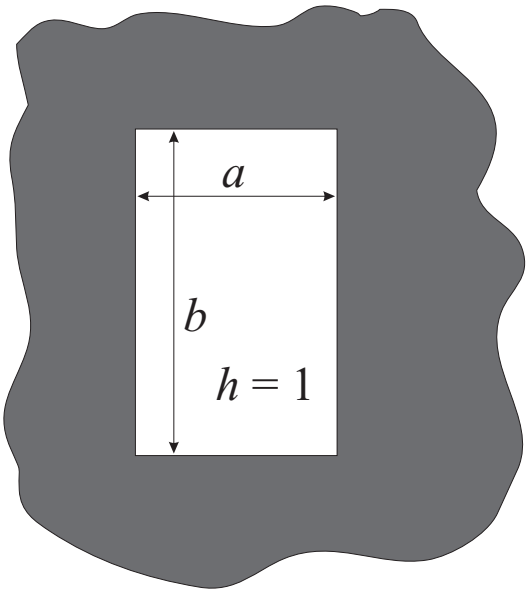


Figure 4.13: Rectangular aperture.

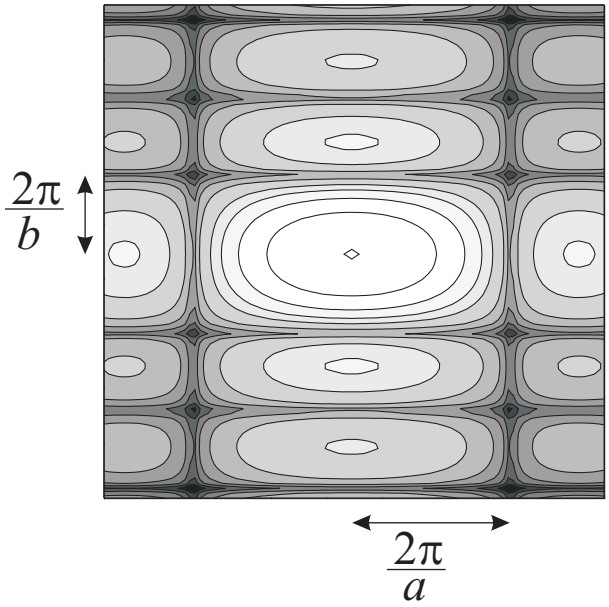


Figure 4.14: Diffracted intensity from a rectangular aperture. Logarithmic shading is used for clarity.

Circular Aperture

This example is important since many apertures in optical systems (telescopes, microscopes, cameras, etc.) are circular. Here h is a function of radius, ρ, and for an aperture of diameter d

$$\begin{aligned} h(\rho) &= 1 \text{ æ} < d/2 \\ &= 0 \text{ otherwise.} \end{aligned}$$

h is not separable in x and y, so the mathematics is more complicated. The result obtained is

$$\psi(\theta) \propto \frac{\psi_0 d^2}{2} \frac{J_1\left(\frac{kd\sin\theta}{2}\right)}{\left(\frac{kd\sin\theta}{2}\right)} \tag{4.48}$$

where J_1 is the 1st order Bessel Function of the first kind. The diffracted intensity has a minimum at $\sin\theta = 1.22\lambda/d$, and the region inside this first minimum is known as the *Airy disc*.

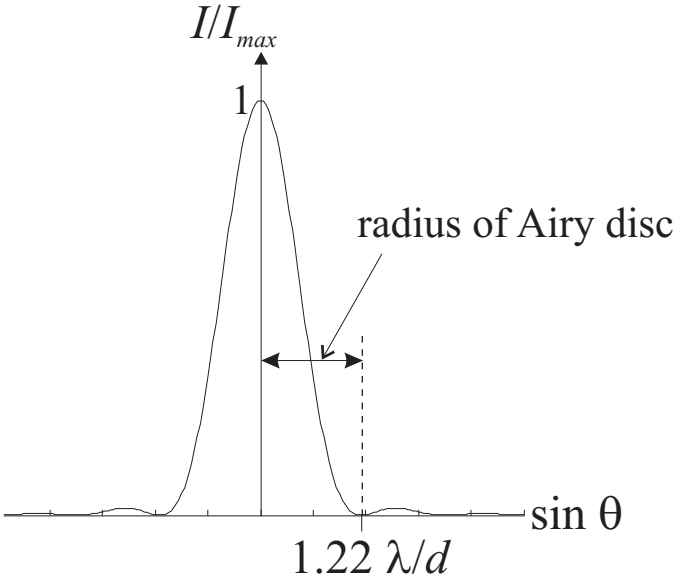


Figure 4.15: The function which describes the *Airy disc*, an example of a *point spread function*.

4.4.6 Resolution of Optical Instruments

Angular resolution of a telescope

In geometrical optics, you saw last year that a perfect lens should image a point object to a perfectly sharp point image. However, this approach neglects diffraction effects. A real lens (or mirror) in a telescope has a finite diameter, and rays which do not pass though the lens will not be focussed. This is equivalent to placing a circular aperture around the lens. There may also be other apertures within the system which obstruct the wavefronts. We saw earlier that any aperture placed between the object and image planes will produce a Fraunhofer diffraction pattern. The incoming (plane) wavefront from a distant point object will thus produce an Airy disc in the image plane, with angular radius $1.22\lambda/D$ where D is the diameter of the objective lens/mirror (assuming that there are no other apertures within the system which further obstruct the wavefronts). The actual radius in the image plane will be $\frac{1.22\lambda}{D}f$ where f is the focal length of the lens.

The Rayleigh criterion can now be applied to obtain the *angular resolution* of the telescope. Two distant point objects (e.g. stars) with angular separation α will produce Airy discs in the image plane with their centres separated by a distance $f\alpha$. These objects can only be resolved if the radius of the Airy discs is less than their separation, hence the angular resolution is $\frac{1.22\lambda}{D}$. Clearly, this neglects all other "geometrical" aberrations, so the "*diffraction limited*" image is the best that can be achieved with a particular value of D. The larger the value of D, the sharper are the images that can be obtained.

Some examples:

Unaided human eye: iris diameter: $D \approx 5\,\text{mm}$: visible light: $\lambda \approx 400\,\text{nm}$ (in aqueous humour): $\alpha \approx 8 \times 10^{-5}\,\text{rad}$.

Jodrell Bank radio telescope: metal dish reflector: $D \approx 64\,\text{m}$: radio waves: $\lambda > 60\,\text{mm}$: $\alpha > 9 \times 10^{-4}\,\text{rad}$.

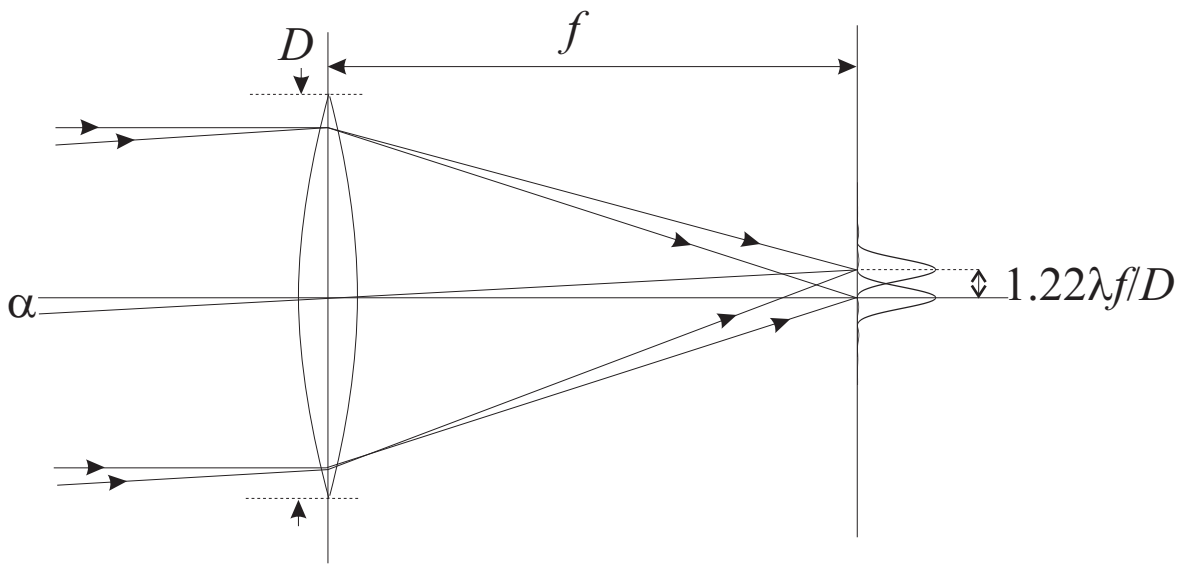


Figure 4.16: Resolution of a telescope.

Arecibo radio telescope: metal dish reflector: $D \approx 305\,\text{m}$: radio waves: $\lambda > 210\,\text{mm}$: $\alpha > 7 \times 10^{-4}\,\text{rad}$.

Keck telescope: optical reflector: $D \approx 8\,\text{m}$: visible light: $\lambda \approx 500\,\text{nm}$: $\alpha \approx 6 \times 10^{-8}\,\text{rad}$.

Spatial resolution of a microscope

Similar considerations determine the *spatial resolution* of a microscope. Here the object is close to the focal point, and the image is formed at large distances. Due to diffraction effects, the angular radius of a point object is $1.22\lambda/D$ as above, where D is the diameter of the objective lens. Two points image separated by Δx in the object will produce beams with an angular separation of $\Delta x/f$, where f is the focal length of the objective lens. Equating these quantities, we find that the separation of two objects which can just be resolved is $\frac{1.22\lambda}{D} f$.

4.4.7 Babinet's Principle

Consider two *complementary apertures* a and b.

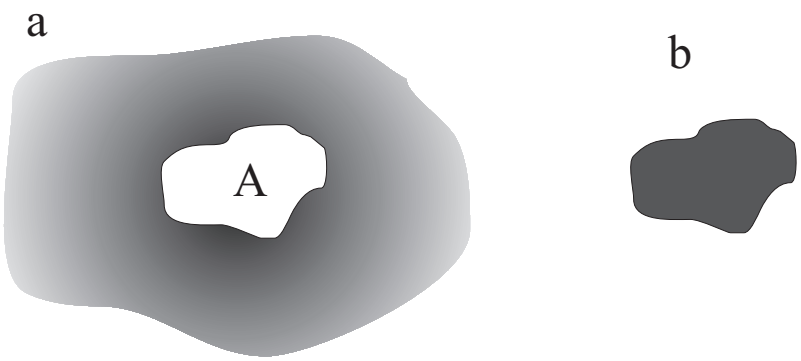


Figure 4.17: Complementary apertures.

The diffracted amplitudes are

$$\psi_a \propto \iint_A e^{-i(px+qy)}\,dx\,dy \tag{4.49}$$

$$\begin{aligned}\psi_b &\propto \iint_{\text{all space}} e^{-i(px+qy)}\,dx\,dy - \iint_{\text{A}} e^{-i(px+qy)}\,dx\,dy \\ &\propto \delta(p,q) - \psi_a. \end{aligned} \tag{4.50}$$

The diffraction intensities are therefore *the same*, except at the origin ($p = 0$, $q = 0$), the direction corresponding to the direction of the incident beam.

4.5 Fresnel Diffraction

Now let us consider a source and screen which are not at a large distance from the aperture. The Fraunhofer condition of linear phase variation with distance across the aperture no longer applies. We saw from Eq. 4.18 that Fraunhofer conditions occur apply when

$$\frac{k(x^2+y^2)}{2R} \ll \pi.$$

For an aperture with maximum dimensions $x_{\max}$ and $y_{\max}$ this is equivalent to requiring that the distance R of the observation screen from the aperture being given by

$$R \gg \rho^2/\lambda \tag{4.51}$$

where $\lambda = 2\pi/k$ and $\rho^2 = x_{\max}^2 + y_{\max}^2$, i.e. ρ is a measure of the maximum dimension of the aperture. The distance ρ^2/λ marks an approximate transition point where the quadratic phase terms in the diffraction integral cannot be ignored and is called the **Rayleigh distance**. For distances much greater than the Rayleigh distance, Fraunhofer diffraction applies, but for distances comparable to or less than the Rayleigh distance we need to do more complex calculations to include the quadratic and possibly higher-order terms in the phase of the integrand in the diffraction integral.

We can make some progress under these conditions by considering a special case, and examining the diffracted intensity at points P which are on the axis defined by the source S and the origin point O in the aperture. Under these conditions, the linear term in the phase will vanish. Calculating the diffraction intensity along a single line may sound uninteresting, but we will soon see that for simple apertures we can do much more than this by changing our choice of origin, O, in the aperture.

