Lecture Notes on

Optics

LIN Shaoze, George

following the lectures given by Axel Kuhn, Hilary 23

Oxford Physics, Second Year Undergraduate Lectures

Contents

Date of compilation: 05/10/23

Sources

The books stated in the second year handbook are

Optical Physics by A. Lipson, S. G. Lipson, and H. Lipson, 4th ed. , (Cambridge University Press, 2011), ISBN 9780521493451;

Optics by E. Hecht, 5th ed. , (Pearson, 2016), ISBN 9781292096964;

Modern Classical Optics by G. Brooker, (Oxford University Press, 2003), ISBN 0198599641.

Principles of Optics by M. Born and E. Wolf, 7th ed. , (Cambridge University Press, 2019), ISBN 9781108477437;

In addition to these, there are two more books recommended by the current lecturer:

Optics and Photonics by F. G. Smith, T. A. King, and D. Wilkins, 2nd ed. , (Wiley, 2007), ISBN 9780470017838;

Introduction to the Physics of Waves by T. Freegarde, (Cambridge University Press, 2013), ISBN 9780521197571.

Notation

Most symbols will be defined in the relevant sections and sometimes defined on diagrams, but there are some symbols that are common and will be used throughout the notes, which are formulated in the following table.

symbol	meaning
f	focal length
$\mathrm{FSR}_{\bar{\nu}}$	instrumental range
$\mathrm{INST}_{\bar{\nu}}$	instrumental width
k	wavenumber $2\pi/\lambda$
OPL	optical path length
$\mathscr{T}_F[u(x)](\beta)$	Fourier Transform
$u(x)$, $U(\beta)$	scalar amplitude
δ, Δ	phase difference
β	spatial frequency
λ	wavelength
$\bar{\nu}$	wavenumber $1/\lambda$

Vectors are in bold font.

Introduction

To understand a diverse range of optical physics, such as quantum optics, one has to be acquainted with classical optics, or the physics of waves. For most of the things we look at, we are not just looking into light, but to a diverse range of waves, in its full generality. Wave phenomena involves diffraction, refraction, interference, which applies to any type of waves that you can think of. One of the extensions of the studies of the propagation of light waves is quantum mechanics, where the matter waves propagates in exactly the same way as light waves.

To start this course, we shall first quickly revise geometric optics, then we move onto wave optics and the phenomenon of interference.

To do that, we actually have to take into account various degrees of simplification. The first thing we are looking at is Huygens' principle, and from there we get onto Fresnel-Kirchhoff diffraction, which is diffraction in the near field, *just* behind some optical element. But for most practical purposes, eventually we are interested in the phenomena that happen far away from our instrument, for instance, in an image produced by a camera, a telescope, or a microscope, and these are then determined by Fraunhofer diffraction, which we compare to a Fourier transform.

Therefore, the next topic is naturally Fourier methods, or the method of "division of wavefront", and the instruments one can produce. Building on top of what you already heard about diffraction gratings, we are going to look into a grating spectrometer into more detail: we are interested in the limits for the resolution, and the selection of gratings to achieve a certain goal. Then we look at spectrometry by division of amplitude. Imagine a wave hitting a half-silvered mirror, and splitting itself in two ways, which are then brought back to each other for interference. If we delay one of the two, for example sending one through an extra distance, thus adding a phase change, you can actually interfere a wave with itself that has been delayed. With that principle, we can build more complicated interferometers, for example, a Fabry-Perot interferometer.

A Fabry-Perot interferometer works by simply having two mirrors and having the light trapped between the two mirrors, which then motivates us to look at the very basic principles of a laser. A laser basically is a resonator for light, which is just a Fabry-Perot interferometer. We trap the light in there, with the light between the two mirrors interacting with the laser medium. The laser medium can be atoms or crystals. If we then combine the reservoir of the light and the resonator of the light, which will allow certain frequencies of the light to circulate inside in, this then provides the energy to have that laser running.

After that we look at polarised light. We shall note that waves in general are either transverse and longitudinal, and light waves specifically are transverse: the electric field vector is perpendicular to the direction of travel. This means that it must behave very differently to sound, which is a longitudinal wave oscillating in the direction of wave travel. This is what we shall investigate in the end. However it is important to note that at the start, we shall assume that light only has a scalar amplitude, i. e. we leave the transverse wave properties of light till the end.

So what is optics good for? The obvious answer would be imaging, and that would be imaging in the widest sense: microscopes, telescopes, and cameras. We can also use optics to actually do something that produces images in a projection, where the best example for it is to use optics to machine something simple — for example using lithography to produce computer chips. On top of that, we have spectroscopy, which uses coherent waves like lasers. Then we have optics used on modern display technologies, and optical coatings on glasses and goggles, which aims to eliminate ghost images due to multiple reflections. Optical fibres are now extremely useful in telecommunications. Last but not least, there is an entire field of physics that looks at the quantum behaviour of light, where we deal with single photons with experiments where we can produce photons one by one, which then leads into the area of quantum computing and quantum communication, where a single photon is an information carrier for a single quantum bit that can exist in a superposition of two quantum states.

Now we shall quickly revise geometric optics.

1 Geometric Optics

§1. Laws of Geometric Optics

We shall survey through some of the very basic laws of geometric optics that is covered with more detail in the first year optics course.

When a ray hits a boundary, the ray can be transmitted or reflected. This is illustrated in figure 1.1. The angle i is the **angle of incidence**, the angle r_l is the **angle of reflection**, and the angle r_a is the **angle of refraction**, usually denoted as the angle that the ray makes with the normal. The ray that hits the boundary satisfies two main laws: the law of reflection

$$i \;=\; r_l \tag{1.1}$$

and the law of refraction, or **Snell's Law**

$$n_1 \sin i \;=\; n_2 \sin r_a \tag{1.2}$$

where n_1 and n_2 are the refractive indices of the two media. An immediate corollary is that, when $n_2 \;<\; n_1$ the light ray will be deviated from the normal, and therefore there exits a certain critical angle i_C, at which the refracted angle r_a is 90°,

$$i_C \;=\; \arcsin\left(n_2/n_1\right). \tag{1.3}$$

If the angle of incidence is larger than i_C, then there will be no refracted beam and all the intensity is reflected, which is a phenomenon called **total internal reflection**.

In full generality, the ray satisfies **Fermat's principle**, which suggests that the light ray that passes two points along its propagation path must travel between the two points via the shortest optical path in between. To understand what "shortest" means, there are two interpretations:

- the **optical path length**, OPL $\;=\;$ $n \times \Delta d$, where n is the refractive index and Δd is the geometrical path length, is minimised;
- the light travels via the path that takes the shortest time.

To illustrate the equivalency of the two statements, note that the time is given as

$$t \;=\; \frac{\text{geometrical path}}{\text{light speed}} \;=\; \frac{\Delta d}{(c/n)} \;=\; \frac{n \times \Delta d}{c} \;=\; \frac{\text{OPL}}{c}, \tag{1.4}$$

and since c is a constant, minimisation of t is indeed equivalent to minimisation of the optical path length OPL.

Summary

1. When light hits a boundary it follows the law of reflection

$$i = r_l \tag{1.5}$$

and Snell's law

$$\sin i \;=\; n \sin r_a, \tag{1.6}$$

which, formulated more generally, gives Fermat's principle, suggesting that light always travels a path that minimises the time and the optical path length OPL $\;=\;$ $n \times \Delta d$.

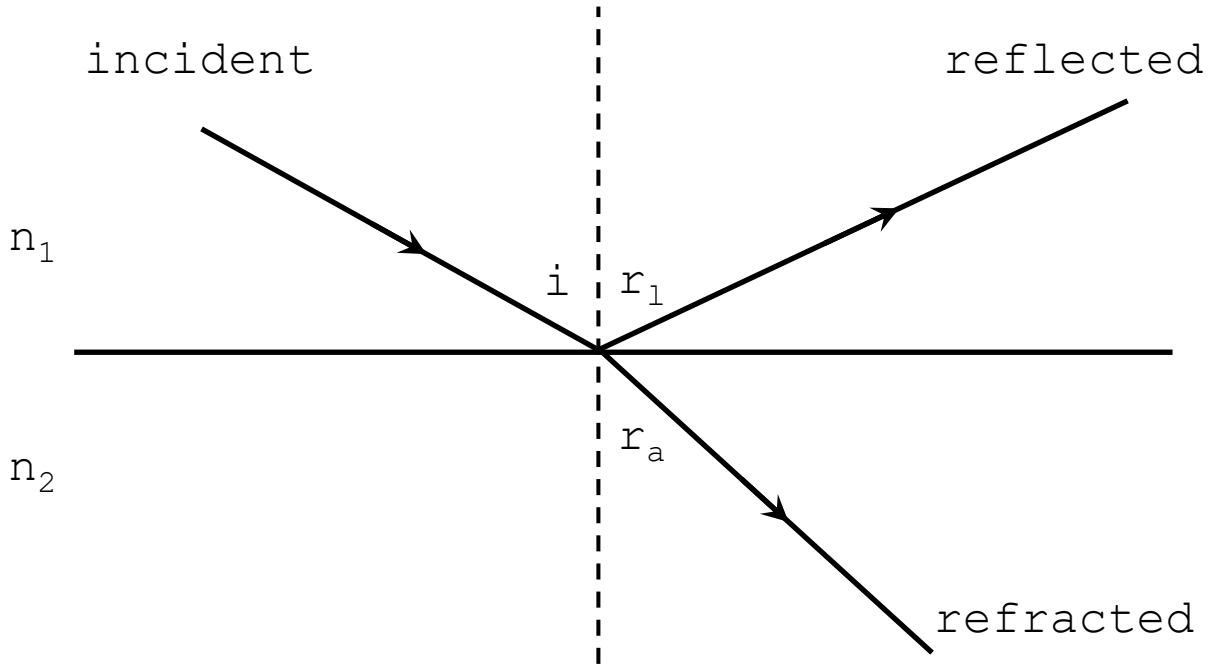

Figure 1.1: The action of a geometric ray at a boundary.

§2. Optical Fibres

An invention that uses Snell's law and total internal reflection is an **optical fibre**, which transmits light by having it totally internally reflecting in a solid core. The construction of such a fibre is given in figure 2. Note that the fibre is cylindrical so we are only looking into a cross-section of it. To give an idea of the scale of the construction, the diameter of the fibre is usually of the order of a few hundred micrometers.

The main idea of an optical fibre is to have light entering one side switching on and off with a very short period. The pattern of the on and off switches encodes 0s and 1s in digital information, and light pulses are transmitted through the fibre which is then detected and changed back into electrical signal on the other side of the fibre. We can therefore see that the key here is to make sure all light pulses travel at the same transverse velocity: as if this is not achieved, then a pulse of a shape of "on-off-on" with the final "on" signal travelling much faster than the first may "catch up" with the first signal, making the middle "off" signal missing in our detection. This problem is called **modal dispersion**.

To make sure that light that transmit through the optical fibre travels at the same transverse speed, we need to have the light mostly travelling along the fibre instead of bouncing back and forth in the core; and this is done by having the core and the cladding having similar refractive indices n_{co} and n_{cl} (where $n_{\text{co}} > n_{\text{cl}}$), giving a very small range of angles for light to transmit across the fibre by total internal reflection, leaving light travelling at large angles to escape the fibre through the cladding. Or, using the parameters given in figure 2, we need the angle χ to be very close to 90°.