The path from S to P via the aperture element at (x, y) is

$$\begin{aligned} r_1 + r_2 &= \sqrt{a^2+x^2+y^2} + \sqrt{b^2+x^2+y^2} \\ &= a + b + \frac{x^2+y^2}{2a} + \frac{x^2+y^2}{2b} \\ &\quad + \text{higher order terms} \end{aligned}$$

Writing

$$\frac{1}{R} = \frac{1}{a} + \frac{1}{b} \tag{4.52}$$

we obtain

$$\text{optical path} = \text{const.} + \frac{x^2+y^2}{2R} \tag{4.53}$$

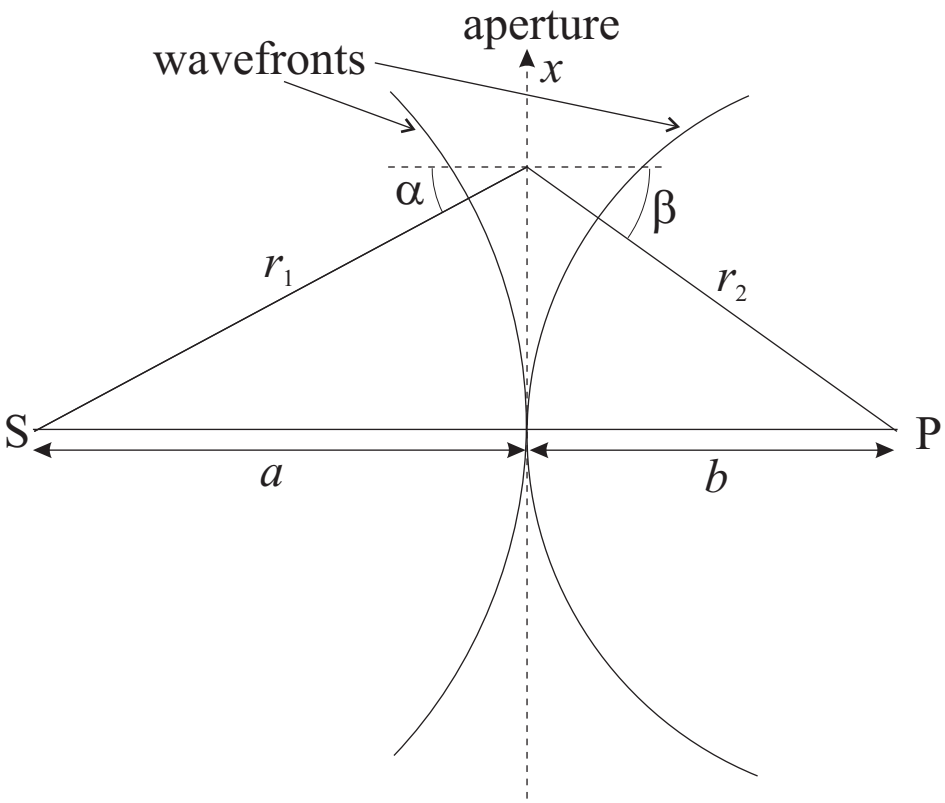


Figure 4.18: Geometry for Fresnel diffraction.

giving a diffracted intensity of

$$\psi_P \propto \iint_\Sigma \frac{h(x,y)K(\theta)\exp\left(ik\frac{x^2+y^2}{2R}\right)}{r_1 r_2}\,dx\,dy. \tag{4.54}$$

We now assume

(a) that the angles to the edge of the aperture are small enough that we can neglect the obliquity factor, and thus take $K(\theta) = 1$.

(b) that the variations in r_1 and r_2 over the aperture are negligible as far as the denominator is concerned: $r_1 \sim a$ and $r_2 \sim b$. NB. variations in $r_1 + r_2 - a - b$ are *not* negligible compared with λ, and so have a significant effect on the phase term.

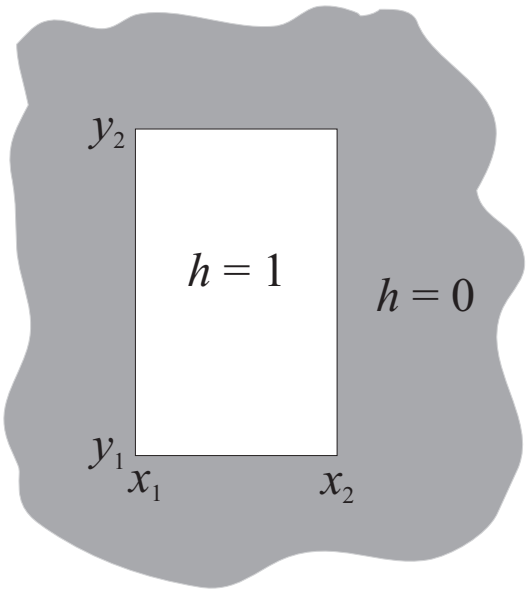
Under these approximations

$$\psi_P \propto \iint_\Sigma h(x,y)\exp\left(ik\frac{x^2+y^2}{2R}\right)\,dx\,dy. \tag{4.55}$$

4.5.1 Separable spatial variables

Consider the case of a uniform rectangular aperture, with $h(x,y)$ such that

The diffraction integral 4.55 is then



$$\begin{aligned}\psi_P &\propto \iint_{\Sigma} h(x,y)\exp\left(ik\frac{x^2+y^2}{2R}\right)dx\,dy \\ &\propto \int_{x_1}^{x_2}\exp\left(\frac{ikx^2}{2R}\right)dx\int_{y_1}^{y_2}\exp\left(\frac{iky^2}{2R}\right)dy.\end{aligned}$$

Changing to scaled variables u and v

$$u = x\sqrt{\frac{2}{\lambda R}}\;;\; v = y\sqrt{\frac{2}{\lambda R}} \tag{4.56}$$

the previous result can then be expressed as

$$\psi_P \propto \int_{u_1}^{u_2}\exp\left(\frac{i\pi u^2}{2}\right)du\int_{v_1}^{v_2}\exp\left(\frac{i\pi v^2}{2}\right)dv. \tag{4.57}$$

It is convenient to define the *Fresnel Integral*

$$\int_0^w \exp\left(\frac{i\pi u^2}{2}\right)du = C(w) + iS(w). \tag{4.58}$$

This integral cannot be evaluated analytically, but can conveniently be represented graphically in the complex plane. A particular value of w determines $C(w)$ and $S(w)$, which gives a point in the complex plane. The locus of these points, as shown in Figure 4.19, is known as the *Cornu spiral*, where w determines the distance from the origin measured *along* the curve.

Note that

$$C(\infty) = 0.5 \quad S(\infty) = 0.5 \tag{4.59}$$

and the curve spirals in very slowly towards this point.

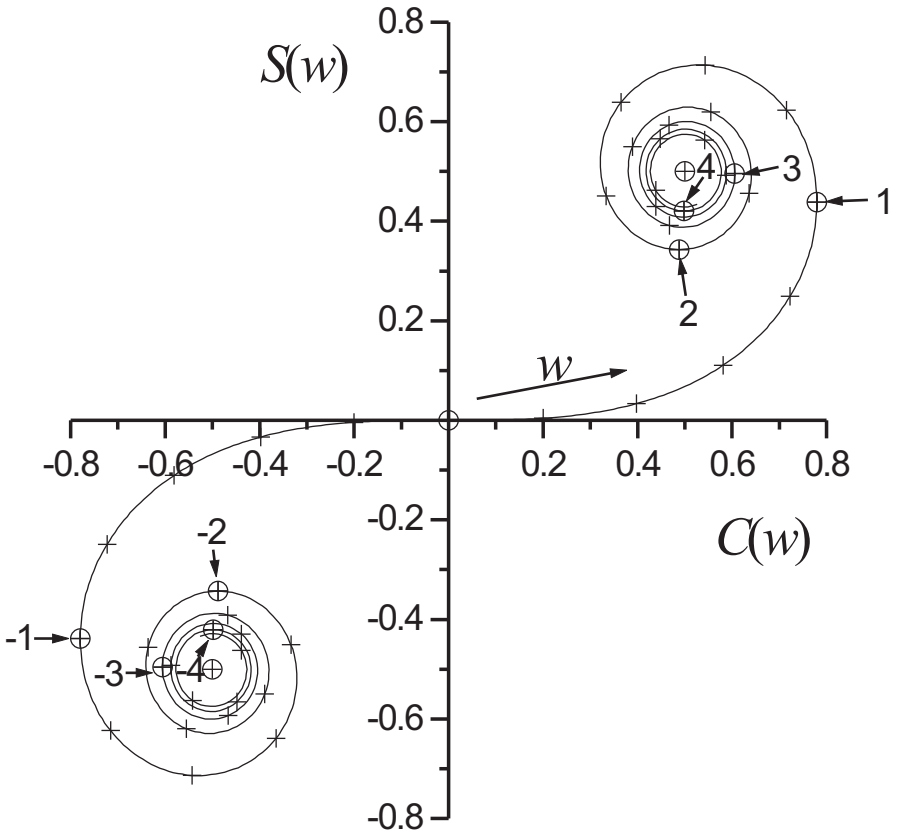


Figure 4.19: The Cornu spiral.

Positions x and y in the aperture can be either positive or negative with respect to the origin, hence negative values of w are also of interest. We note that

$$C(-w) = -C(w)\,; \;\; S(-w) = -S(w). \tag{4.60}$$

We will now see how to use the Cornu spiral to determine the diffraction pattern from simple apertures. We consider first a slit or edge extending in the y direction. The relevant diffraction integral is now just

$$\psi_P \;\propto\; \int_{w_1}^{w_2} \exp\left(\frac{i\pi u^2}{2}\right) du \tag{4.61}$$

$$= \; C(w_2) - C(w_1) + i(S(w_2) - S(w_1)) \tag{4.62}$$

with

$$w_1 = x_1\sqrt{\frac{2}{\lambda R}}\,; \;\; w_2 = x_2\sqrt{\frac{2}{\lambda R}}. \tag{4.63}$$

This result is equivalent to the vector *spanning* the Cornu spiral from point w_1 to w_2. This vector can be normalised by the amplitude resulting from an unobstructed wavefront, given by ψ_u, the spanning vector from $w = -\infty$ to $w = \infty$, which has length $\sqrt{2}$. The diffracted intensity is proportional to the square of the length of the spanning vector concerned.

4.5.2 Diffraction from a single straight edge

Consider a wavefront obstructed by a straight edge as shown in Figure 4.21.

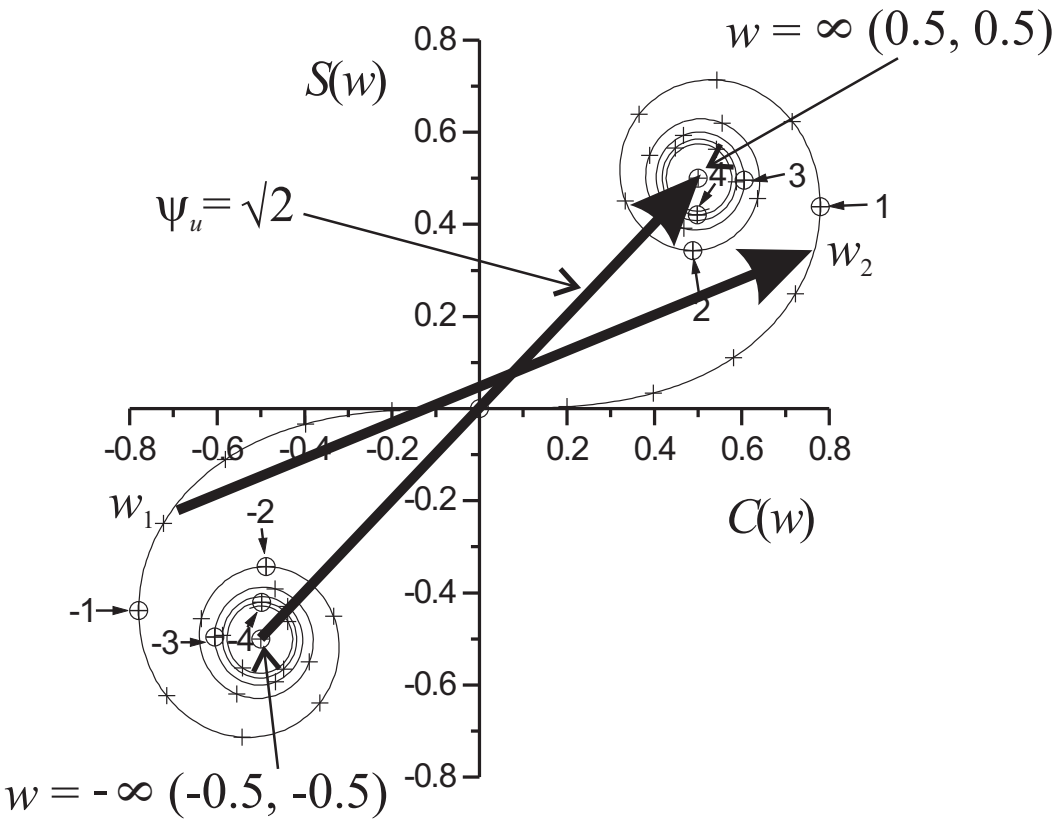


Figure 4.20: Diffraction amplitudes given by the vector spanning the Cornu spiral.

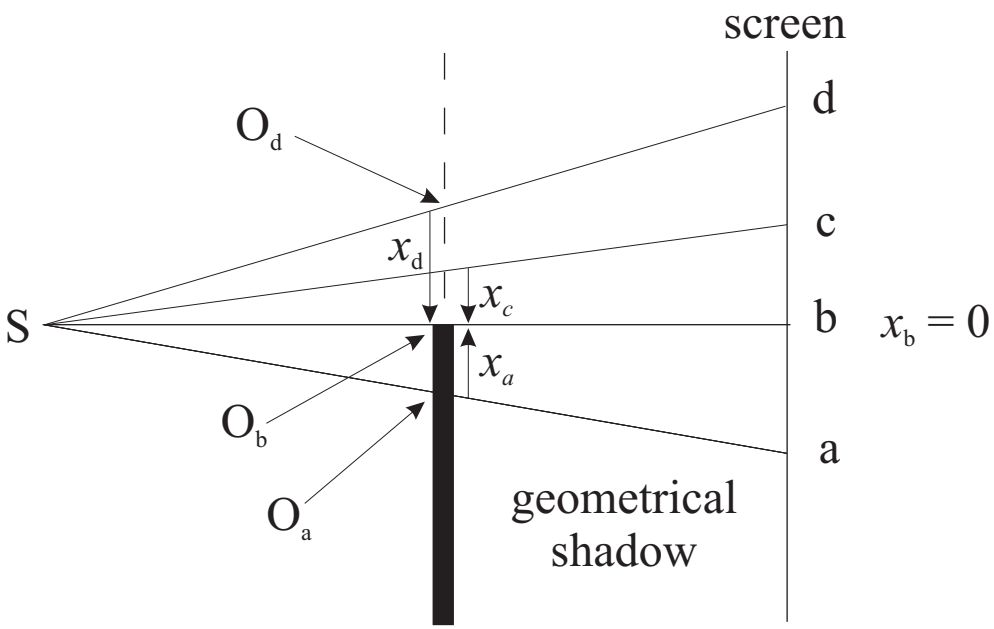


Figure 4.21: Wavefront obstructed by a straight edge.

To calculate the diffracted wave at point b, we define the origin O_b in the aperture plane to be at the edge of the obstruction. Fresnel conditions are satisfied since S, O_b and b are in a straight line. The diffracted wave is calculated from Eq. 4.61 by integrating over the aperture from $x = 0$ to $x = \infty$, i.e. $x_1 = 0$, $x_2 = \infty$. To calculate the diffracted wave at other points on the screen, we can use a trick: we simply move the origin so that it is still in between S and our observation point and Fresnel conditions are still satisfied. For observation point c, for example, the integral over the aperture is then from $x = x_c$ (which is negative) to $x = \infty$, i.e. $x_1 = -x_c$, $x_2 = \infty$. In summary:

a	$x_2 = \infty$	$x_1 = x_a > 0$	
b	$x_2 = \infty$	$x_1 = x_b = 0$	the geometrical edge
c	$x_2 = \infty$	$x_1 = x_c < 0$	
d	$x_2 = \infty$	$x_1 = x_d; w_d \sim -1.26$	maximum $\lvert\psi_P\rvert$

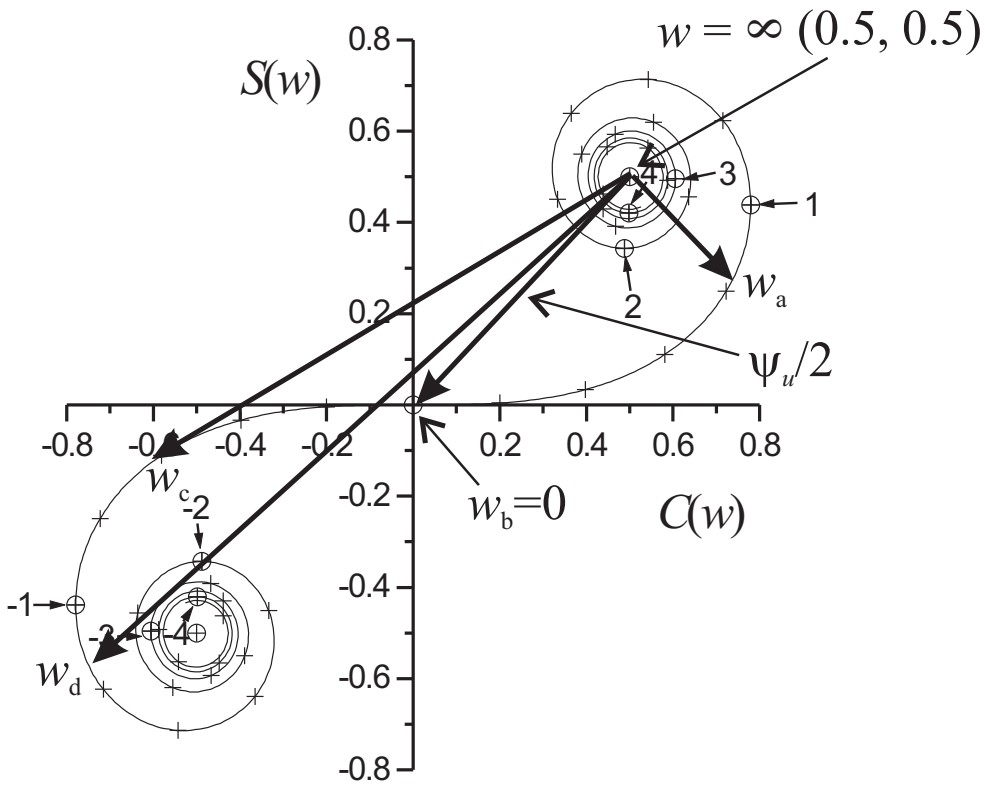


Figure 4.22: Diffraction amplitudes for diffraction by a straight edge.

The resulting diffracted waves are shown as vectors in the Cornu spiral in Figure 4.22, giving a diffraction pattern of the form shown in Figure 4.23.

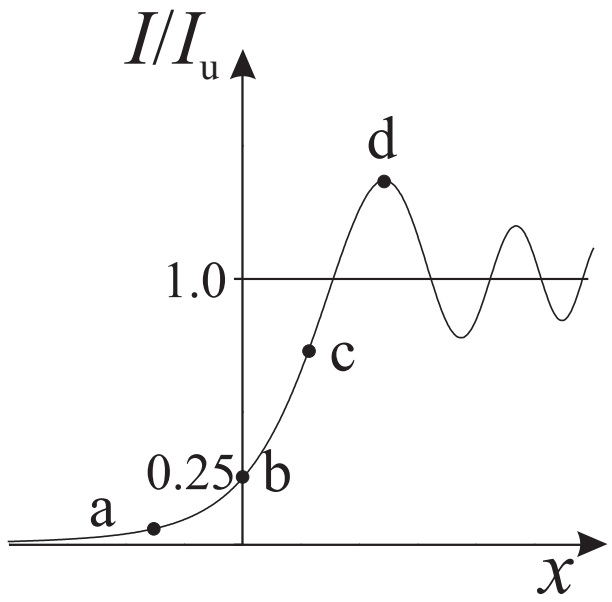


Figure 4.23: Fresnel diffraction pattern for a straight edge.

Note that

(i) well outside the shadow the intensity is I_u, the same as for an unobstructed wavefront;

(ii) the intensity at the geometric edge is 25% of I_u;

(iii) the intensity falls rapidly inside the geometric shadow;

(iv) there are fringes *outside* the geometric shadow, producing a maximum intensity 138% of I_u at d, just outside the edge.

4.5.3 A finite slit

Here we use a similar approach to the single edge, except that the integral does not extend to infinity.

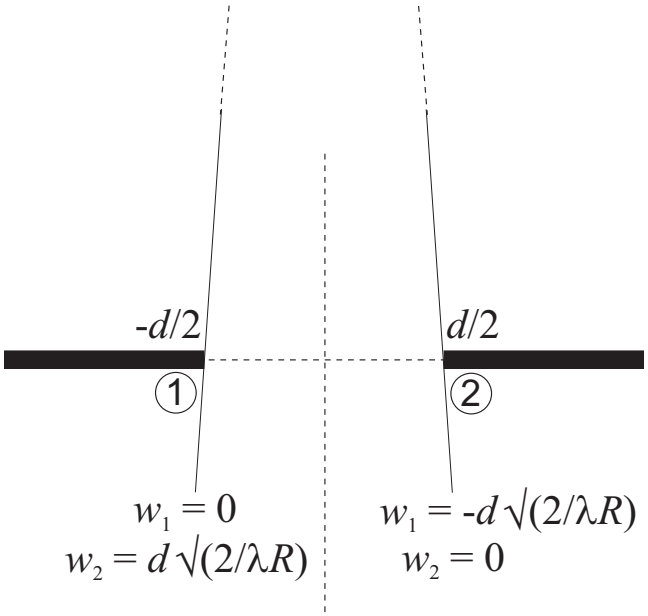


Figure 4.24: A slit as two straight edges.

For an observation point opposite one edge, $w_1 = 0$ and $w_2 = d\sqrt{(2/\lambda R)}$. For an observation point opposite the other edge, $w_1 = -d\sqrt{(2/\lambda R)}$. For an arbitrary observation point, we integrate from w_1 to w_2, where $w_1 - w_2 = \Delta w = d\sqrt{(2/\lambda R)}$. So the spanning vector on the Cornu spiral is between two points separated by a fixed length along the curve (Figure 4.25).

For small Δw, as shown in Figure 4.25, the spanning vector decreases monotonically in length as the observation point moves away from the centre of the slit, and the diffraction pattern has the form shown in Figure 4.26.

As $R \to 0$ or $d \to \infty$ (i.e. as the screen becomes "close" to the aperture), Δw becomes large, and we get two single edge patterns (Figure 4.27).

For intermediate situations, it may be that there is in fact a minimum on the axis of the system (see Figures 4.25 and 4.28) with fringing either side.

Do not forget that the diffraction pattern changes as we move away from the aperture (since R changes).

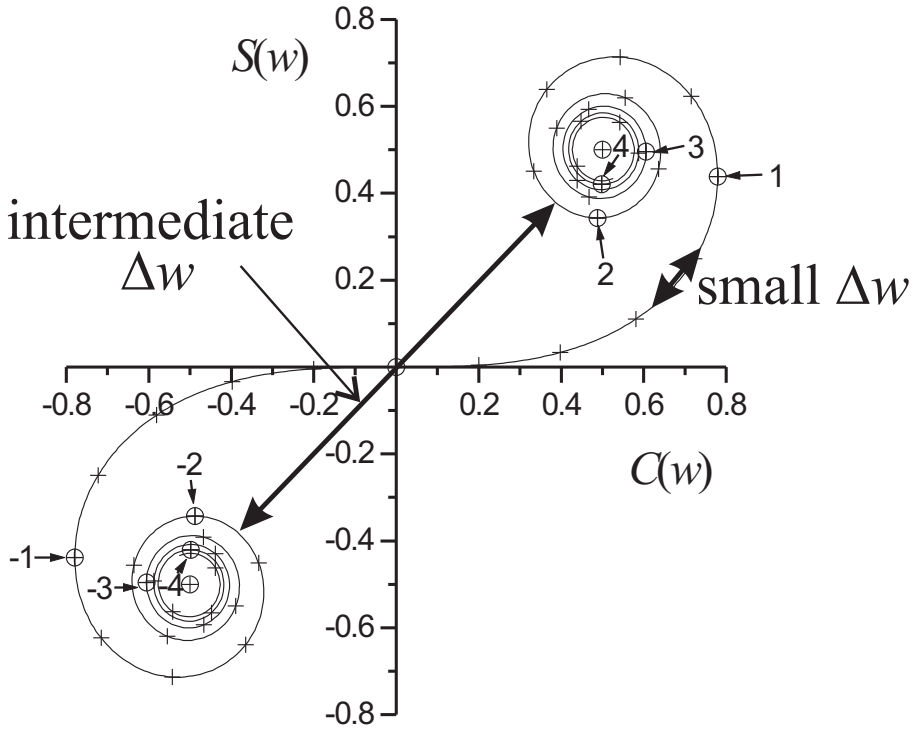


Figure 4.25: The Cornu spiral used to deduce the Fresnel diffraction pattern of a slit.

4.5.4 A circular aperture

Consider a circular aperture of radius a. Recall the general formula (Eq. 4.54) for the diffraction pattern

$$\psi_P \propto \iint_{\Sigma} \frac{h(x,y)K \exp\left(ik\frac{x^2+y^2}{2R}\right)}{r_1 r_2}\, dx\, dy. \tag{4.64}$$

This time we will retain the obliquity factor, and will retain the variation in r_1 and r_2 across the aperture. We will only consider observation points on the optical axis joining the source S with the centre of the aperture. Dividing the aperture into annular elements of radius ρ and thickness $d\rho$

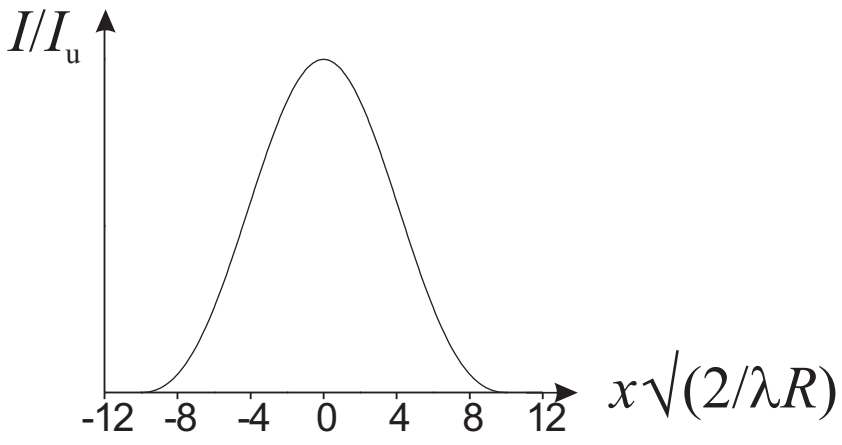


Figure 4.26: Fresnel diffraction for a narrow slit ($\Delta w = 0.2$).