Let us do some quantitative analysis. Since the light inside the core must totally internally

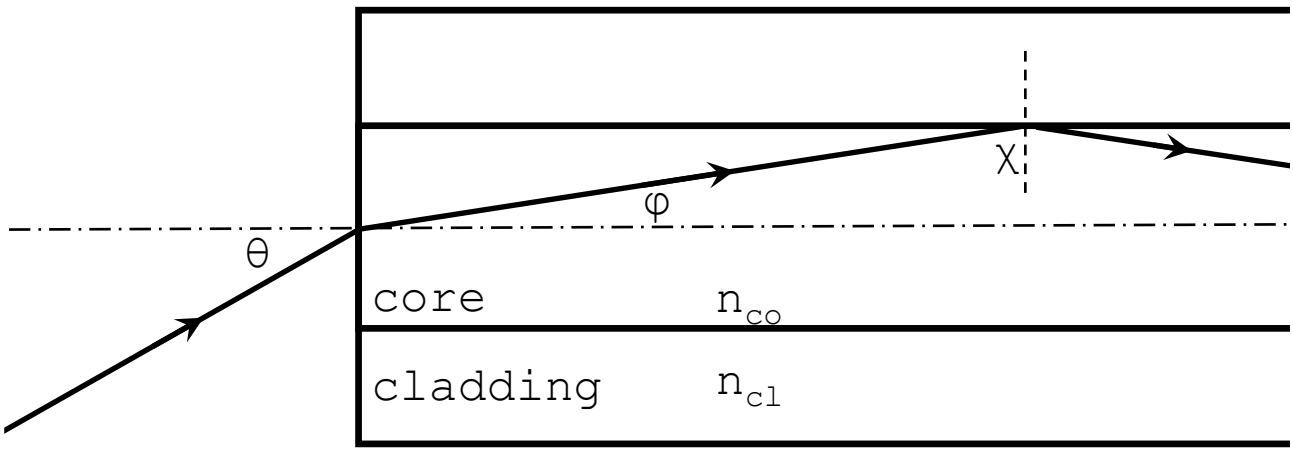

Figure 2.1: The construction of an optical fibre.

reflect, there is a lower limit of χ, which is the critical angle at the core-cladding boundary

$$\chi_{\min} \;=\; \arcsin\left(n_{\rm cl}/n_{\rm co}\right). \tag{2.1}$$

This then casts an upper limit of φ,

$$\varphi_{\max} \;=\; \arccos\left(n_{\rm cl}/n_{\rm co}\right), \tag{2.2}$$

hence, by Snell's Law, we have

$$\theta_{\max} \;=\; \arcsin\left(\sqrt{{n_{\rm co}}^2 - {n_{\rm cl}}^2}\right). \tag{2.3}$$

Sometimes one uses $\theta_{\max}$ to define the **numerical aperture** for an optical fibre

$$\mathrm{NA} \;=\; \sin\theta_{\max} \;=\; \sqrt{{n_{\rm co}}^2 - {n_{\rm cl}}^2}, \tag{2.4}$$

which is a parameter that limits the acceptance angle of this optical fibre.

Summary

1. Optical fibres are deigned to transmit information by sending light pulses from one side and receiving light pulses from the other side. It is made of a core and a cladding with very similar refractive indices to trap light with a very narrow range of angle χ to transmit through the fibre by total internal reflection. A commonly used parameter for a fibre is the numerical aperture
$$\mathrm{NA} \;=\; \sin\theta_{\max} \;=\; \sqrt{{n_{\rm co}}^2 - {n_{\rm cl}}^2}, \tag{2.5}$$
which limits the angle of incidence for the light entering the fibre.

§3. Imaging by a Sphere

Now let us revise the rules of basic ray-tracing that one should be acquainted in the first year course, including some of the approximations that one is taking when doing so. This is best done by using an example. Let us consider a parallel bundle of rays entering a sphere with

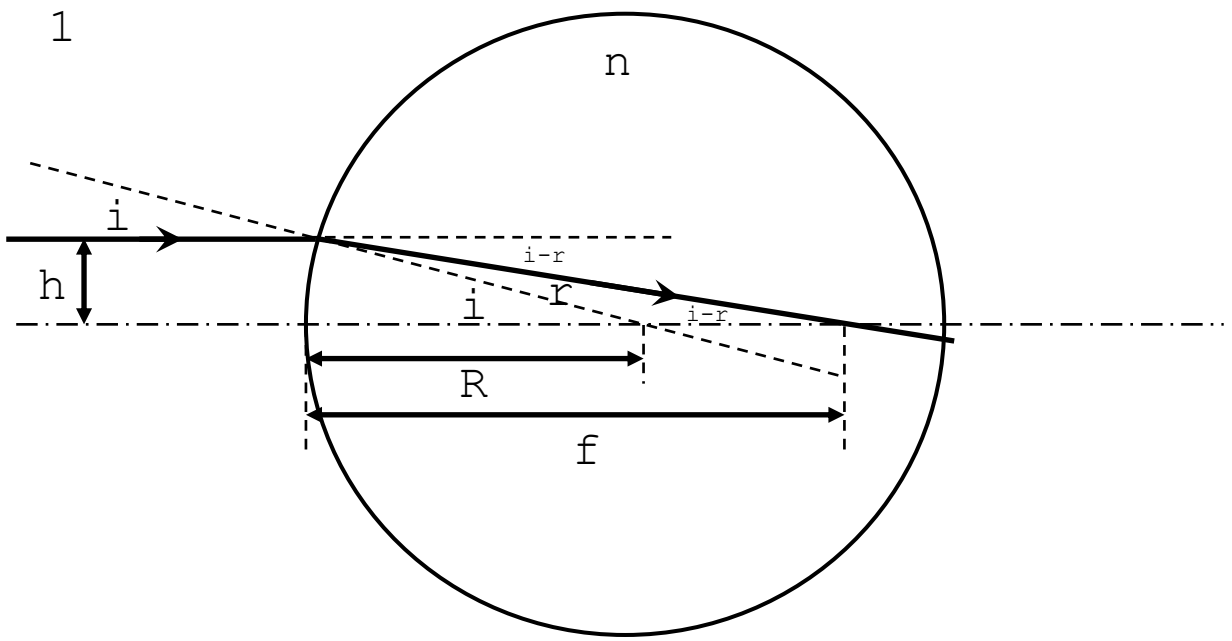

Figure 3.1: Imaging by a sphere.

refractive index n from a medium with refractive index 1, which can be thought as air, as a good approximation. Let us also suppose that the bundle is thin and the centre of the bundle is incident normally on the sphere. The first step would be to draw a diagram, and add a few geometrical parameters, which is done in figure 3.1.

When faced with an optical system as such, the best way is tackle it is to take one of the rays in the parallel bundle that is off-centred, let us say the ray at a distance h from the centre, also shown in figure 3.1. This is then focussed by the sphere and meets the centre ray at a distance f after its entrance. We use the paraxial approximation throughout, which means that $h \ll R$ and i is small, and hence $i = h/R$ and $\sin i = i$. Also we have $i - r$ small, and therefore $i - r = h/f$. Hence,

$$\frac{h}{R} = i = nr = n\left(i - \frac{h}{f}\right) = nh\left(\frac{1}{R} - \frac{1}{f}\right). \tag{3.1}$$

Rearranging this equation, we have

$$\frac{1}{f} = \frac{n-1}{n} \times \frac{1}{R}, \tag{3.2}$$

which is independent of h. This suggests that the sphere acts somewhat similar to a convex lens, and equation 3.2 acts as a simple lens maker's equation for the system.

Now let us see the limit of the paraxial approximation. If we rotate the system, then since the system has spherical symmetry, it is impossible for all light at any angle image at at the same point, instead the point that they focus on must form a sphere with radius f, illustrated by 3.2, which cannot be predicted using the paraxial approximation.

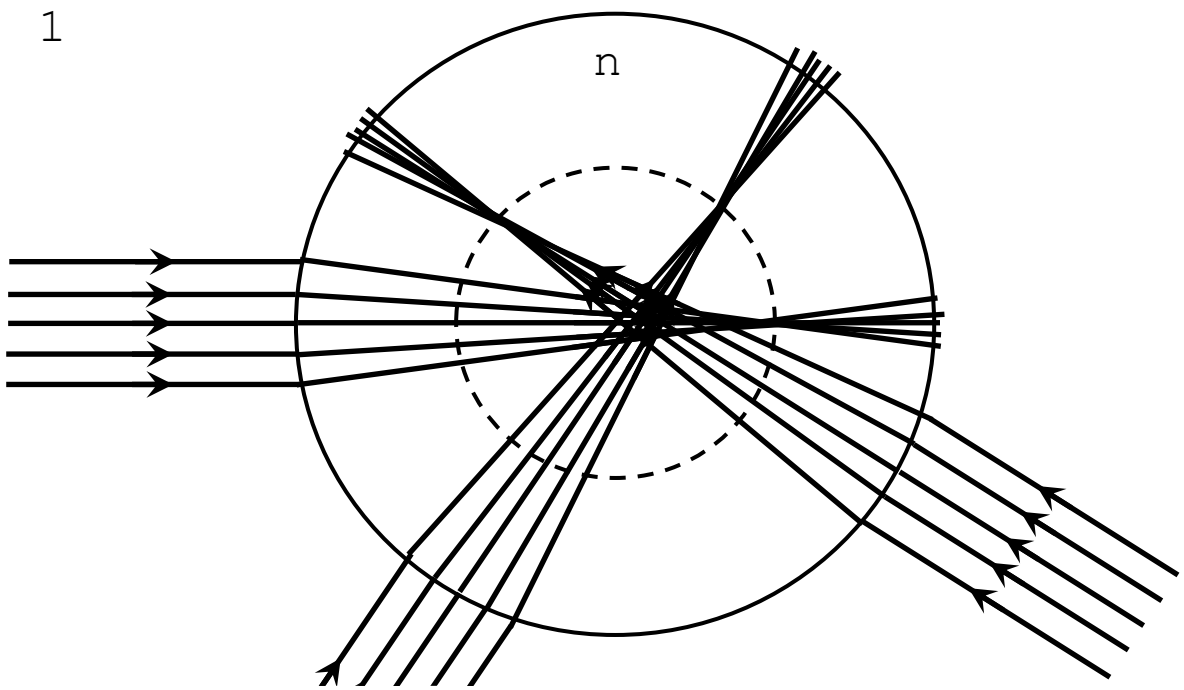

Figure 3.2: Spherical symmetry of the setup.

From here we are not far from a thin convex lens, which can be thought of two of such spherical surfaces sandwiched together. We shall look at these in the next section.

Summary

1. When faced with a simple system that we need to do ray-tracing, use an off-centred ray and paraxial approximation. However we do need to note the limitations of paraxial approximation and use the symmetry of the system when the paraxial approximation ever ceases to work.

§4. Thin Lens

Basic Properties of the Thin Lens

The first property of a thin lens is that an on-axis light ray through it follows the **lens maker's equation**, i. e. all the light that originates at a point on the principal axis u will hit the principal axis again on the other side of the lens at a distance v, where

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}, \tag{4.1}$$

where f is the **focal length** of the thin lens. If we consider a thin lens with refractive index n is bounded by two spherical surfaces with radii R_1 and R_2, and light enters the system from a medium with refractive index 1, then

$$\frac{1}{f} = (n-1) \times \left(\frac{1}{R_1} + \frac{1}{R_2} \right). \tag{4.2}$$

The derivation can be found in the first year lecture notes, and hence will not be repeated.

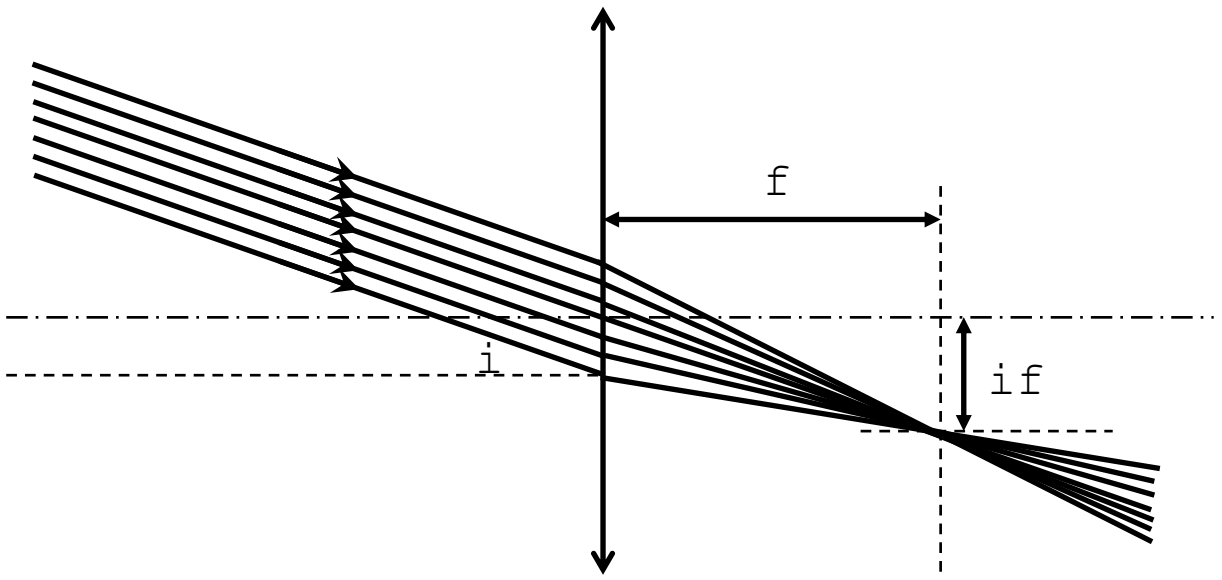

Figure 4.1: A parallel bundle of rays incident obliquely on a thin lens.

The lens maker's equation illustrates what happens for a light source at the principle axis, but as we have seen previously that the most simple way of ray-tracing is to consider an off-centred ray. To work out what happens for these rays, we have four simple rules (under the condition that we can use paraxial approximation):

- the ray that passes through the centre of the thin lens goes straight through the lens;
- the rays that are parallel to the principle axis upon entrance are focused to the **focal point** of the thin lens, located on the principal axis on the other side of the thin lens that is a focal length away from the lens;
- the rays that passes through the focal point before entering the lens become a parallel bundle of rays upon exit;
- a parallel bundle of rays incident on the lens at an angle i will be focussed at a focal length away, at a distance $x = if$. This is illustrated in figure 4.1. An qualitative proof of this is shown in §6.

We have stated Fermat's principle in the previous section yet we have not had a chance to apply it. We shall do that now in the context of a thin lens, to try and proof lens maker's equation in the first year in an alternative way.

Lens Maker's Equation from Fermat's Principle

Let us now consider a ray that originates from point A on the principal axis and hits the principal axis again at point B, through the thin lens, hitting it at point C. This is illustrated in figure 4.2. Let us now consider the optical path length of the light as a function of h. This gives

$$\mathrm{OPL}(h) = \sqrt{h^2 + u^2} + \sqrt{h^2 + v^2} + d(h) \times (n - 1), \tag{4.3}$$

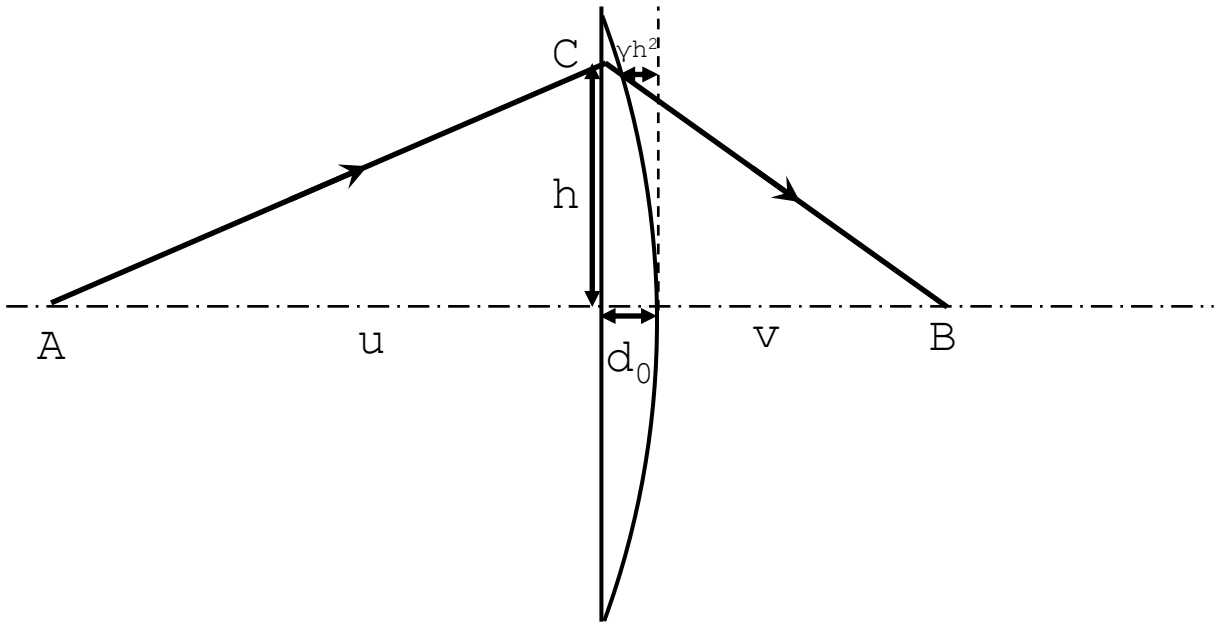

Figure 4.2: Using Fermat's principle to find the lens maker's equation.

where we assume the ray travels in the lens for a short section $d(h)$. Note that the optical path length in that section is longer than the optical path length in air, and therefore we need to correct to that by first subtracting the optical path length in air, then add onto the optical path length in the lens, and hence gives a total contribution of $d(h) \times (n-1)$. If we let the centre of the length to have thickness d_0, then we may approximate $d(h) = h_0 - \gamma h^2$, just by approximating the lens surface as a parabola. Using Fermat's principle, this optical path length must be the same as the optical path length of the ray that travels though the centre of the lens, as only if that's the case, light travelling through both paths are allowed. This gives

$$\text{OPL} = \sqrt{h^2+u^2} + \sqrt{h^2+v^2} + \left(d_0 - \gamma h^2\right) \times (n-1) = u + v + d_0 \times (n-1). \tag{4.4}$$

Next using the paraxial approximation, so we can binomially expand the square root to second order of h/u and h/v since both ratios are much smaller than unity, we have

$$\frac{1}{u} + \frac{1}{v} = 2\gamma \times (n-1), \tag{4.5}$$

the lens maker's equation with $f = 1/[2\gamma \times (n-1)]$. Just by considering the simple model where the lens is spherical on one side with radii R and flat on the other side, we have

$$\gamma h^2 = \sqrt{h^2+R^2} - R^2 = h^2/(2R) \quad \Rightarrow \quad \gamma = 1/(2R), \tag{4.6}$$

and hence we have the lens maker's formula as

$$\frac{1}{u} + \frac{1}{v} = (n-1) \times \frac{1}{R}. \tag{4.7}$$

A more sophisticated analysis will then give the full lens maker's formula that assumes that the lens is spherical on both sides with different radii.

Summary

1. There are four simple rules for ray-tracing for light that passes through a lens, described in the main section of the text, used for light rays that does not path the principal axis.

2. For light that passes through the principal axis, the path follows the lens maker's equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}. \tag{4.8}$$

§5. Compound-Lens Systems

Aperture Stop

Previously we have talked about thin lens, however thin lenses in real life are far from ideal. For example, ideally, using the thin lens to image a parallel bundle of rays along the principle axis will lead to the rays focussed at a focal length f, however this is usually not the case in the real world: some rays will meet shorter than the focal length and some rays will meed further than the focal length. This is called **spherical aberration**. For a parallel bundle of rays with the central ray hitting the lens at the centre, this effect can also happen, and is called **coma aberration**. To reduce these effects, sometimes when designing a system, we block some of the light such that light only travels through the centre part of the lens. These blocks may limit the amount of light travelling through the optical setup, and the cross-section through the block where the amount of light is limited is called the **aperture stop**. If there are no mechanical blocks, then the aperture stop is just the objective lens of the system.

For ray-tracing in a complicated system, we usually consider how two rays travel through the system. They are

- the chief ray, the ray that hits the centre of the aperture stop;
- and the marginal ray, the ray that hits the edge of the aperture stop.

Of course, describing the method of ray-tracing without examples is difficult, and therefore we shall demonstrate these effects with the compound microscope and the astronomical telescope.

The Compound Microscope

Let us look at a simple model of the compound microscope used a small object, say with a height h_o, illustrated in figure 5.1, which is simply made up of an objective lens with a focal length f_o and an eyepiece lens f_e which are a **tube length** L apart. For simplicity let us not cover any parts of the lens and assume that it is ideal, and therefore the aperture stop would just be the objective lens. For a microscope we simply assume that the object we are looking at is very small, and as a result both the marginal rays can be thought as they are emitted from a point on the principal axis. Therefore we assume that they meet up with the chief ray at a distance v governed by the lens maker's equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f_o}, \tag{5.1}$$

therefore forming an image with height h_i. An alternative way of thinking about this is to trace the ray that is parallel to the principal axis when the rays enter the objective lens, and the result would be the same — the parallel ray meets the principal axis at a distance f_o, and will meet the chief ray at where the image is formed. To magnify again, we may add an eyepiece

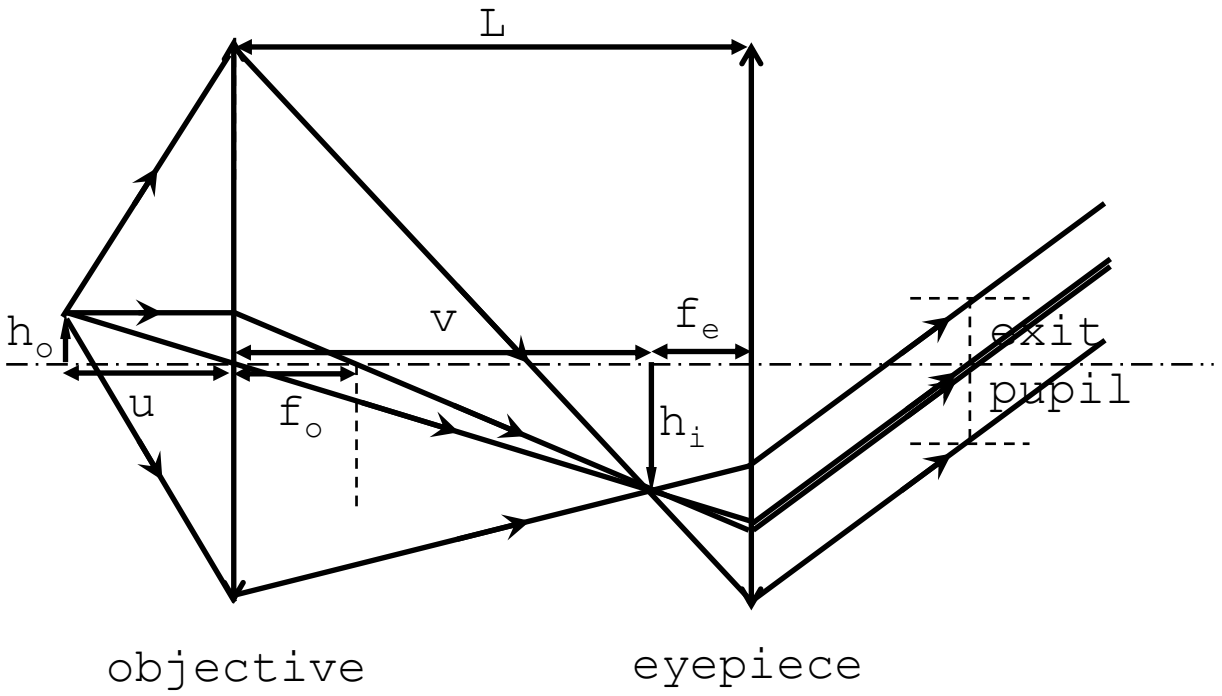

Figure 5.1: A simple model of a compound microscope.

lens, which we place at a distance f_e away from the image. Then since they all meet at a point at the focal length f_e, using the final rule in §2 in reverse, we note that they must be a parallel bundle of rays upon exit. When the chief ray hits the principal axis, we denote the cross-section of the exiting bundle of rays as the **exit pupil** (note that the exit pupil is simply the image of the aperture stop, in this case it is the image of the objective lens). We denote the **objective magnification** M_o as the ratio of the sizes of the image and the object, which in this case is given by, by the theory of geometry of similar triangles,

$$M_o \;=\; h_i/h_o \;=\; v/u. \tag{5.2}$$

We note that there is no image formed since the output is a parallel bundle of rays, and therefore we need a third lens to form another image, which in this case is simply the eye. Therefore, in order to get all the detail of the object into the eye and hence form an image on the retina, one need to make sure that the eye is placed at the exit pupil, with the size of the lens of the eye to be larger than the diameter of the exit pupil, but the closer the better, to get a good magnification. If the image is to be projected onto a charge-coupled device (CCD) camera (which is exactly the same technology used as phone cameras), then we will need to add a third lens before the sensor and make sure that the size of the third lens is at least as large as the exit pupil, and the image will then be formed at the focal length of the third lens.