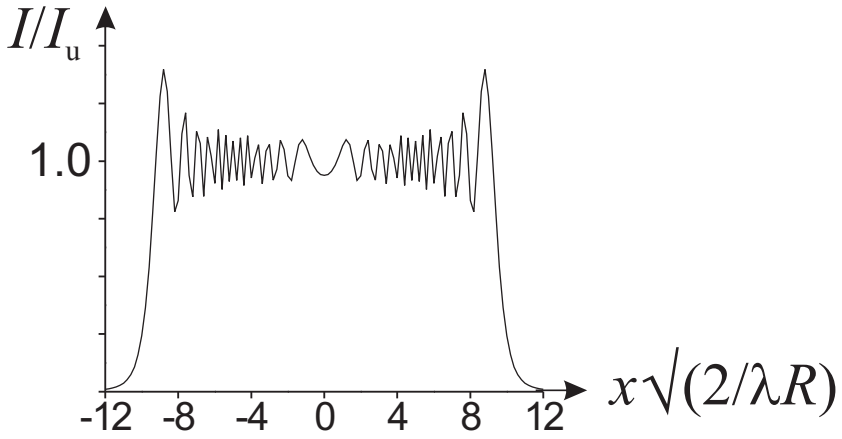


Figure 4.27: Fresnel diffraction for a wide slit ($\Delta w = 20$).

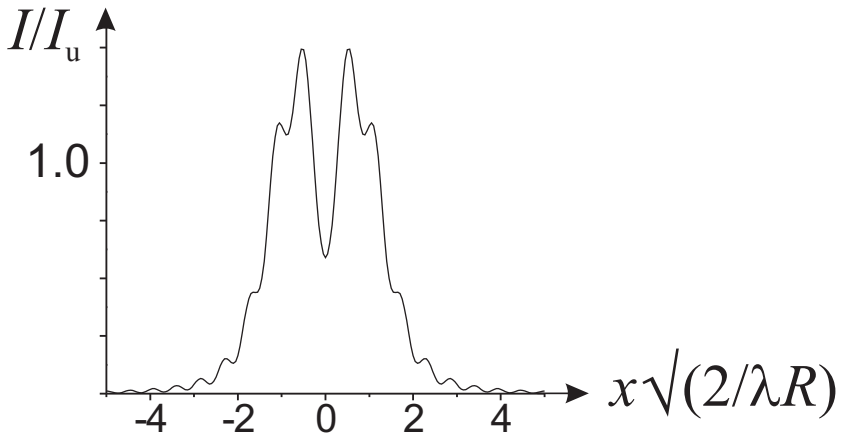


Figure 4.28: Fresnel diffraction for a narrow slit ($\Delta w = 3.8$).

we can replace $x^2 + y^2$ with ρ^2 and $dx\,dy$ with $2\pi\rho\,d\rho$ to give

$$\psi_P \propto \int_{\rho=0}^{\rho=a} \frac{K}{(a^2+\rho^2)^{1/2}(b^2+\rho^2)^{1/2}} \exp\left(\frac{ik\rho^2}{2R}\right) 2\pi\rho\,d\rho. \tag{4.65}$$

It is convenient to make a substitution:

$$\rho^2 = s\,; \;\; 2\rho\,d\rho = ds \tag{4.66}$$

and to replace k by $2\pi/\lambda$, to give

$$\psi_P \propto \int_{s=0}^{s=a^2} \frac{K}{(a^2+s)^{1/2}(b^2+s)^{1/2}} \exp\left(\frac{i\pi s}{\lambda R}\right) \pi\,ds. \tag{4.67}$$

This integral can be again be evaluated graphically using the phasor diagram shown in Figure 4.29. Note that:

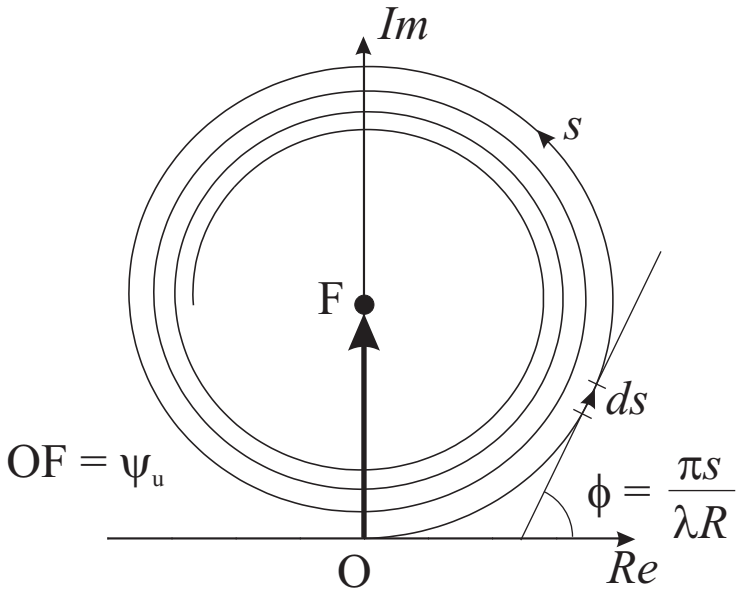


Figure 4.29: Phasor diagram for Fresnel diffraction and circular symmetry.

(a) the phase ($\phi = \pi s/(\lambda R)$) varies linearly with s, and the elemental contributions to the integral are approximately ds, so as s increases the phasor diagram is (approximately) a circle, just as for 1-d Fraunhofer diffraction.

(b) the term before the exponential is a slowly decreasing function of s (the denominator obviously increases with s, and $K = \frac{1}{2}(\cos\theta_S + \cos\theta_P)$ decreases with s). Thus the radius of the circle in the phasor diagram gradually decreases with s.

To calculate the diffracted amplitude we consider the spanning vector from point O to the point a distance s along the curve. The curve spirals in towards point F, so in the absence of an obstruction/aperture the integration range is from O to F, corresponding to $s = \infty$. OF is therefore the amplitude of the unobstructed wavefront.

For finite apertures/obstructions the diffraction integral (i.e. the spanning vector) varies considerably:

$$\begin{array}{lll} \phi = 2n\pi & \rho^2 = 2n\lambda r & \Rightarrow \ \psi \approx 0 \\ \phi = (2n+1)\pi & \rho^2 = (2n+1)\lambda r & \Rightarrow \ \psi \approx 2\psi_u \end{array}$$

4.5.5 Fresnel half-period zones

In the circular geometry under discussion, these are concentric circular zones in the aperture plane, over which the phase at the observation point changes by π.

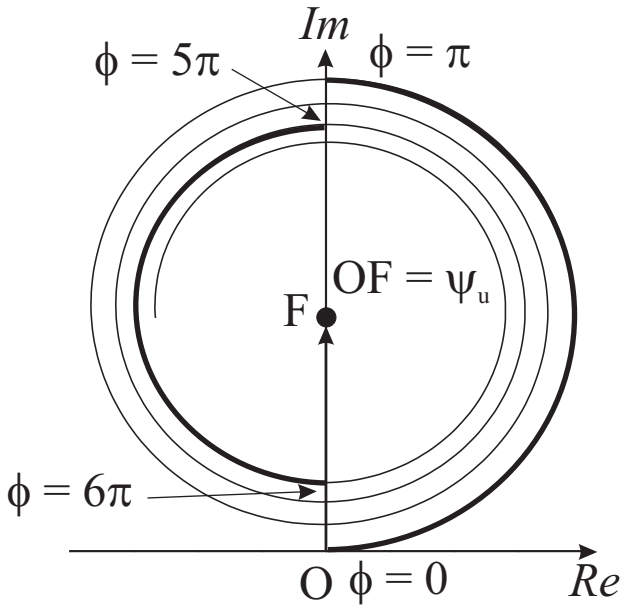


Figure 4.30: Fresnel half-period zones.

We define the 1st zone as the circular region in the aperture plane which satisfies

$$0 \leq \phi(\rho) \leq \pi \tag{4.68}$$

This corresponds to

$$\rho^2 \leq \lambda R. \tag{4.69}$$

The nth zone corresponds to

$$(n-1)\pi \leq \phi(\rho) \leq n\pi \tag{4.70}$$

$$\sqrt{(n-1)\lambda R} \leq \rho \leq \sqrt{n\lambda R}. \tag{4.71}$$

Note that the *area of each zone is the same*

$$\pi\left(\rho_n^2 - \rho_{n-1}^2\right) = \pi\lambda R. \tag{4.72}$$

If we neglected the obliquity factor and the variation of r_a and r_b, each zone would contribute equally to the amplitude at P, and the phasor diagram would be exactly circular.

Odd numbered zones *add* to, and *even* numbered zones *subtract* from the overall amplitude at P.

So, for an observation point P on the optic axis of a circular aperture of radius a, the aperture includes N zones, given by $a^2 = n\lambda R$.

N ODD: bright spot at P; $\psi \sim 2\psi_u$; $I \sim 4I_u$

N EVEN: dark spot at P: $\psi \sim 0$; $I \sim 0$

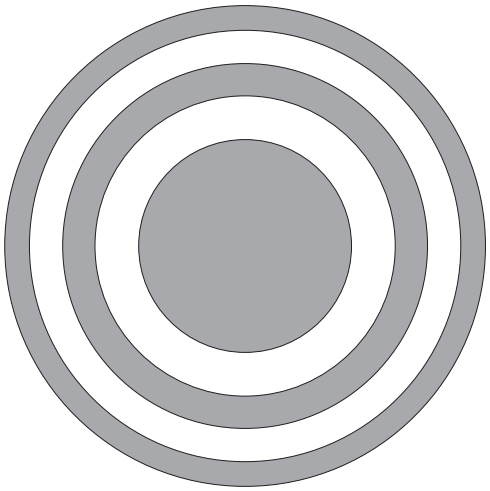


Figure 4.31: Fresnel half-period zones across the wavefront; circular geometry.

For large apertures, the small angle approximations begin to break down. The "spiralling in" in the phasor diagram due to the $1/r$ terms in the diffraction integral becomes noticeable, and this effect is further enhanced since $K < 1$. The width of the outer zones also increases, as the higher-order terms in Eq. 4.52 become important.

4.5.6 Circular obstruction: Poisson's Spot

Consider a circular obstruction on the axis. For an obstacle of radius a, the inner zones up to $\rho = a$, i.e.

$$\phi_a = \frac{\pi a^2}{\lambda R} \tag{4.73}$$

are obscured, but all the outer zones are clear. So, the limits of integration for the diffraction integral are $\rho = a$ to $\rho = \infty$, and the relevant spanning vector in the phasor diagram is from F to A as shown in Figure 4.32.

Provided ϕ_a is not too large (i.e. provided a is not too large) the intensity is therefore close to that expected in the absence of any obstruction. This leads to a bright spot on the axis, known as *Poisson's spot*.

Close to the obstruction, the diffraction angles also become large so that K falls, and no spot is observable.

4.5.7 Off-axis intensity for a circular aperture/obstacle

The Fresnel zones are centred on the source-observation point straight line SOP (Figure 4.33), so as P shifts off the axis of the aperture, the zone structure moves with it across the aperture. The situation is modelled *approximately* (but quite well) by assuming that S and P are fixed, and the aperture shifts sideways across the zone structure, as shown in Figures 4.33 and 4.34.

Take as an example an aperture which covers the first two zones when on axis (Figure 4.34(1)). This produces (in principle), zero intensity on the axis itself.

In (3), the 1st zone gives a full contribution, but the 2nd and 3rd are each partially obstructed, and tend to cancel each other. So there is a maximum in the intensity.