We note that, since we have a parallel bundle of rays on exit, it is difficult to come up with a definition of the magnification for the compound microscope as a whole like how we wrote down the objective magnification. Instead we have to change our logic and write down an "angular magnification", that is, the ratio of the *angular size* of the object and what is seen by the eye. Noting that the largest angular size subtended by the object without the microscope is

$$\alpha \;=\; h_o/D \tag{5.3}$$

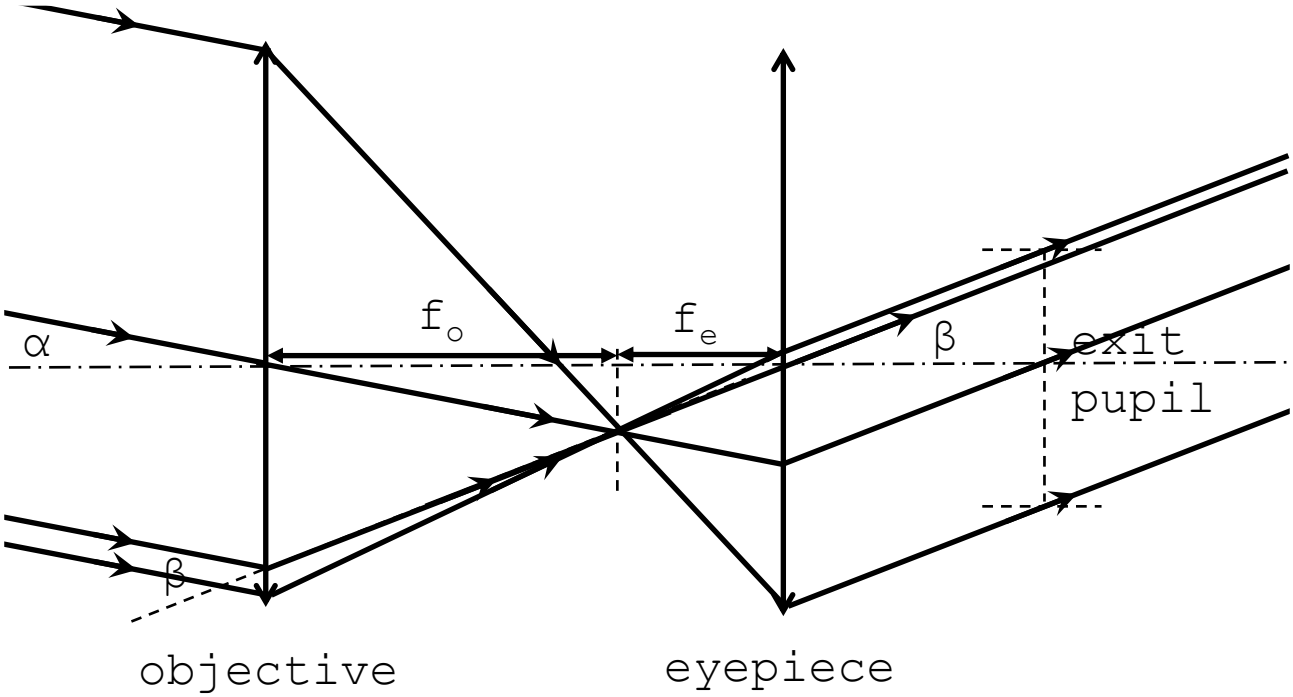

Figure 5.2: A simple model of an astronomical telescope.

where $D \approx 25\,\text{cm}$ is the **near point** of the eye, that is, the shortest distance of an object that the human eye can resolve. On the other hand, the outgoing ray has an angle

$$\beta = h_i/f_e, \tag{5.4}$$

using the rule that a lens does not change the direction of a ray that passes through its centre. Then, a sensible overall magnification of the microscope is given as

$$M = \frac{\beta}{\alpha} = \frac{h_i}{f_e} \times \frac{D}{h_o} = M_o M_e, \qquad M_e = \frac{D}{f_e} = \frac{h_i}{f_e} \times \frac{D}{h_i}. \tag{5.5}$$

Here M_e is the **eyepiece magnification**, defined analagous to the overall magnification. Since real compound microscopes have focal lengths f_o and f_e much shorter than the tube length L, we may make the approximation $v \approx L \gg f_o$. Then, using equation 5.1, we have $u \approx f_o$, and thus we may formulate the objective magnification as

$$M_o = L/f_o \tag{5.6}$$

and therefore the overall magnification is

$$M = M_o M_e = (LD)/(f_o f_e). \tag{5.7}$$

The Astronomical Telescope

The next example is an astronomical telescope, shown in figure 5.2. The astronomical telescope is used to image distant objects e. g. stars or planets, which, since they are so far away, we simply assume that the light from them is a parallel bundle of ray that comes from the same angle. This is formed with an objective lens with a focal length f_o and an eyepiece lens with a

focal length f_e that are $f_o + f_e$ apart, and the image is formed at a distance f_o from the objective and f_e from the eyepiece. Again we do not mask out any light so the aperture stop is simply the objective lens. Noting that the rays leaving the system in an astronomical telescope is also a parallel bundle, similar to the case of a compound microscope, giving an angular magnification

$$M = \beta/\alpha = f_o/f_e, \tag{5.8}$$

where the final equality stated uses the geometry of similar triangles. We can also match the exit pupil with the lens of the eye, however in practice we would like to look into the camera right next to the eyepiece lens. To do this, we would need to modify the setup by adding a **field lens**, placed at exactly where the image forms, such that the chief ray crosses the principal axis at exactly the position of the eyepiece, i. e. we select the field lens to have a focal length f_f where

$$\frac{1}{f_f} = \frac{1}{f_o} + \frac{1}{f_e} \tag{5.9}$$

by the lens maker's equation.

Summary

1. To avoid aberrations associated with the thin lens, we block some of the light. This block can limit the light going through the system, and the cross-section where the amount of light is most limited is called the aperture stop. When we do ray-tracing for complicated systems, we usually only focus on the chief ray and the marginal ray.

2. The compound microscope magnifies a small object. The exit pupil is the cross-section of the bundle of light rays where the chief ray crosses the principal axis. We may define the magnifications

   - objective magnification $M_o = h_i/h_o = v/u = L/f_o$;
   - eyepiece magnification $M_e = D/f_e$;
   - overall magnification $M = M_o M_e = (LD)/(f_o f_e)$.

3. The astronomical telescope magnifies light from a distant objects where the light rays are parallel. The angular magnification is defined as $M = \beta/\alpha = f_o/f_e$. To place the exit pupil right next to the eyepiece lens, we need to add a field lens at the position of the image.

§6. Ray-Transfer Matrices

Ray-Vectors and Ray-Transfer Matrices

We now develop a formalism where we can ray-trace more algebraically under paraxial approximation. Note that a light ray is a straight line, and therefore any point on a ray can be parametrised by two parameters, the distance away from the principal axis x and the angle between the ray and the principle axis θ. These two parameters naturally forms a vector, called a **ray-vector**

$$\begin{pmatrix} x \\ \theta \end{pmatrix}. \tag{6.1}$$

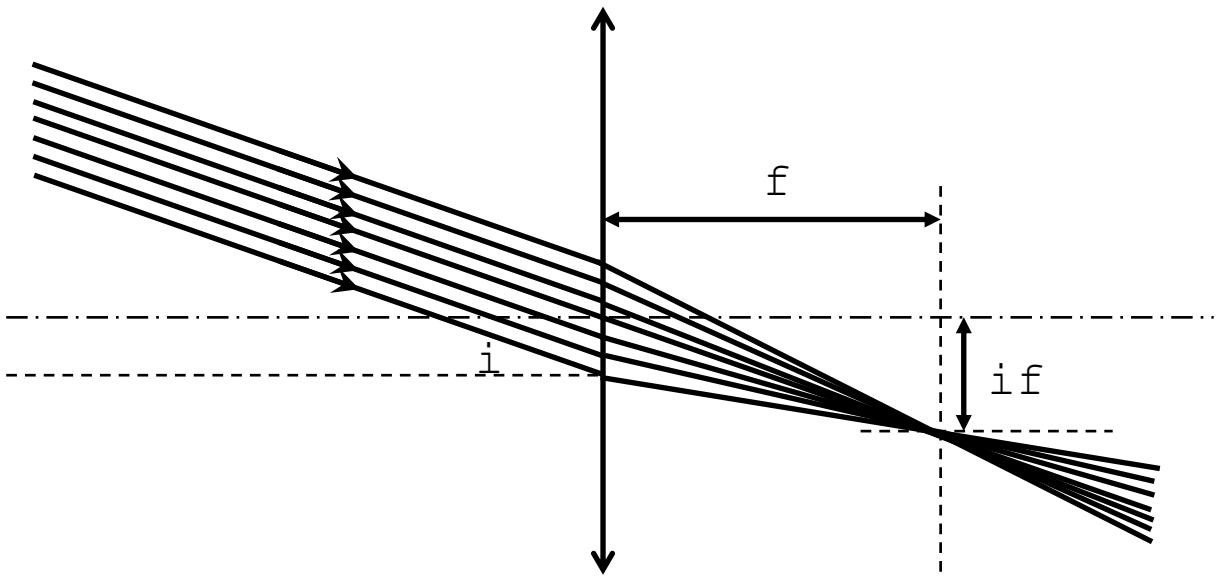

Figure 6.1: A repeat of figure 4.1: a parallel bundle of rays incident obliquely on a thin lens.

To describe how the ray evolves *along* the principal axis, we use matrices. As an example, propagation of a ray with a ray-vector $(x_1, \theta_1)^{\mathrm{T}}$ by a distance d along the principal axis gives θ unchanged, and the distance of the ray away from the principal axis $x_2 = x_1 + d\theta_1$. This leads to

$$\begin{pmatrix} x_2 \\ \theta_2 \end{pmatrix} = S_d \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix}, \quad S_d = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}, \tag{6.2}$$

where S_d is the **ray-transfer matrix** of light propagation through distance d.

Ray-Transfer Matrix of a Thin lens

We now derivethe ray-transfer matrix of propagation through a thin lens. The first observation would be that the ray will have the distance away from the principal axis x unchanged after going through the thin lens, and therefore we only need to care about the angle. An on-axis ray satisfies the lens maker's equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}, \tag{6.3}$$

which forms a good starting point for our analysis. A ray with ray-vector $(x, \theta_1)^{\mathrm{T}}$ before passing through the lens and $(x, \theta_2)^{\mathrm{T}}$ after passing through the lens has $u = x/\theta_1$ and $v = -x/\theta_2$ (note that the ray points *towards* the principal axis after focussed by the lens which gives a negative θ_2), and therefore the lens maker's equation becomes

$$\frac{\theta_1}{x} - \frac{\theta_2}{x} = \frac{1}{f} \quad \Rightarrow \quad \theta_2 = \theta_1 - \frac{1}{f} \times x, \tag{6.4}$$

which means that a lens with a focal length has a ray transfer matrix

$$S_f = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}. \tag{6.5}$$

Equipped with this technology we can ray-trace algebraically through a system of lenses with different focal lengths separated at different distances. For example, a ray with ray vector $(x, \theta_1)^{\mathrm{T}}$ at a distance d_2 from a lens with a focal length f_2, passing through a lens with a focal length f_1 separated from the first lens by d_1, then travelling through a distance d_0, has an exiting ray vector

$$\begin{pmatrix} x_2 \\ \theta_2 \end{pmatrix} \;=\; S_{d_0} S_{f_1} S_{d_1} S_{f_2} S_{d_2} \begin{pmatrix} x_1 \\ \theta_1 \end{pmatrix}. \tag{6.6}$$

As we can see, this formalism takes care of all the rays — not just a number of special rays — which means that it forms a more powerful method than the previous rules, which only deals with very special rays propagating through a lens.

To give a concrete example of the usage of this formalism, let us try to show that a parallel bundle of rays passing through a lens at an angle i will be focused at a distance f on the other side, illustrated by figure 6.1 (which is a copy of figure 4.1, where this problem is first encountered). The input ray-vector is given by $(x, i)^{\mathrm{T}}$ with any arbitrary x, and this is travelled through a lens with focal length f, and then propagated through a distance f along the principal axis. This means that the output vector can be calculated using

$$\begin{pmatrix} 1 & f \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} x \\ i \end{pmatrix} \;=\; \begin{pmatrix} if \\ i - x/f \end{pmatrix}. \tag{6.7}$$

Indeed,the rays meets at a distance if away from the principal axis but with different exiting angles. We note that this suggests that bundles of light that are incident at different angles will be focussed at different distances from the principle axis, and sometimes this phenomena is referred to as "a thin lens changes angles to distances".

Summary

1. A light ray can be described by a ray-vector $(x, \theta)^{\mathrm{T}}$, and the light ray passing through different optical elements can be described by ray-transfer matrices. The ray-transfer matrix for a light ray propagating through a distance d along the principal axis is given by

$$S_d \;=\; \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}. \tag{6.8}$$

2. The ray-transfer matrix of a thin lens is

$$S_f \;=\; \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}, \tag{6.9}$$

   and light travelling through many lenses can be described by an initial ray-vector acted on the product of many ray-transfer matrices, which is applicable to any light ray.

2 Building a Scalar Theory of Light

§7. Maxwell's Equations and the Wave Equation of light

Previously we have seen what we can do by just treating light as "rays", but in fact light are waves, and geometric optics cannot describe light travelling through very small apertures, where the image is very limited by the diffractive aspects of light as a wave. To solve these problems, we need a proper treatment of what light is, which means back to the drawing board and look into Maxwell's equations.

Maxwell's Equations and the Wave Equation for Light in Linear Materials

In linear materials with no free charges or currents, Maxwell's equations read

$$\text{(Gauß)} \qquad \operatorname{div}\mathbf{D} \;=\; 0; \tag{7.1}$$

$$\operatorname{div}\mathbf{B} \;=\; 0; \tag{7.2}$$

$$\text{(Faraday)} \qquad \operatorname{curl}\mathbf{E} \;=\; -\partial_t\mathbf{B}; \tag{7.3}$$

$$\text{(Ampère)} \qquad \operatorname{curl}\mathbf{H} \;=\; \partial_t\mathbf{D}. \tag{7.4}$$

Using the above Maxwell's equations, and the relations in linear materials

$$\mathbf{D} \;=\; \varepsilon_0\varepsilon_r\mathbf{E}; \tag{7.5}$$

$$\mathbf{B} \;=\; \mu_0\mu_r\mathbf{H}, \tag{7.6}$$

we have

$$\operatorname{curl}\operatorname{curl}\mathbf{E} \;=\; -\partial_t\operatorname{curl}\mathbf{B} \;=\; -\mu_0\mu_r\partial_t\operatorname{curl}\mathbf{H} \;=\; -\mu_0\mu_r\partial_t^2\mathbf{D} \;=\; -\mu_0\mu_r\varepsilon_0\varepsilon_r\partial_t^2\mathbf{E}. \tag{7.7}$$

However, we also have, using vector identities and Gauß's Law

$$\operatorname{curl}\operatorname{curl}\mathbf{E} \;=\; \operatorname{grad}\operatorname{div}\mathbf{E} - \Delta\mathbf{E} \;=\; -\Delta\mathbf{E}, \tag{7.8}$$

where Δ is the Laplacian with respect to the three dimensions of space. Equating equations 7.7 and 7.8, we obtain the wave equation

$$\left(\Delta - \frac{1}{v^2}\partial_t^2\right)\mathbf{E} \;=\; \mathbf{0}, \qquad v \;=\; \frac{1}{\sqrt{\mu_0\mu_r\varepsilon_0\varepsilon_r}} \;=\; \frac{1/\sqrt{\mu_0\varepsilon_0}}{\sqrt{\mu_r\varepsilon_r}} \;=\; \frac{c}{n} \tag{7.9}$$

for light in a linear material.

From a Wave Equation to a Helmholtz Equation

To solve the wave equation, one approach is to separate the time components out of the equation to achieve a Helmholtz equation, which is dependent on space purely. The ansatz that separates the time part out in this case is

$$\mathbf{E} \;=\; \mathbf{E}_{\text{sp}}\mathrm{e}^{-\mathrm{i}\omega t}, \quad \omega \;=\; 2\pi/f, \tag{7.10}$$

where $\mathbf{E}_{\text{sp}}$, or the spatial part of the electric field, is a function dependent on time only, and f is the frequency of the light. Substituting this into the wave equation gives a Helmholtz equation

$$\left(\Delta + \frac{\omega^2}{v^2}\right)\mathbf{E}_{\text{sp}}\mathrm{e}^{-\mathrm{i}\omega t} \;=\; \mathbf{0} \quad\Rightarrow\quad \left(\Delta + k^2\right)\mathbf{E}_{\text{sp}} \;=\; \mathbf{0}, \quad k \;=\; \frac{\omega}{v} \;=\; 2\pi f \times \frac{n}{c} \;=\; \frac{2\pi}{\lambda}, \tag{7.11}$$

where λ is the wavelength of the light. Note that since the frequency of light is invariant in a linear material, the wavelength of the light in a linear material is $\lambda = c/(nf) = \lambda_0/n$, where λ_0 is the wavelength of light in a vacuum.

The Wave Vector and the Wavenumber

We note that in the Helmholtz equation we have got an expression for the **wavenumber** k, which is the modulus of the **wave vector**. We shall see that, in special solutions such as the plane wave solutions, the wave vector $\mathbf{k}$ points at the direction of propagation of the wave. We are also able to define the wavenumber in a vacuum k_0 satisfying $k_0 = 2\pi/\lambda_0$, which gives rise to the relation between k and k_0 as

$$k = \frac{2\pi}{\lambda} = \frac{2\pi n}{\lambda_0} = nk_0. \tag{7.12}$$

We note that it is also very common to define

$$\bar{\nu} = 1/\lambda \tag{7.13}$$

as our wavenumber, and therefore in any literature in optics, when the term "wavenumber" appears, we will need to take a closer look for

- whether it is referring to $k = 2\pi/\lambda$ or $\bar{\nu} = 1/\lambda$,
- and whether it is defined using the wavelength in a vacuum or the wavelength in the linear medium.

In this set of notes, as we are looking through many different fields of optics, we shall refer "wavenumber" to k and $\bar{\nu}$ interchangeably, following the most common object to use in each field of optics that we look into.

Summary

1. By manipulating the Maxwell's equations, we can derive the wave equation for the electric field, with wavespeed $v = c/n$.
2. By separating the time component, we can find a Helmholtz equation for the spatial part of the electric field
$$\big(\Delta + k^2\big)\mathbf{E}_{\text{sp}} = \mathbf{0}. \tag{7.14}$$
3. k in the Helmholtz equation is the wavenumber $k = 2\pi/\lambda$, the modulus of the wave vector $\mathbf{k}$. Sometimes we also define the wavenumber as $\bar{\nu} = 1/\lambda$.

§8. Plane Wave Solutions

Equivalence Between Light and Matter Waves

Before solving the Helmholtz equation, we shall demonstrate that the approach that we have taken is applicable to a wide range of waves, not only just light waves. One such example is matter waves, whose time-independent part propagates through space following the time-independent Schrödinger equation

$$\left(\frac{\mathbf{p}^2}{2m} + V\right)|\psi\rangle = E\,|\psi\rangle \quad \Rightarrow \quad \left[\Delta + \frac{2m(E-V)}{\hbar^2}\right]\psi = 0, \tag{8.1}$$

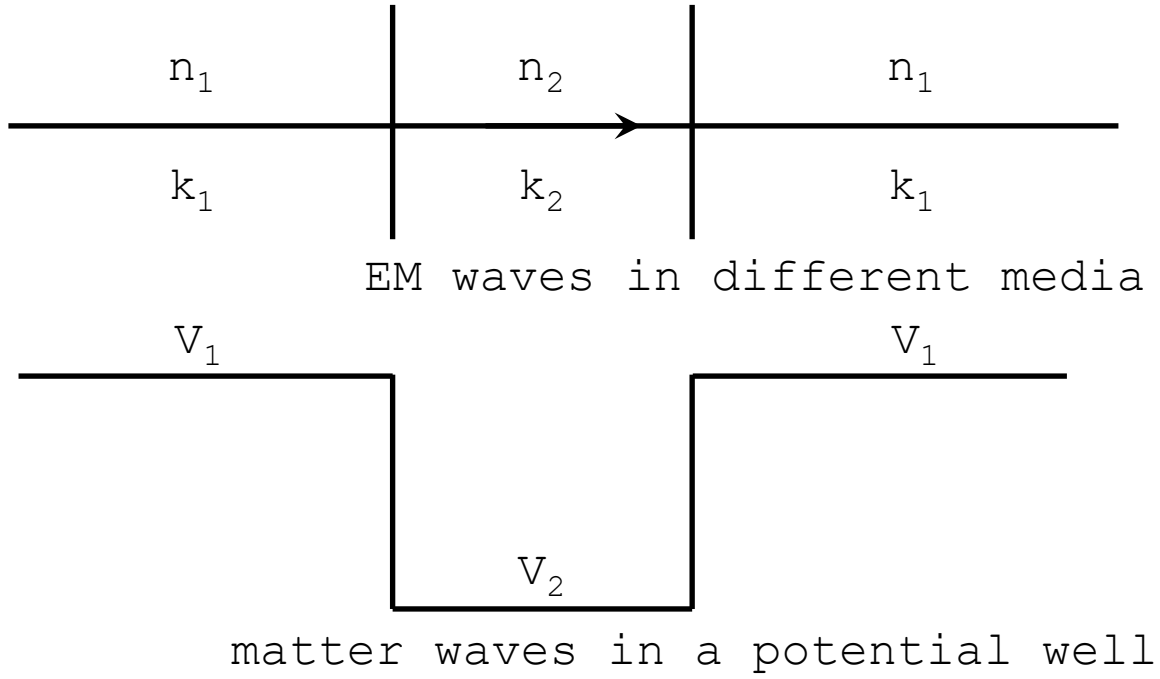

Figure 8.1: Correspondence between an electromagnetic wave travelling through media with different refractive indices and matter waves scattered through potential wells with different depths.

where the wavefunction $\psi \;=\; \langle \mathbf{x}|\psi\rangle$, and $\mathbf{p}$ is the momentum operator with

$$\langle \mathbf{x}|\, \mathbf{p}\, |\psi\rangle \;=\; -\mathrm{i}\hbar \,\mathrm{grad}\, \psi. \tag{8.2}$$

This is a Helmholtz equation

$$\big(\Delta + k^2\big)\psi \;=\; 0 \tag{8.3}$$

with wavenumber k, where

$$\hbar^2 k^2 \;=\; 2m(E-V). \tag{8.4}$$

Since matter waves have momentum $\hbar k$, we note that $\hbar^2 k^2/(2m) \;=\; (E-V)$ is therefore the kinetic energy of the particle. When one solves for solutions of matter waves, one often looks for the waves as scattering states of a potential well, which has a correspondence with electromagnetic waves in different media. One of the examples of such correspondence would be the setup illustrated in figure 8.1. The solutions for E_y, the electric field perpendicular to the direction of wave propagation in the three regions, follows analagous mathematical patterns as that of ψ, the wavefunction, since the boundary conditions of the equations at the boundaries of media and the potential well must be matched for both E_y and ψ.