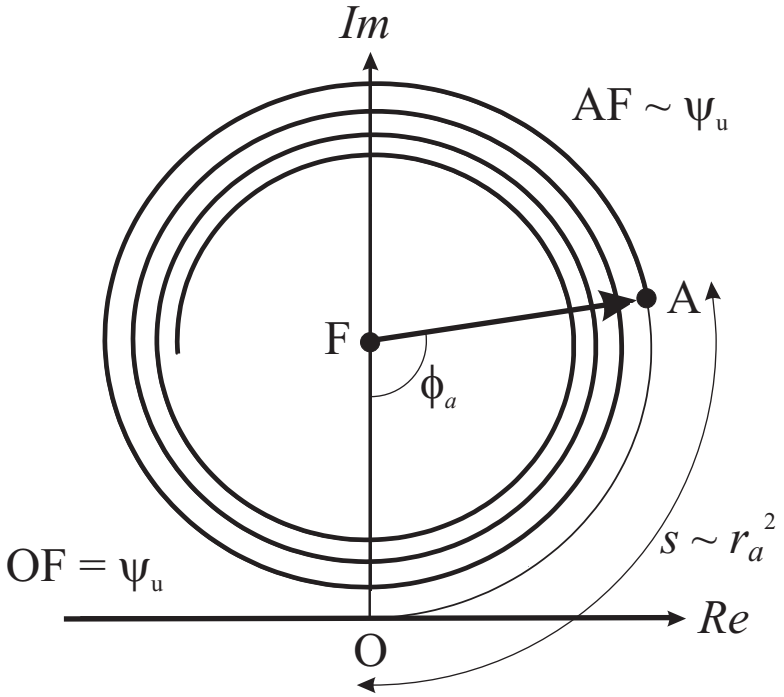


Figure 4.32: Phasor diagram for a circular obstacle.

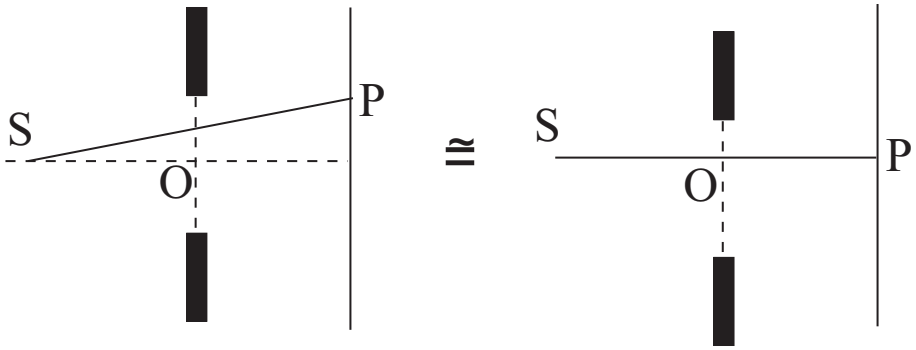


Figure 4.33: Fresnel diffraction for points of the axis.

Hence, there is an oscillation in intensity as P moves off-axis. Circular symmetry means this must correspond to *circular fringes*. The periodicity of the fringes is approximately equal to the zone width at the edge of the aperture.

For larger apertures, the zone widths are smaller at the edge, so the fringe spacing is smaller.

A long way off-axis, the aperture is clear for sectors of a large number of very narrow zones. The aperture covers most of the area of each of these zones, so they contribute rather little to ψ, decreasing as P goes further off-axis. So the intensity at P falls rapidly far from the optic axis of the aperture.

4.5.8 Lenses used to produce Fresnel conditions

Earlier we established that Fraunhofer conditions (ϕ linear in x) can be established by putting the source and screen in *conjugate planes* of an optical system, with the aperture anywhere (in principle) in between.

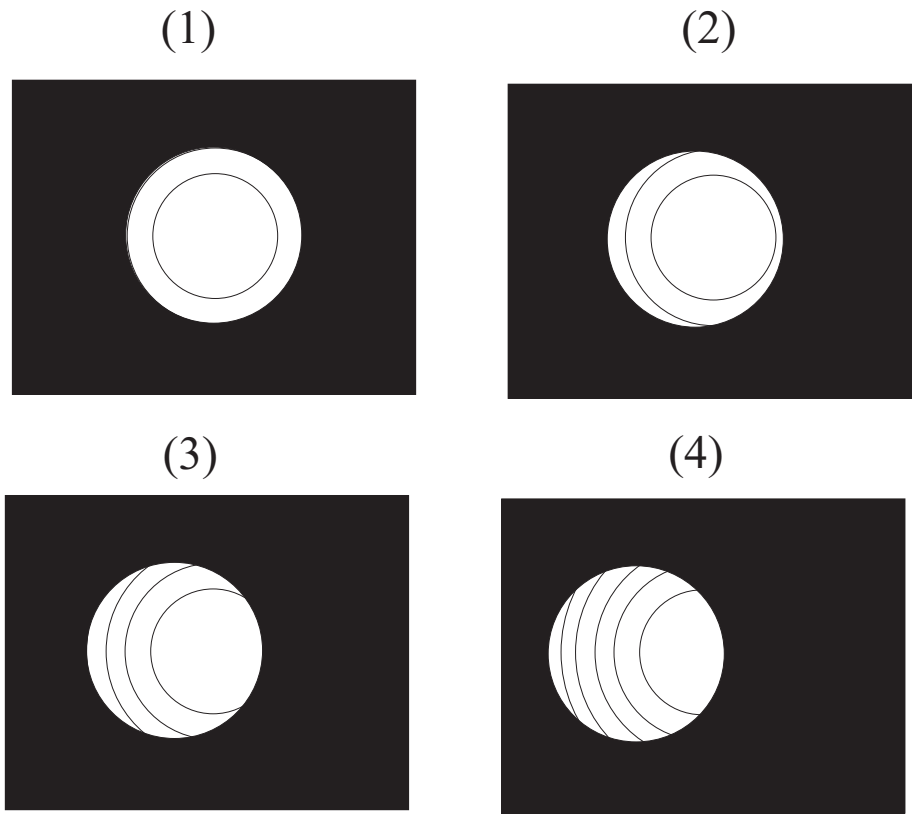


Figure 4.34: Off-axis diffraction determined by zone obstruction by aperture.

If the source/screen/lens is moved away from these positions, this condition of course breaks down, and higher order terms in x have to be taken into account.

So, Fresnel conditions (ϕ quadratic in x) can be achieved by *defocussing* such an arrangement. This will be illustrated in the Practical Class.

4.5.9 The Fresnel zone plate

A zone plate is an aperture which blocks out alternate Fresnel half-period zones.

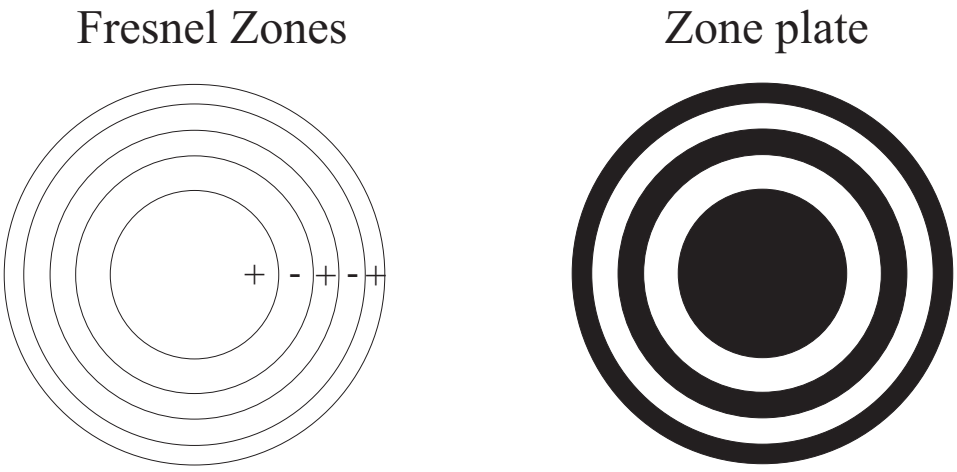


Figure 4.35: Fresnel zone plate.

The aperture on the right will, for the observation point P, block out the 1st, 3rd, 5th ... zones shown on the left if designed as follows in Figure 4.36 with

$$\rho_1 = \sqrt{\lambda R},\ \rho_2 = \sqrt{2\lambda R},\ \rho_3 = \sqrt{3\lambda R},\ \text{ etc.} \tag{4.74}$$

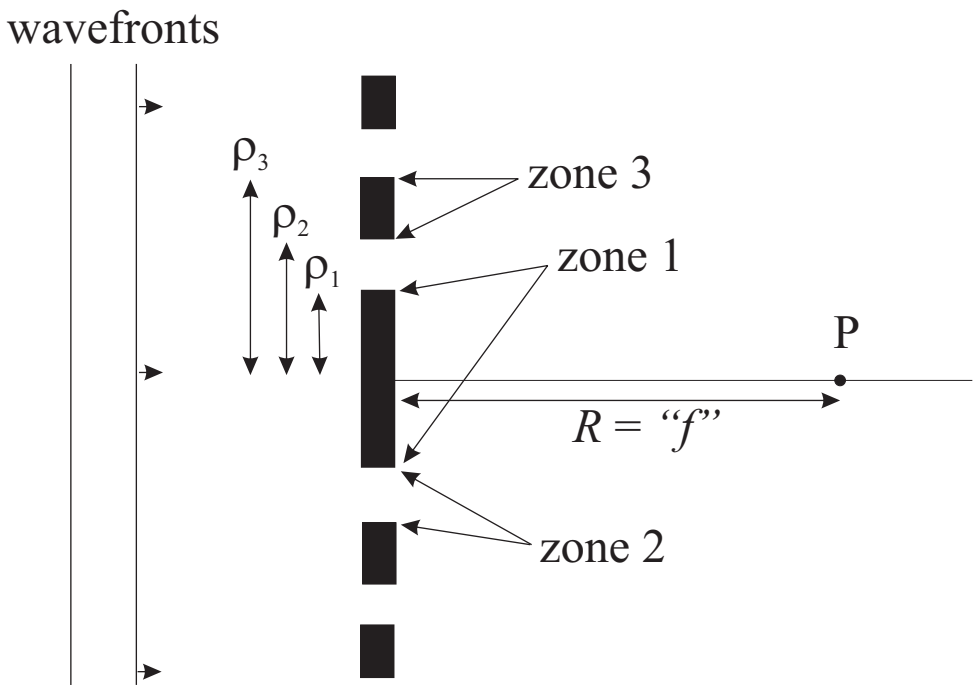


Figure 4.36: Fresnel zone plate; calculation of dimensions.

The effect on the phasor diagram is shown in Figure 4.37.

So, the net amplitude at P is

$$\begin{aligned} \psi_P &= \psi_2 + \psi_4 + \psi_6 + \dots \\ &\approx 2N\psi_u \end{aligned} \tag{4.75}$$

where N is the number of open zones in the plate. Thus the intensity is

$$I_P \approx 4N^2 I_u. \tag{4.76}$$

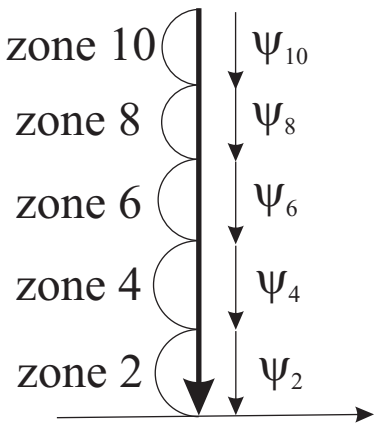


Figure 4.37: Fresnel zone plate; phasor diagram.

So, an incident plane wave is "brought to a focus" at P, and the zone plate acts a *lens* with focal length

$$f = \frac{\rho_1^2}{\lambda} = \frac{\rho_n^2}{n\lambda} \tag{4.77}$$

Clearly, because $f \propto 1/\lambda$, this is a highly chromatic lens, and suitable only for its design wavelength.

Suppose now we move the observation point P along the axis towards the zone plate, with the zone plate dimensions fixed as before. The Fresnel zones for the new position R′ of P, however, become smaller in size, with

$$\rho_1' = \sqrt{\lambda R'}, \rho_2' = \sqrt{2\lambda R'}, \rho_3' = \sqrt{3\lambda R'}, \text{ etc.} \tag{4.78}$$

When $R' = f/2$,

$$\begin{aligned} \rho_1' &= \sqrt{\lambda f/2} \\ \rho_2' &= \sqrt{\lambda f} = \rho_1 \\ \rho_3' &= \sqrt{3\lambda f/2} \\ \rho_4' &= \sqrt{2\lambda f} = \rho_2. \end{aligned}$$

So the open area of the plate

$$\rho_1 \leq \rho \leq \rho_2 \tag{4.79}$$

now allows through *two* zones, 3 and 4, for the new position of P. Hence $\psi_P \sim 0$.

More generally (for m an integer):

(i) when $R' = f/2m$, each open area of the zone plate admits an even number of Fresnel zones of opposite phase, so $\psi_P \to 0$. So there are points of zero intensity on the axis at $R' = f/2m$.

(ii) when $R' = f/(2m+1)$, each open area of the zone plate admits an odd number of Fresnel zones, with a net contribution of one zone for each of the N open areas. So $\psi_P \to 2N\psi_u$ in principle, but this is reduced by the decreasing K for the outer zones. However, there are clearly maxima at $R = f/(2m+1)$.

So, there are *subsidiary foci*, and, schematically, the intensity varies along the optic axis as shown in Figure 4.38.

Figure 4.38 neglects the obliquity factor, which would reduce, eventually to zero, the intensity of the maxima at $f/(2m+1)$ as $m \to \infty$.