Plane Waves

Now we are motivated that the Helmholtz equation is valid for not just electromagnetic radiation, we can formulate such an object called **scalar amplitude**, denoted by u (it looks like most literature tend to choose their own symbol for the scalar amplitude, where anther popular candidate of this choice is ψ or Ψ), and it stands for whatever object we plug into the Helmholtz equation. For example, it could be ψ, E_y, H_z, or other innovative objects. This means that the main equation to tackle would be

$$\big(\Delta + k^2\big)u \;=\; 0. \tag{8.5}$$

The simplest way to solve this would be in Cartesian coordinates, in which

$$\Delta \;=\; \partial_x^2 + \partial_y^2 + \partial_z^2, \tag{8.6}$$

which gives the **plane wave solution** as

$$u \;=\; u_0 \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}, \tag{8.7}$$

and $\mathbf{k}$ is exactly the wave vector we have talked about, with $|\mathbf{k}| \;=\; k$. However, we note that $\mathbf{k}$ can be pointing to any direction, and any of these forms a solution to the wave equation. Adding on the previously separated time dependence, we have

$$u \;=\; u_0 \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{r}-\omega t)}. \tag{8.8}$$

Note that, either when $\mathbf{r}$ changes at a fixed time (i. e. with t fixed), or when t changes at a fixed position (i. e. with $\mathbf{r}$ fixed), the value $\mathbf{k}\cdot\mathbf{r}-\omega t$ changes. We usually denote $\mathbf{k}\cdot\mathbf{r}-\omega t$ as the **phase** of the wave, and **wavefronts** as the surfaces for which light has a constant phase, i. e. the planes for which

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{k}\cdot\mathbf{r}-\omega t) \;=\; 0. \tag{8.9}$$

For special solutions with $\mathbf{k}$ along the z-direction, we have $k\mathrm{d}z/\mathrm{d}t \;=\; \omega$, i. e.

$$z \;=\; \frac{\omega}{k}t + \text{constant} \;=\; v_p t + \text{constant} \tag{8.10}$$

where $v_p \;=\; \omega/k$ is the **phase velocity** of the light. Since the expression of z does not have any spatial dependence, we note that wavefronts, in this case, would be planes perpendicular to the wave vector $\mathbf{k}$, which is always the case for linear materials.

Summary

1. Light waves and matter waves satisfy exactly the same Helmholtz equation, and therefore their solutions are mathematically analagous.
2. We therefore denote the object that satisfies the Helmholtz equation as u, the scalar amplitude, and solve for u. The solution of the Helmholtz equation in Cartesian coordinates are plane wave solutions, with wavefronts, the planes of constant phase, perpendicular to the wave vector $\mathbf{k}$.

§9. Intensity

Field Direction

We have already built a scalar theory of light, yet it is probably too early to call it a day. This is because $\mathbf{E}$ is a vector and u is a scalar, therefore for $\mathbf{E}$ to satisfy the Helmholtz equation, we have to solve for each of the three components of $\mathbf{E}$. However, are there any limitations upon the three electric field components? To resolve this, we shall send the plane wave solutions back into Maxwell's equations, which we have

$$\operatorname{div}\mathbf{D} \;=\; 0 \quad\Rightarrow\quad \mathrm{i}\mathbf{k}\cdot\mathbf{D} \;=\; 0 \tag{9.1}$$

$$\operatorname{div}\mathbf{B} \;=\; 0 \quad\Rightarrow\quad \mathrm{i}\mathbf{k}\cdot\mathbf{B} \;=\; 0 \tag{9.2}$$

$$\operatorname{curl}\mathbf{E} \;=\; -\partial_t\mathbf{B} \quad\Rightarrow\quad \mathrm{i}\mathbf{k}\wedge\mathbf{E} \;=\; \mathrm{i}\omega\mathbf{B}; \tag{9.3}$$

$$\operatorname{curl}\mathbf{H} \;=\; \partial_t\mathbf{D} \quad\Rightarrow\quad \mathrm{i}\mathbf{k}\wedge\mathbf{H} \;=\; -\mathrm{i}\omega\mathbf{D}. \tag{9.4}$$

Since in an isotropic material, we have $\mathbf{E}$ and $\mathbf{D}$ parallel and $\mathbf{H}$ and $\mathbf{B}$ parallel, from the above equations we can deduce that

$$\mathbf{E} \perp \mathbf{k}, \quad \mathbf{H} \perp \mathbf{k}, \quad \mathbf{E} \perp \mathbf{H}. \tag{9.5}$$

As a result, electromagnetic waves are transverse, for which we shall delay the consequences of that all the way till the polarisation part at the very end of this set of notes. Now we can keep this caveat in mind so we know what we are doing when we are investigating into the scalar theory of light — for all practical applications of scalar wave theory this additional property can be disregarded.

Energy Density

The energy density of an electromagnetic wave is

$$\rho_{em} = \frac{1}{2}(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}). \tag{9.6}$$

The electric components and the magnetic components are exactly the same in a vacuum, which means that in a vacuum we have

$$\rho_{em} = \mathbf{E}\cdot\mathbf{D}. \tag{9.7}$$

Previously we have suggested that the temporal part of $\mathbf{E}$ and $\mathbf{D}$ are fast oscillations shaped as $\mathrm{e}^{\mathrm{i}\omega t}$, and hence taking the real part, we have

$$\rho_{em} = \mathbf{E}_{\mathrm{sp}}\cdot\mathbf{D}_{\mathrm{sp}}\cos^2(\omega t). \tag{9.8}$$

Since we have the average of the fast time oscillation

$$\left\langle \cos^2(\omega t)\right\rangle_t = \frac{1}{2}, \tag{9.9}$$

we have the time average of the energy density of the electromagnetic wave as

$$\langle\rho_{em}\rangle_t = \frac{1}{2}\mathbf{E}_{\mathrm{sp}}\cdot\mathbf{D}_{\mathrm{sp}} = \frac{1}{2}\varepsilon_0{\mathbf{E}_{\mathrm{sp}}}^2. \tag{9.10}$$

From this we can work out the intensity of light.

Intensity and Irradiance

The intensity of light is defined as the energy per unit area per unit time. Therefore

$$I = \frac{\text{energy}}{\text{area}\times\text{time}} = \frac{\text{energy}}{\text{volume}}\times\frac{\text{distance}}{\text{time}} = \langle\rho_{em}\rangle_t\times\frac{c}{n} \tag{9.11}$$

where c/n is the velocity of the light. An alternative definition of the intensity of light using the **Poynting vector** is

$$\mathbf{S} = \mathbf{E}\wedge\mathbf{H}, \tag{9.12}$$

which is a vector parallel to the wave vector $\mathbf{k}$ in isotropic media. The intensity is then given by

$$I = \langle|\mathbf{S}|\rangle_t = \langle|\mathbf{E}\wedge\mathbf{H}|\rangle_t. \tag{9.13}$$

To show the equivalence of the two definitions of the intensity, we can consider this in a vacuum (i. e. $n = 1$), which gives

$$\langle|\mathbf{E}\wedge\mathbf{H}|\rangle_t = \frac{1}{2}|\mathbf{E}_{\mathrm{sp}}\wedge\mathbf{H}_{\mathrm{sp}}| = \frac{1}{2Z_0}{\mathbf{E}_{\mathrm{sp}}}^2 = \frac{1}{2}\sqrt{\frac{\varepsilon_0}{\mu_0}}{\mathbf{E}_{\mathrm{sp}}}^2 = \frac{1}{2}c\varepsilon_0{\mathbf{E}_{\mathrm{sp}}}^2 = \langle\rho_{em}\rangle_t\times c, \tag{9.14}$$

equivalent to the intensity defined using ρ_{em}.

Note that there is an analagous definition of the intensity called the irradiance. The difference between the two is that the intensity is the energy per unit time per unit area of the *light* and the irradiance is the energy per unit time per unit area of the *detector*. This means that, if the detector is placed at an oblique angle at the beam, then we have the irradiance of the light smaller than the intensity of the light, since the area of the detector must be larger than the cross-sectional area of the light.

When we have a wave represented by a scalar amplitude u, which is usually representative of the spatial part of the electric field strength E_{sp}, we have the intensity I proportional to u^2. If we are using the complex notation, then this u^2 will become u^*u, where u^* is the complex conjugate of u. Since practically one can only detect the relative intensity in different parts of space (as almost all optical equipments absorbs some of the light, it is practically impossible to track through the total amount of light through the optical system), it is common to omit the constant dimensional factors that make up I, and just write

$$I \;=\; u^*u, \tag{9.15}$$

which we will use in this course.

Summary

1. Light in an isotropic linear material is transverse, with $\mathbf{E} \perp \mathbf{k}$, $\mathbf{H} \perp \mathbf{k}$, and $\mathbf{E} \perp \mathbf{H}$.
2. The energy density of the light is given as

$$\rho_{\text{em}} \;=\; \frac{1}{2}(\mathbf{E}\cdot\mathbf{D} + \mathbf{B}\cdot\mathbf{H}). \tag{9.16}$$

   In a vacuum, the time averaged energy density is given as

$$\langle \rho_{em} \rangle_t \;=\; \frac{1}{2}\varepsilon_0 {\mathbf{E}_{\text{sp}}}^2. \tag{9.17}$$

3. The intensity is the energy per unit area per unit time of the light, with two equivalent expressions

$$I \;=\; \langle \rho_{em} \rangle_t \times \frac{c}{n} \;=\; \langle |\mathbf{S}| \rangle_t \,, \tag{9.18}$$

   where $\mathbf{S}$ is the Poynting vector

$$\mathbf{S} \;=\; \mathbf{E} \wedge \mathbf{H}. \tag{9.19}$$

   If we are using scalar amplitudes, then we have

$$I \;=\; u^*u. \tag{9.20}$$

§10. Spherical Wave Solutions

Spherical Wave Solutions to the Helmholtz Equation

We note that the Helmholtz equation

$$\left(\Delta + k^2\right)u \;=\; 0 \tag{10.1}$$

is linear. This means that although we have solved the equation by finding the eigenfunction of the Laplacian Δ in Cartesian coordinates, a superposition of these eigenfunctions must also be a solution to the Helmholtz equation, yet it is no longer an eigenfunction of the Laplacian under Cartesian coordinates. For a superposition of eigenfunctions, although we have wavefronts perpendicular to $\mathbf{k}$, we do allow $\mathbf{k}$ to have spatial dependence, and therefore the wavefronts are no longer planes.

Although all of these superposition solutions can be found by superposing plane wave solutions, some of them can be found using other methods. One particular method would be to find the eigenfunctions of the Laplacian Δ in other coordinate systems. As an example, spherical waves solutions naturally emerges as an eigenfunction of the Laplacian Δ in spherical polar coordinates. By symmetry, spherical waves should have an "source" where the wavefronts propagates from, and since the material that it travels in is assumed to be isotropic in the current discussion, it is reasonable to assume that the scalar wave u should not dependent on the angular variables, i. e. it should only be dependent on the radial coordinate r. This suggests that we can write the Laplacian as just the part associated with r, i. e.

$$\Delta \;=\; \frac{1}{r^2}\partial_r r^2 \partial_r. \tag{10.2}$$

Therefore the Helmholtz equation now reads

$$\left(\frac{1}{r^2}\partial_r r^2 \partial_r + k^2\right) u \;=\; 0. \tag{10.3}$$

The solutions then read, with the temporal part added back in,

$$u \;=\; \frac{u_0}{r} \mathrm{e}^{\mathrm{i}(\pm kr - \omega t)}, \tag{10.4}$$

corresponding to waves moving away and towards from the source corresponding to $+$ and $-$ signs on the exponent. The wave vector $\mathbf{k} \;=\; k\mathbf{e}_r$ (where $\mathbf{e}_r$ is the unit vector along the radial direction) is now along the radial direction, and is perpendicular to the spherical wavefronts.