Hence the lens is not a very good one, even at its design wavelength. However, at some wavelengths (e.g. X-rays, for which the refractive indices $n \sim 1 + \delta$) it may be the only option.

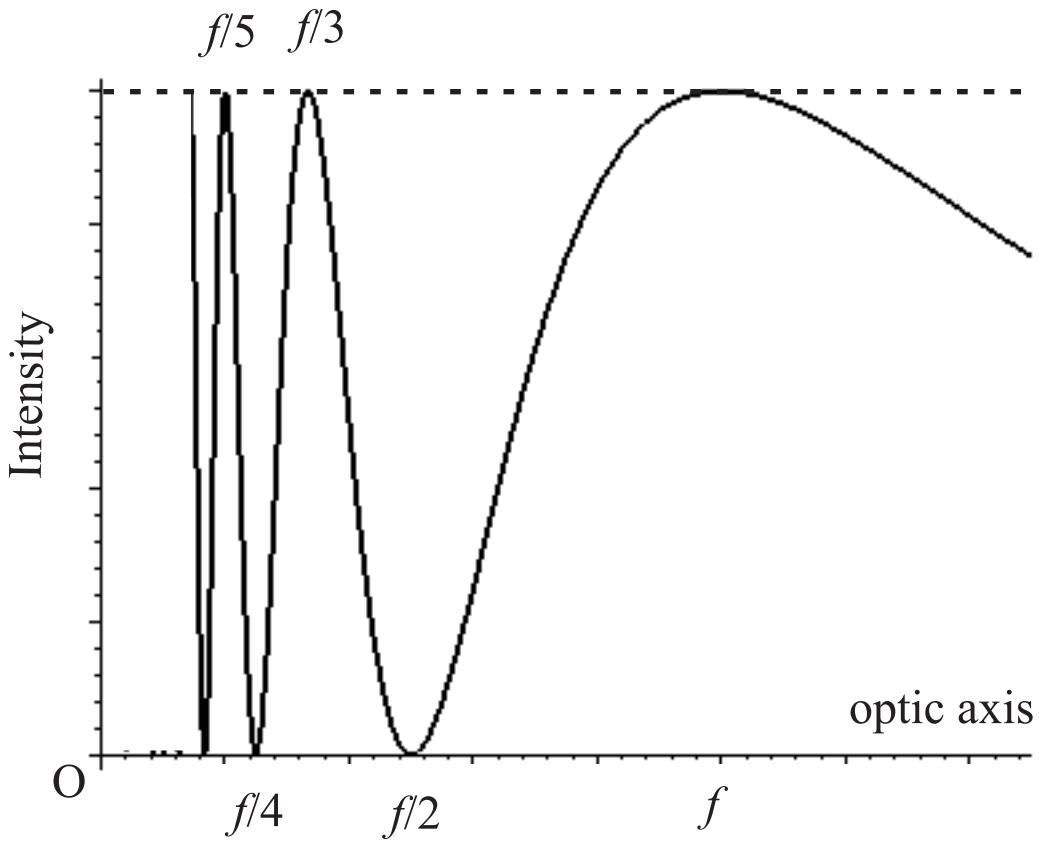


Figure 4.38: Subsidiary foci for a Fresnel zone plate.

5 Interference

5.1 Conditions for interference

Consider a wave Ψ which is the superposition of two monochromatic waves at angular frequencies ω_1 and ω_2

$$\Psi = \Re\left[\psi_1 e^{-i\omega_1 t} + \psi_2 e^{-i\omega_2 t}\right].$$

Most optical detectors respond to the intensity I of the wave, which is proportional to the square of Ψ. Expanding the real part of a complex quantity A as $\Re(A) = \frac{1}{2}(A + A^*)$ we get, after a little working

$$\begin{aligned} I \propto \Psi^2 =& \tfrac{1}{2}|\psi_1|^2 + \tfrac{1}{2}|\psi_2|^2 + \Re\left[\psi_1\psi_2^* e^{i[\omega_2-\omega_1]t}\right] \\ &+ \tfrac{1}{2}\Re\left[\psi_1^2 e^{-2i\omega_1 t}\right] + \tfrac{1}{2}\Re\left[\psi_2^2 e^{-2i\omega_2 t}\right] + \Re\left[\psi_1\psi_2 e^{-i[\omega_1+\omega_2]t}\right] \end{aligned}$$

The last three terms in this expression are varying on a timescale which is more rapid than the response time of most detectors, and so these terms average to zero during the response time. We can expand the coefficients ψ_1 and ψ_2 into their amplitudes and phases in the form $\psi = ae^{i\phi}$, giving a mean intensity

$$\langle I\rangle \propto \tfrac{1}{2}\langle a_1^2\rangle + \tfrac{1}{2}\langle a_2^2\rangle + \langle a_1 a_2 \Re\left[e^{i[\phi_1-\phi_2-(\omega_1-\omega_2)t]}\right]\rangle \tag{5.1}$$

where the angle brackets denote averaging over the integration time of the detector. The observable condition which tells us that we are seeing interference phenomena is that we can see a *reduction* in intensity when we add two beams together. This is clearly only possible if the third term in Eq. 5.1 is non-zero; in other words this term is the interference term and the other two terms are present whether or not interference is occurring.

If the detector is averaging the intensity over some finite period τ, we will not see interference if $(\omega_1 - \omega_2)\tau \gg 1$, in other words to see interference we need to have $\omega_1 = \omega_2$ to a good degree of approximation. For example, optical detectors typically have a response time $\tau > 10^{-9}$s whereas in the optical regime $\omega \sim 10^{15}$ rad/s, so the frequencies need to be equal to within a part per million or better.

In practice at optical wavelengths, the phases ϕ_1 and ϕ_2 of independent sources (usually atoms) vary randomly and rapidly. Even phase of the light emitted from a laser, which is typically the most stable light source available, can change every few nanoseconds, so *interference is typically only seen when light from a single source (e.g. atom) is split and then recombined*: only in this case can $\phi_1 - \phi_2$ be stable for long enough that interference is seen.

There are two generic ways to do this splitting and recombining of the wavefront from a light source. In the previous chapter, we saw many examples of interference in the context of diffraction. We considered a wavefront incident on an aperture, producing secondary waves which overlap and interfere. The incident wavefront is thus modulated *spatially* by the aperture, corresponding to *wavefront division*. Here, we will consider cases of interference where the incident wavefronts are divided in *amplitude*, for example by reflection and transmission at an interface. The waves produced then travel through different optical paths before they subsequently overlap, producing an interference pattern. This type of interference is known as *amplitude division*.

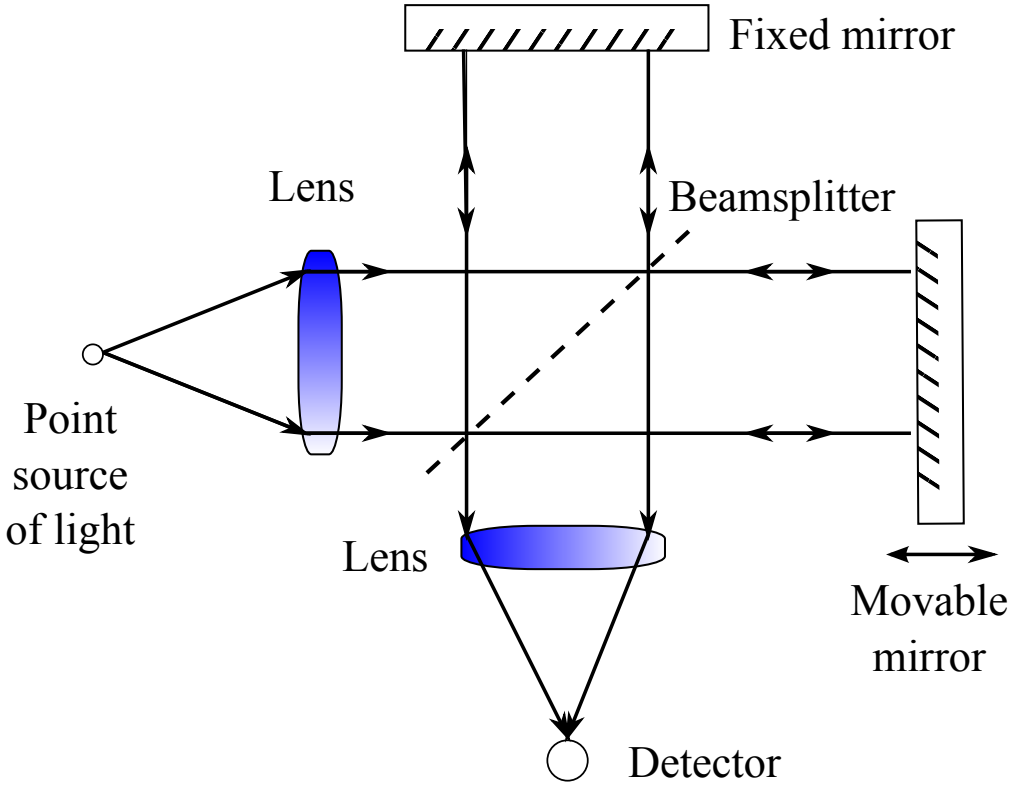


Figure 5.1: The Michelson interferometer

5.2 The Michelson Interferometer

The Michelson interferometer is perhaps the simplest amplitude-splitting interferometer, but it has many important applications, including high- precision metrology, Fourier Transform spectroscopy and the Michelson- Morley experiment.

The interferometer uses a beam-splitter to split the incident beam into two components which travel along different paths, and then recombine the beams on the same beamsplitter, as shown in Figure 5.1. After recombination, we get interference between two copies of the beam which have travelled different distances. Constructive or destructive interference will occur depending on the difference between the two paths. Normally one path is kept fixed (the so-called reference path) and the other path is varied by moving a mirror on a precision stage. The interferometer must be made very rigid, because vibrations of order a wavelength will "smear" the fringes out.

The interferometer is usually operated with an extended light source for visual work. Light from different parts of the source travels at different angles through the interferometer to the eye and the differential delays will in general be a function of this angle. In addition, the mirrors may be tilted to introduce different delays for beams hitting the mirrors at different points. The combination of these effects means that a spatially-varying intensity (i.e. fringes) can be seen when the mirrors are both held fixed.

The arrangement shown in Figure 5.1 is commonly used if an electronic detector such as a photodiode is being used to detect the fringes. In this case the system is illuminated with the collimated beam from a point source and the mirrors are aligned so that the path difference is constant across the beam. As a result, the interference pattern is either all light or all dark depending on the path difference between the two arms. The fringe pattern i.e. the sinusoidal variation in intensity, is seen by varying the displacement of one of the mirrors as a function of time and recording the intensity as a function of the mirror position. This point-source temporally-varying geometry is easier to analyse than the extended-source spatially-varying geometry and so we concentrate on the former

here.

5.2.1 Monochromatic fringes

For light with wavenumber $k = 2\pi/\lambda$ the measured intensity can be derived from Eq. 5.1 with $\omega_1 = \omega_2$

$$\langle I \rangle \propto \tfrac{1}{2}\langle a_1^2 \rangle + \tfrac{1}{2}\langle a_2^2 \rangle + \langle a_1 a_2 \Re\left[e^{i\delta}\right] \rangle, \tag{5.2}$$

where $\delta = \phi_1 - \phi_2$ is the phase difference between the interfering beams. This difference is given by $\delta = kx$ where x is the difference in the paths travelled by the two beams. It is important to note that x is *twice* the difference in the distances of the two mirrors from the beamsplitter as the beams traverse these distances twice. We assume here that the beamsplitter is symmetric and therefore introduces no additional relative phase shift between the beams. If we assume that each of the interfering beams has a constant intensity $I_0/2$ then this expression becomes

$$I(x) = I_0 \left(1 + \Re\left[e^{ikx}\right]\right), \tag{5.3}$$

where we have dropped the angle brackets: averaging over the response time of the detector is implicit from now on.

Thus if we vary x linearly as a function of time, then the intensity seen at the output of the interferometer will vary sinusoidally, i.e. we will get a fringe pattern in time.

5.2.2 Interference with broadband light

If the beam being observed is not truly monochromatic, then light at each wavelength will form its own set of fringes — there will be no fringes corresponding to interference between different wavelengths, because the condition $\omega_1 - \omega_2 \ll 1/\tau$ will not in general be satisfied. The total intensity pattern observed will therefore be the sum of the intensities of all of the fringe patterns at the different wavelengths. These patterns will be sine waves of different frequencies and so adding them will tend to result in a blurred pattern. If the intensity of light in the wavenumber range $k \to k + dk$ is given by $2S(k)dk$ (the reason for the factor of 2 will become apparent shortly) then the total intensity observed will be the sum of all the sine waves weighted by the intensity at the relevant wavelength:

$$I(x) = 2\int_0^\infty S(k)\left(1 + \Re\left[e^{ikx}\right]\right) dk \tag{5.4}$$

For mathematical convenience, we will also define $S(k)$ for negative k, such that $S(-k) = S^*(k)$. The fringe pattern as a function of the path difference x can then be written

$$\begin{aligned} I(x) &= \int_{-\infty}^\infty S(k)\left(1 + e^{ikx}\right) dk \\ &= I_1 + \int_{-\infty}^\infty S(k) e^{ikx}\, dk \end{aligned} \tag{5.5}$$

where I_1 is a constant equal to the total intensity of the light

$$I_1 = \int_{-\infty}^\infty S(k)\, dk. \tag{5.6}$$

5.2.3 Fourier transform spectroscopy

We recognise the second term in Eq. 5.5 as being the inverse Fourier transform of $S(k)$. To find S, we can therefore measure the intensity as a function of the pathlength difference $I(x)$, subtract off the constant term and take the Fourier transform

$$S(k) \propto \int_{-\infty}^{\infty} (I(x) - I_1) e^{-ikx}\, dx. \tag{5.7}$$

This method is used in modern infrared spectrometers: the *Fourier transform infrared spectrometer* (FTIR) is a standard tool for characterising organic molecules by their vibrational frequencies, and uses a Michelson interferometer with automatic scanning of one mirror, accompanied by software to perform the Fourier transform.