We shall also note that equation 10.4 is also a solution to the Helmholtz equation in cylindrical polar coordinates, with

$$\Delta \;=\; \frac{1}{r}\partial_r^2 r. \tag{10.5}$$

However note that although mathematically they are the same form, physically the wavefronts are in fact different, as in spherical polar coordinates $r \;=\; \sqrt{x^2+y^2+z^2}$, but in cylindrical polar coordinates $r \;=\; \sqrt{x^2+y^2}$. This means that instead of spherical wavefronts, eigenfunctions of the Laplacian Δ in cylindrical coordinates have cylindrical wavefronts. Again, the wave vector $\mathbf{k} \;=\; k\mathbf{e}_r$ (note that now $\mathbf{e}_r$ is different to the previous $\mathbf{e}_r$ in spherical polar coordinates) is perpendicular to these cylindrical wavefronts.

Superposition and Interference

Previously we have suggested that the superposition of two solutions to the wave equation also forms a solution, and let us try to find some of these. Suppose that we have two sources located at A and B, each emitting dissipating spherical waves, then we can write

$$u_A \;=\; \frac{u_{0A}}{r} \mathrm{e}^{\mathrm{i}(kr-\omega t)}; \tag{10.6}$$

$$u_B \;=\; \frac{u_{0B}}{r} \mathrm{e}^{\mathrm{i}(kr-\omega t+\delta)}, \tag{10.7}$$

where δ accounts for the phase difference of the two emitters. Then, at a third point P that is at a distance a from point A and a distance b from point B, the scalar wave at P is given as

$$u_P = u_{aA}\mathrm{e}^{\mathrm{i}\delta_A} + u_{bB}\mathrm{e}^{\mathrm{i}\delta_B}, \tag{10.8}$$

where $u_{aA} = u_{0A}/a$, $u_{bB} = u_{0B}/b$, and the phases δ_A and δ_B accounts for both the phase caused by the source and the phase that is accumulated as the wave travels from points A and B to point P. The intensity is therefore

$$\begin{aligned} I &= u_{aA}{}^{*}u_{aA} + u_{bB}{}^{*}u_{bB} + 2\mathrm{Re}\left[u_{aA}u_{bB}{}^{*}\right]\cos\left(\delta_B - \delta_A\right) \\ &= I_A + I_B + 2\mathrm{Re}\left[u_{aA}u_{bB}{}^{*}\right]\cos\left(\delta_B - \delta_A\right), \end{aligned} \tag{10.9}$$

where $I_A = u_{aA}{}^{*}u_{aA}$ and $I_B = u_{bB}{}^{*}u_{bB}$ are the intensity of the sources located at A and B respectively, and the final term $2\mathrm{Re}\left[u_{aA}u_{bB}{}^{*}\right]\cos\left(\delta_b - \delta_a\right)$ that deviates from $I_A + I_B$ gives rise to the effect of **interference**, which is the main effect we shall be looking into in the next section of the course.

Summary

1. We can find solutions to the Helmholtz equation by finding eigenfunctions of the Laplacian Δ in other coordinate systems. When this is done in spherical polar coordinates, the solutions are
$$u = \frac{u_0}{r}\mathrm{e}^{\mathrm{i}(\pm kr - \omega t)}. \tag{10.10}$$
These solutions have spherical wavefronts. The solutions that arises by this method applied to cylindrical polar coordinates have the same mathematical form, but with a different definition of r, which gives rise to cylindrical wavefronts.

2. We can find solutions to the Helmholtz equation by superposing known eigenfunctions. The total intensity will in general deviate from the sum of the individual intensities, which is the effect of interference.

3 Theory of Interference

§11. Basic Two-Source Interference

Previously we have built the theory based on the scalar amplitude u from Maxwell's equations, now its time to unleash the full power of the theory that we have developed. In this section of the notes we shall investigate the interference effects associated with different geometric setups, and build a theory that analyses these effects systematically.

Arbitrariness of Overall Phase

We shall first note that the only detectable physical identity for any light ray would is intensity of the ray. Hence, if the scalar amplitude of light at one certain point is given as $u\mathrm{e}^{\mathrm{i}\delta}$, then the intensity of light is given as

$$I \;=\; u^*\mathrm{e}^{-\mathrm{i}\delta}u\mathrm{e}^{\mathrm{i}\delta} \;=\; u^*u, \tag{11.1}$$

where the overall phase vanishes. Therefore in optics it is common to ignore an overall phase of the scalar amplitude. However we shall take special notice that, if the scalar amplitude is a superposition of multiple contributions, each carrying a phase different to one another, then the relative phases between these contributions must *not* be ignored, as these exactly give rise to the interference effect.

Two-Slit Interference

We shall now move back to the problem of interference of two sources. We shall now align the two sources along the x-axis with a separation d with the principal axis running along the y-axis, which, if the two sources emit light with the same frequency and the same phase, would form the two-slit interference problem, which is exactly what we shall try to analyse. This is shown in figure 11.1. To analyse this problem, we consider light that exits the system at a fixed angle θ. We note that, from the geometrical setup, there is a path difference $d\sin\theta$ between the two sources, which gives rise to a phase difference $\delta \;=\; kd\sin\theta$ between the two slits, where k is the wavenumber. Therefore the exiting scalar amplitude is

$$u_{\text{out}} \;=\; u_0 + u_0\mathrm{e}^{\mathrm{i}\delta} \;=\; \text{overall phase} \times u_0\big(\mathrm{e}^{\mathrm{i}\delta/2} + \mathrm{e}^{-\mathrm{i}\delta/2}\big) \;=\; \text{overall phase} \times 2u_0\cos\left(\frac{\delta}{2}\right). \tag{11.2}$$

This gives the exiting intensity as

$$I_{\text{out}} \;=\; I_0\cos^2\left(\frac{\delta}{2}\right) \;=\; I_0\cos^2\left(\frac{kd\sin\theta}{2}\right) \;=\; I_0\cos^2\left(\frac{\pi d\sin\theta}{\lambda}\right), \tag{11.3}$$

where $I_0 \;=\; 4u_0^*u_0$, the maximum intensity. Note that when δ is an integer multiple of 2π (which in general corresponds to the optical path length between the two sources corresponds to an integer multiple of the wavelength *in vacuum* λ_0), we have maximum intensity output, which is a condition that we refer to as **constructive interference**. When δ is π away from an integer multiple of 2π (or that the optical path length between the two sources corresponds to $\lambda_0/2$ away from an integer multiple of the wavelength in vacuum λ_0), then we have minimum intensity output, which is a condition that we refer to as **destructive interference**.

Summation and Integration of Phasors

We may think of u as a vector on the complex plane, which allows us to add up complex numbers geometrically. The vectors formed by these complex numbers are called **phasors**. Since the overall phase is arbitrary, it is common to assign u_0 with no phase, leading to a vector

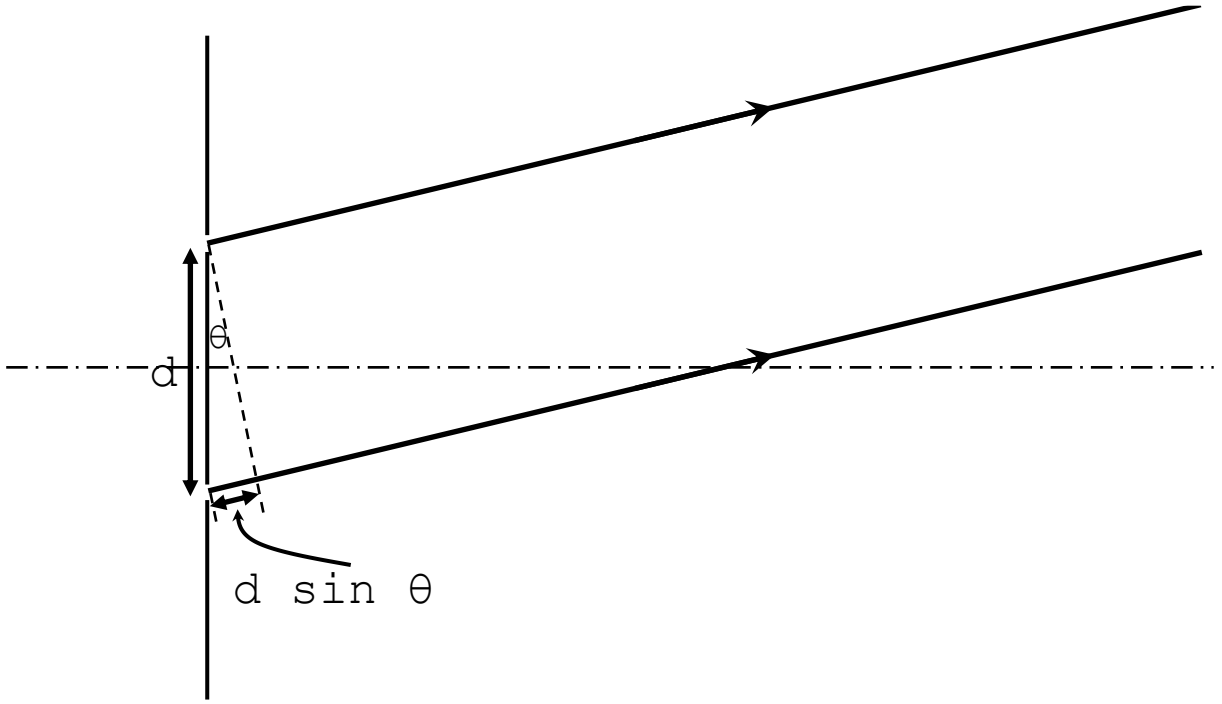

Figure 11.1: The two-slit interference problem.

along the real axis, with $u_0 e^{i\delta}$ as a vector with modulus u_0 but at an angle δ from the real axis. This is shown in figure 11.2. From the phasor diagram, simply by exploiting the geometry of the relation between interior and exterior angles of an isosceles triangle and trigonometric relationships, we can read off the overall phasor to have a magnitude $2|u_0|\cos^2(\delta/2)$, which immediately gives

$$I_{\text{out}} \;=\; \left[2|u_0|\cos^2\left(\frac{\delta}{2}\right)\right]^2 \;=\; I_0\cos^2\left(\frac{\delta}{2}\right). \tag{11.4}$$

We note that this approach is can be generalised to an interference problem between any finite number of light sources. However if we are faced with an extended source that is not divided into a finite number of slits, the scenario can be more tricky. Algebraically, this can be done by integrating the scalar amplitude. Geometrically, the method to use in that case would be to artificially sub-divide the plane into an infinite number of slits, each with an infinitesimally small slit size, and equipped with a different phase. Since the slit size is infinitesimal, the length of the phasors corresponding to each artificially sub-divided slit will also be infinitesimal. This means that an integration of scalar amplitudes is geometrically identical to drawing a continuous curve on the complex plane, where adding the contribution from each sub-divided slit is equivalent to extending the curve on the complex plane infinitesimally along the direction of the phase of light coming out of that the sub-divided slit. which we shall see examples of this later on.

Summary

1. The overall phase of a scalar amplitude u is arbitrary and undetectable, however the relative phases between the different contributions to the output amplitude is important.