Fourier Transform Spectroscopy (FTS) is used for getting very-high- precision measurements of narrow spectral lines, because it is capable of higher spectral resolution than a spectrometer based on a prism or a diffraction grating. It is also easier to calibrate to high precision, because all the wavelengths of light pass through the same parts of the apparatus. The disadvantage is that the FTS relies on making many sequential measurements of $I(x)$, and this may take longer than taking a single image from a grating spectrograph.

Example: A Fourier Transform Spectrograph has a maximum mirror displacement of 50cm. What is its spectral resolution at a wavelength of $\lambda_0 = 1\,\mu$m?

Solution: With a maximum mirror displacement of 50cm, the maximum change in pathlength difference will be 1 m. We assume that this is disposed symmetrically about the position of zero pathlength difference, i.e. x has a range of ± 50 cm. Knowing $I(x)$ over this range is equivalent to knowing $I(x)$ over the infinite range and then deliberately truncating it with a tophat function of width $W = 1$m.

The reconstructed spectrum will therefore be convolved with the Fourier Transform of the tophat function (a sinc function). Any spectral information at finer resolutions will be blurred by this convolving function. The width of the blurring sinc function in spatial frequency space will be $\Delta k = 2\pi/W$. We can express this as a wavelength resolution using the fact that $k = 2\pi/\lambda$ so that $dk/k = -d\lambda/\lambda$

$$|\Delta\lambda| = \lambda \frac{\Delta k}{k} = \frac{\lambda^2 \Delta k}{2\pi} = \frac{\lambda^2}{W} \tag{5.8}$$

This result means that we can in principle resolve (i.e. distinguish) two spectral lines which are only 1 picometre apart in wavelength when the central wavelength is 1 micrometre.

5.2.4 Fringe visibility

If we have a light source with two closely-spaced wavelength components at $k_0 \pm \Delta k$, each monochromatic component will produce a fringe pattern with a slightly different fringe spacing. The observed interference pattern will be the sum of the two individual patterns. Around $x = 0$, the two fringe patterns approximately coincide, so the overall pattern will show well-defined maxima and minima. However, as x increases, a phase shift develops between the two patterns, and eventually (if the two components have the same intensity) the fringes become invisible since the maxima in one set of fringes overlap with the minima in the other set. The larger the spacing between the wavelength components, the more rapidly fringes disappear as x is increased.

In our double-sided convention for a spectrum, $S(k)$ can be represented as pairs of closely-spaced

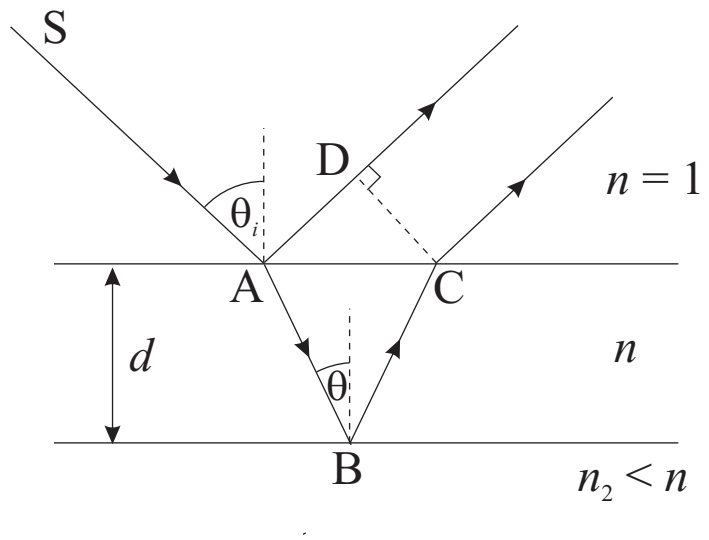


Figure 5.2: Interference in a thin film.

delta functions spaced about $\pm k_0$. This can be written as a convolution

$$S(k) = [\delta(k - k_0) + \delta(k + k_0)] * [\delta(k - \Delta k) + \delta(k + \Delta k)] \tag{5.9}$$

Thus the intensity pattern will contain a product of two cosine functions

$$I(x) = I_1[1 + \cos(k_0 x)\cos(\Delta k x)] \tag{5.10}$$

The high frequency cosine $\cos(k_0 x)$ represents the fringes while the low frequency envelope $\cos(\Delta k x)$ modulates the amplitude of these fringes. We can define the *fringe contrast* or *fringe visibility* as the ratio of the local high-frequency fringe amplitude to the mean intensity

$$\text{Visibility} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}. \tag{5.11}$$

At the position of the first zero in the low-frequency envelope, the fringe contrast will be zero and this can be used to find Δk.

Measuring the fringe visibility allows the fine structure of atomic lines to be investigated. For example in the 1890s Michelson was able to confirm that the sodium D line at 589.3 nm was in fact a doublet with $\Delta\lambda = 0.6$ nm, whereas the cadmium line at 634.8 nm was highly monochromatic, showing no minimum in fringe visibility for values of x up to 0.4 m.

5.3 Thin Film Interference

Amplitude division interference occurs when light waves are incident on a uniform thin film of material. Here, interference occurs between the wave reflected at the top surface of the film and the wave which is initially transmitted, then reflected at the bottom surface and finally transmitted at the top surface, as shown in Figure 5.2.

To calculate the intensity produced, we must consider the phase difference between paths AD and ABC. Taking into account the different refractive indices we have an optical path difference of

$$\begin{aligned} n(\text{AB} + \text{BC}) - \text{AD} &= 2n\,\text{AB} - \text{AD} \\ &= \frac{2nd}{\cos\theta} - 2d\tan\theta\sin\theta_i \end{aligned}$$

Using Snell's law, $\sin\theta_i = n\sin\theta$ we obtain

$$\begin{aligned} n(\mathrm{AB}+\mathrm{BC})-\mathrm{AD} &= \frac{2nd}{\cos\theta}\left(1-\sin^2\theta\right) \\ &= 2nd\cos\theta \end{aligned} \tag{5.12}$$

This path difference introduces a phase difference

$$\delta = 2n\,d\,k\cos\theta = \frac{4\pi nd}{\lambda}\cos\theta. \tag{5.13}$$

The reflection at the upper surface occurs at a high-impedance to low-impedance boundary, while that at the lower surface occurs at a low-impedance to high-impedance boundary. This introduces an extra phase difference of π between the two beams. For simplicity, let us make the (unjustified) assumption that both beams have the same intensity $I_0/2$. In this case Eq. 5.2 gives the intensity

$$I(\delta) = I_0\left(1-\Re\left[e^{i\delta}\right]\right). \tag{5.14}$$

These characteristic fringes have maxima and minima as a function of angle θ as the two waves interfere constructively and destructively. Where the two interfering beams differ in strength, the fringes are still present with the same period, but are less distinct since they sit on top of a constant background. These fringes are known as *fringes of equal inclination*, and may be observed by using a lens to focus the pattern. A particularly easy situation to visualise is the case where the film is illuminated by an extended source at close to normal incidence (Figure 5.3). Here, circular fringes (*Haidinger fringes*) are produced, with each ring corresponding to a constant value of θ.

At a given value of θ, maximum transmission occurs at the wavelength corresponding to $\delta = (2m+1)\pi$, where m is an integer. This corresponds to

$$n\,d\cos\theta = \frac{(2m+1)}{4}\lambda. \tag{5.15}$$

Note that the resonance wavelength (the wavelength for constructive interference) decreases as θ increases. This is somewhat counterintuitive, since the path length inside the film increases. However, this is more than compensated for by the additional path length along AD.

A more commonly observed type of interference fringes is seen for films of non-uniform thickness. These are known as *fringes of equal thickness*. For near-normal incidence, bright fringes are seen in the regions of the film where the thickness satisfies the condition

$$2nd = (m+\tfrac{1}{2})\lambda. \tag{5.16}$$

This type of fringe pattern is often seen in soap films or in thin films of oil on water. When illumination is with white light, each wavelength component produces its own set of fringes with a different period. The superposition of these fringes leads to a complex pattern of colours (interference colours) across the film.

A particular example of fringes of equal thickness occurs when an air gap is formed between two spherical surfaces (or a spherical surface and a plane surface), as shown in Figure 5.4. This produces circular fringes known as *Newton's rings*.

5.4 The Fabry-Pérot etalon

In the discussion of thin-film interference above, we have neglected the effect of reflections at the film-air interface. For a simple film this is normally a good approximation, however we can

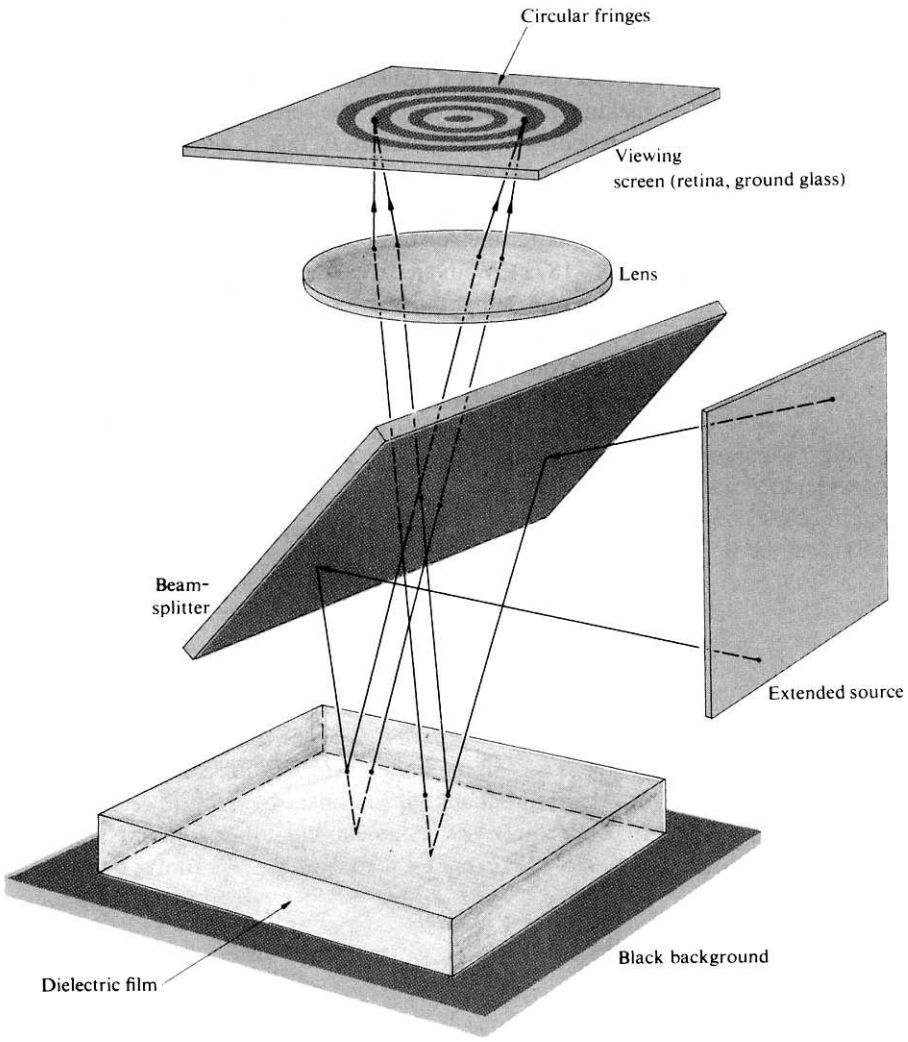


Figure 5.3: Haidinger fringes (from Hecht, *Optics*).

design a structure with much larger reflection coefficients at either side of the film, where we must consider interference between multiple beams. An example of this type of structure is a film of air sandwiched between two mirrors, known as a *Fabry-Pérot etalon*. This etalon produces a fringe pattern with much sharper features than a simple thin film, and is useful in high-resolution optical spectroscopy.

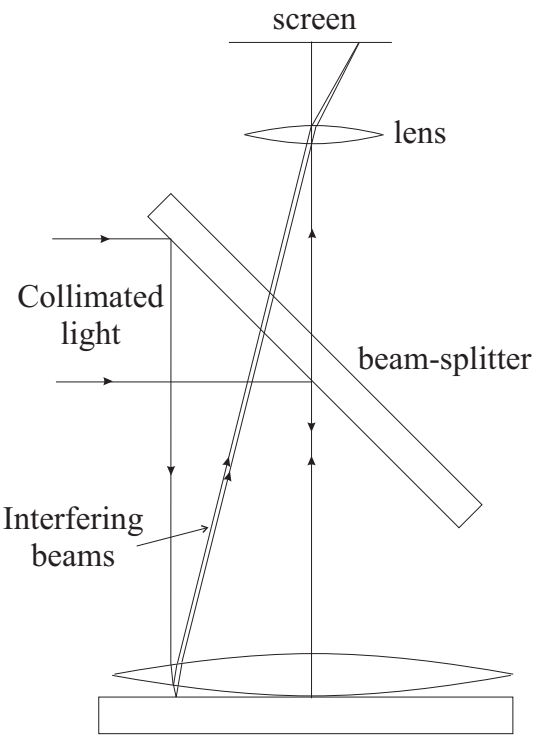
Consider a thin film of air of thickness d surrounded by mirrors. For light incident at angle θ the mirrors have an amplitude reflection coefficient r (which we assume is real). The amplitudes of the waves generated after successive reflections inside the cavity are shown in Figure 5.5. Each successive beam emerging from the etalon acquires a factor of $r^2 = R$ in amplitude and a phase shift δ, where

$$\delta = \frac{4\pi d}{\lambda}\cos\theta. \tag{5.17}$$

(from Eq. 5.13 with $n = 1$). Note that we can neglect the effect of the glass supporting the mirrors, since this introduces a constant phase shift into all the beams.

Taking account of the phase shift, the total amplitude emerging from the etalon is given by

$$A = T(1 + Re^{i\delta} + R^2e^{2i\delta} + R^3e^{3i\delta}\ldots). \tag{5.18}$$



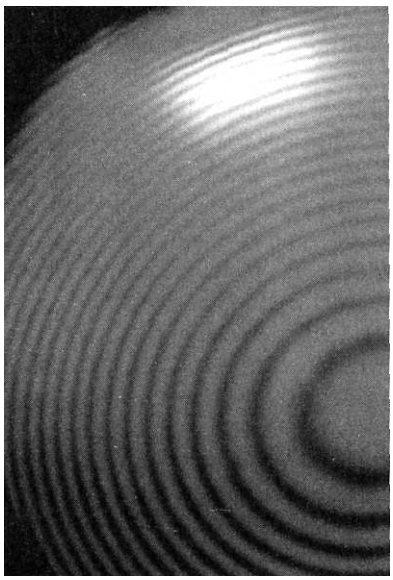


Figure 5.4: Experimental arrangement for observing Newton's rings formed between a spherical surface and a flat surface (left) and the rings seen (right, from Hecht, *Optics*).

This is a geometric series with a total

$$A = \frac{T}{1 - Re^{i\delta}}. \tag{5.19}$$

The total intensity transmission of the etalon, $I_{\text{out}}/I_{\text{in}}$ is thus

$$\begin{aligned} |A|^2 &= \frac{T^2}{(1 - R\cos\delta)^2 + R^2\sin^2\delta} \\ &= \frac{T^2}{1 - 2R\cos\delta + R^2(\cos^2\delta + \sin^2\delta)} \\ &= \frac{T^2}{1 + R^2 - 2R\cos\delta}. \end{aligned}$$

Writing $(1 - R)^2 = 1 + R^2 - 2R$ we get

$$\begin{aligned} |A|^2 &= \frac{T^2}{(1 - R)^2 + 2R(1 - \cos\delta)} \\ &= \frac{T^2}{(1 - R)^2}\frac{1}{1 + [4R/(1 - R)^2]\sin^2(\delta/2)}. \end{aligned} \tag{5.20}$$

The fringe pattern as a function of δ is shown in Figure 5.6. Maxima occur whenever $\delta = 2m\pi$ where m is an integer. At normal incidence ($\theta = 0$) this corresponds to an integer number of half wavelengths between the mirrors.

Although a Fabry-Pérot etalon will in general produce circular fringes in a similar manner to a thin film, it is now more usual to use the etalon at normal incidence (often with a laser as the incident beam) and to detect the transmitted intensity at angles very close to $\theta = 0$. By varying the spacing of the mirrors and using a photodiode to measure the transmitted intensity it is possible to measure the spectrum of the incident light, with each wavelength component producing its own set of peaks as a function of d.

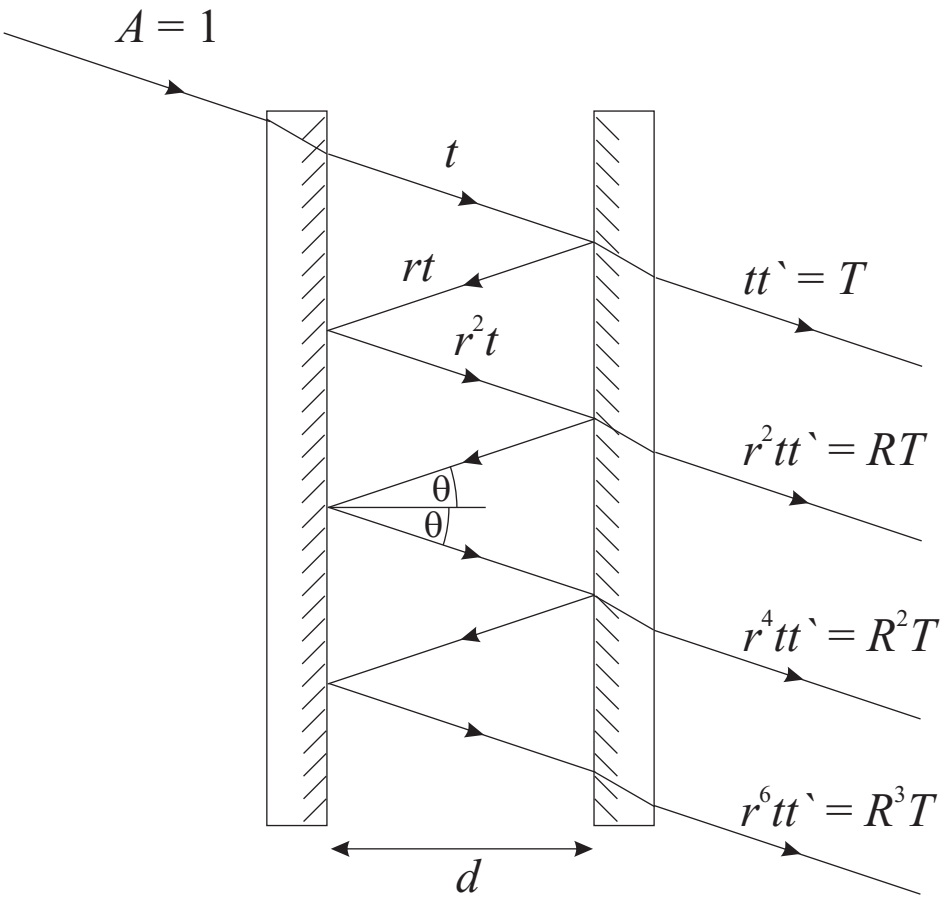


Figure 5.5: Multiple-beam interference in a Fabry-Pérot etalon of thickness d. Amplitudes of the emerging beams are shown; the amplitude reflection coefficient of the mirrors is r, where $r^2 = R$, and the amplitude transmission coefficients at the interfaces shown are t and t', where $tt' = T$.

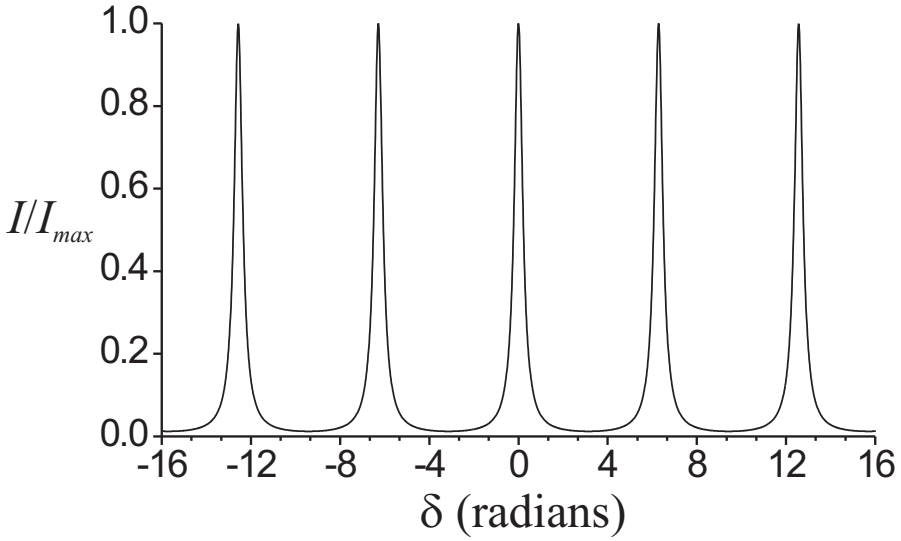


Figure 5.6: Transmission function for a Fabry-Pérot etalon with $R = 0.8$.

Clearly the sharpness of the peaks as a function of δ is critical in determining the resolution of the Fabry-Pérot spectrometer. From Eq. 5.20, the maximum value of I_{out}/I_{in} is $T^2/(1-R)^2$ when $\delta = 2m\pi$. We define the half-width of the peaks, $\delta_{1/2}$, as the change in δ required for the intensity

to reduce to half its maximum value, hence

$$\frac{4R}{(1-R)^2}\sin^2(\delta_{1/2}/2) = 1. \tag{5.21}$$

Since $\delta_{1/2}$ is typically small, $\sin(\delta_{1/2}/2) \approx \delta_{1/2}/2$, hence

$$\delta_{1/2} = \frac{1-R}{R^{1/2}}. \tag{5.22}$$

It is useful to define the finesse, $\mathcal{F}$ as the ratio of separation of successive peaks in δ (equal to 2π) to their full-width at half maximum ($2\delta_{1/2}$).

$$\mathcal{F} = \frac{2\pi}{2\delta_{1/2}} = \frac{\pi R^{1/2}}{1-R}. \tag{5.23}$$

The finesse is a measure of the quality of the etalon, and becomes high as R approaches 1. For a typical etalon with metal mirrors, R might be 0.95, corresponding to $\mathcal{F} = 61$. R can be increased further by using multiple-layer dielectric stacks as mirrors.

The chromatic resolving power of the etalon is defined as $\lambda/\Delta\lambda$, where $\Delta\lambda$ is the minimum wavelength difference between two spectral components which can just be resolved in the region of λ. Differentiating

$$\delta = \frac{4\pi d\cos\theta}{\lambda} \tag{5.24}$$

gives

$$d\delta = -\frac{4\pi d\cos\theta}{\lambda^2}d\lambda. \tag{5.25}$$

We assume that two spectral components can be distinguished if they are separated in δ by an amount $2\delta_{1/2}$, thus (forgetting the minus sign)

$$2\delta_{1/2} = \frac{4\pi d\cos\theta}{\lambda^2}\Delta\lambda. \tag{5.26}$$

Hence

$$\frac{\lambda}{\Delta\lambda} = \frac{2\pi d\cos\theta}{\lambda\delta_{1/2}}. \tag{5.27}$$

Recalling that, at a maximum in intensity, $2d\cos\theta = m\lambda$, we obtain

$$\frac{\lambda}{\Delta\lambda} = \frac{m\pi}{\delta_{1/2}} = m\mathcal{F}. \tag{5.28}$$

Since m is typically large, the resolving power can be very high. For example, for $R = 0.95$, $d = 5\,\text{mm}$ and $\lambda = 500\,\text{nm}$, the resolving power is more than 10^6, comparable with a grating spectrometer with a very large grating.

When working with large values of d, and hence large m, care must be taken when measuring the spectra of sources containing a significant range of wavelength components. If the spectrum is too wide, the signal from one end of the spectrum with order m will overlap with the signal from the other end of the spectrum with order $m+1$. The wavelength difference at which overlapping takes place is known as the *free spectral range*.

At normal incidence, the intensity maxima are given by

$$2d = m\lambda, \tag{5.29}$$

and differentiating this gives

$$-\frac{2d}{\lambda^2}\,d\lambda = dm. \tag{5.30}$$

Where m is large, we can use this equation to find the wavelength change $(\Delta\lambda)_{fsr}$ corresponding to $\Delta m = 1$, i.e.

$$\frac{2d}{\lambda^2}(\Delta\lambda)_{fsr} = \frac{m}{\lambda}(\Delta\lambda)_{fsr} = 1. \tag{5.31}$$

The free spectral range is thus given by

$$(\Delta\lambda)_{fsr} = \frac{\lambda}{m}. \tag{5.32}$$

For the example above, the free spectral range around $\lambda = 500\,\text{nm}$ is only $0.025\,\text{nm}$. Etalons are therefore most suitable for measuring the fine structure of narrow spectral lines.