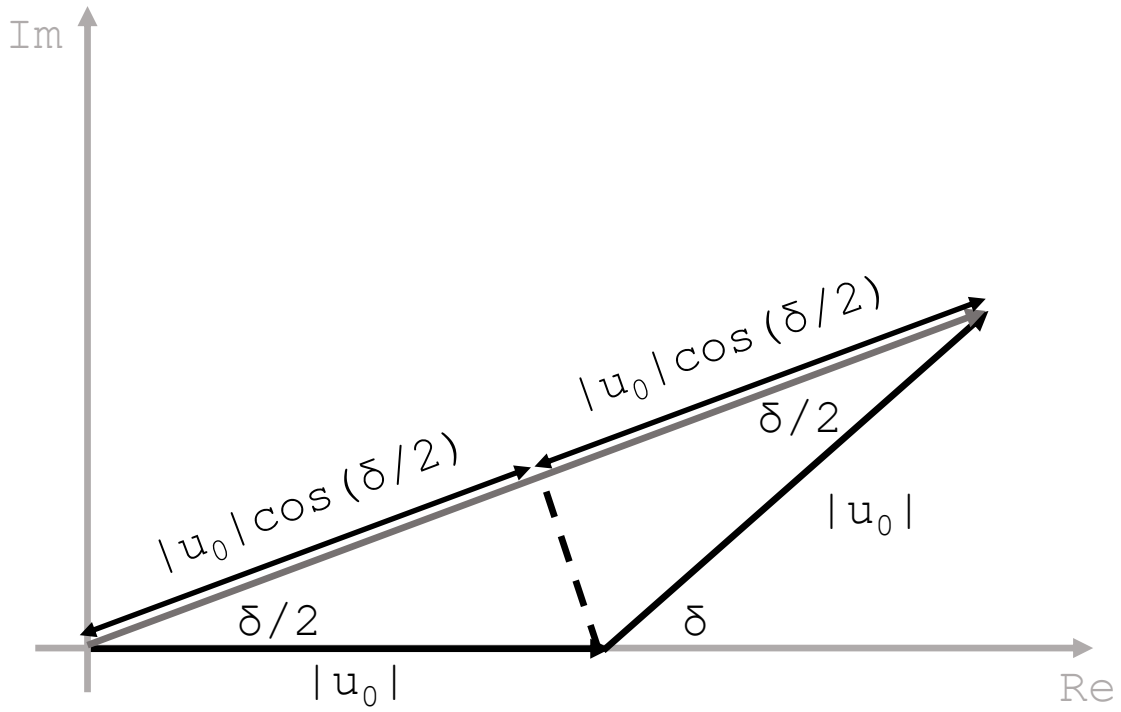

Figure 11.2: The phasor representation of the two-slit interference problem.

2. The scalar amplitude and intensity of two slits separated by a distance d is given as

$$u_{\text{out}} = 2u_0 \cos\left(\frac{\delta}{2}\right), \tag{11.5}$$

$$I_{\text{out}} = I_0 \cos^2\left(\frac{\delta}{2}\right), \tag{11.6}$$

where $I_0 = 4u_0^* u_0$, and $\delta = (kd\sin\theta)/2 = (\pi d\sin\theta)/\lambda$.

3. It is possible to represent scalar amplitudes as vectors called phasors on the complex plane, where summation of scalar amplitudes is geometrically identical to vector addition. Integration over scalar amplitudes is geometrically identical to drawing continuous curves on the complex plane.

§12. Fresnel-Kirchhoff Integral

Huygens' Principle

Now let us think about how light propagates, concentrating on the effects of superposition and interference. We shall first consider most general problem, then consider its special cases. The most general problem one can formulate is to find the scalar amplitude at a point **R** given that we know the scalar amplitude of light on a source plane S, which could be planar or curved, illustrated in figure 12.1. Doing this using Maxwell's equations requires us to consider superposing solutions that needs to be chosen very carefully, which is practically very difficult to do.

As a result, from here on we shall steer away from Maxwell's equation and choose another more convenient and more historical starting point, called **Huygens' Principle**. This approach is purely based on the analysing the system using physical intuition, which contains

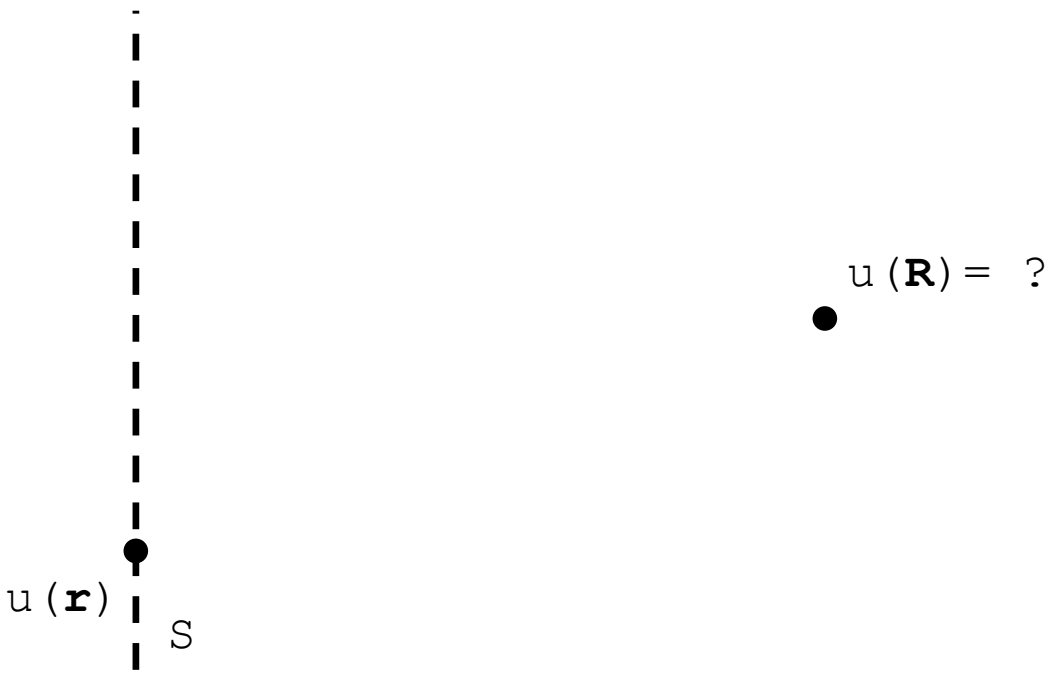

Figure 12.1: Finding the scalar amplitude at position $\mathbf{R}$ that originates as superposing waves that comes from a plane S.

statements that, in hindsight, does not match the elements of our current understanding using Maxwell's equations, however analysing it this way does give results that matches the experiment and the predictions of the more complete theory based on the analysis of Maxwell's equations. However, the mathematics used in this theory is somewhat simpler, and therefore it is still very useful to learn this method.

The statement of Huygens is that every point on a wavefront can be considered as an emitter of spherical waves — a theory that is clearly incorrect, as according to the theory of electromagnetism, waves travel themselves and do not need to be re-emitted. However for reasons stated above, let us stick with it. Let us assume that the point on the source plane S have position vectors $\mathbf{r}$, and is equipped with scalar amplitudes $u(\mathbf{r})$. Using Huygens' principle, let us integrate these spherical waves whose form are given by equation 10.4, which gives

$$u(\mathbf{R}) \;=\; \iint_S \frac{u(\mathbf{r})}{|\mathbf{R}-\mathbf{r}|} \times \mathrm{e}^{\mathrm{i}k|\mathbf{R}-\mathbf{r}|}\,\mathrm{d}a, \tag{12.1}$$

where $\mathrm{e}^{\mathrm{i}k|\mathbf{R}-\mathbf{r}|}$ is the phase accumulated when the spherical waves travels from $\mathbf{r}$ to $\mathbf{R}$.

Fresnel-Kirchhoff Integral

However, we shall note that our current theory does not take into account that waves only travels *forwards* and not *backwards*: if we consider the wave that moves from S to $\mathbf{R}$ as a plane wave which originates in the source plane S, then clearly the scalar amplitude of the point behind the source plane S should clearly be 0, as all light is travelling forwards and no light is hitting the point in behind. If we assume waves to be spherical, then this cannot be accounted for. To correct for that, we add an **obliquity factor** $\eta(\theta_i, \theta_o)$, where θ_i and θ_o are defined in figure 12.2, where we assume that the light is travelling from behind the source plane. This

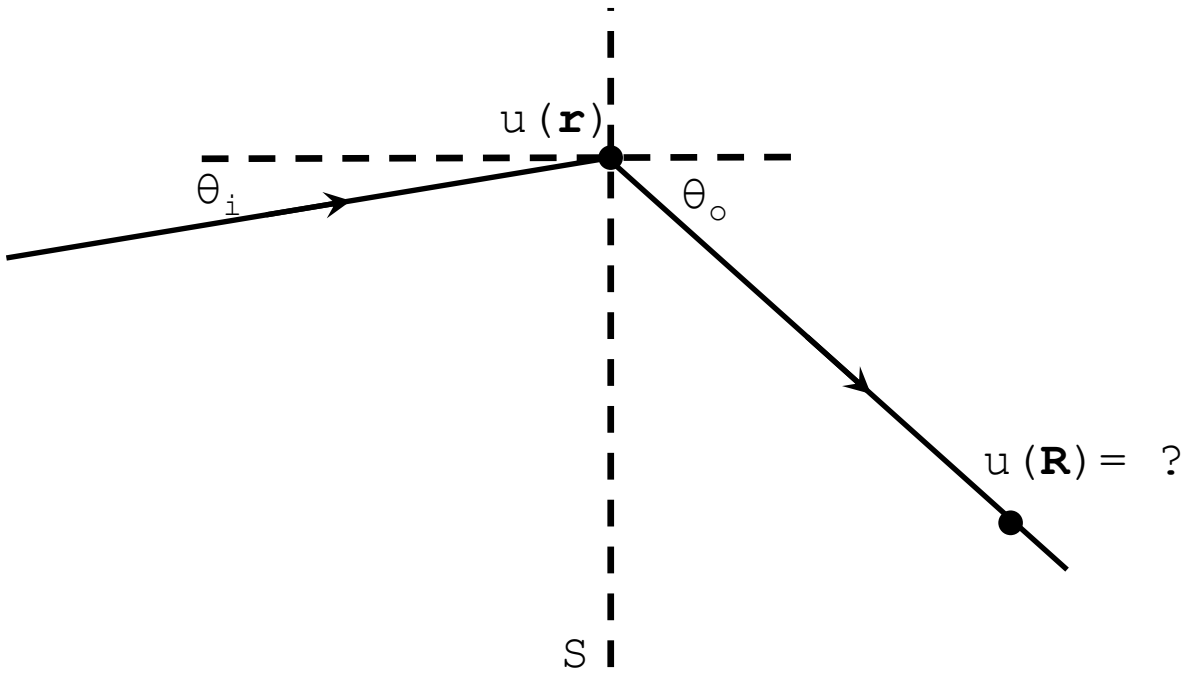

Figure 12.2: The definitions of θ_i and θ_o in the obliquity factor $\eta(\theta_i, \theta_o)$.

gives

$$u(\mathbf{R}) \;=\; \frac{1}{\mathrm{i}\lambda}\iint_S u(\mathbf{r}) \times \frac{\eta(\theta_i,\theta_o)}{|\mathbf{R}-\mathbf{r}|} \times \mathrm{e}^{\mathrm{i}k|\mathbf{R}-\mathbf{r}|}\,\mathrm{d}a, \tag{12.2}$$

the **Fresnel-Kirchhoff integral**. Note that it is conventional to add a $1/(\mathrm{i}\lambda)$ factor in front — since the detectable quantities are all relative, an addition of a constant factor will not change the physics. The addition of this factor is to accommodate for the fact that a collimated beam of light should not change its intensity as it is propagating through free space, which some justification will be given in §15.

A common choice for the obliquity factor is given as

$$\eta(\theta_i,\theta_o) \;=\; \frac{1}{2}(\cos\theta_i + \cos\theta_o). \tag{12.3}$$

Simply verifying the extreme cases, we have

- $\eta(0,0) \;=\; 1$: all light passes through if we are looking along the direction of light travel;
- $\eta(0,\pi) \;=\; 0$: there are no light through if we are looking back towards where light comes from,

which achieves the goal that we have originally set for the obliquity factor, and hence it is a good choice.

Summary

1. Huygens' principle states that we are able to view each point on the wavefront as an emitter for spherical waves.

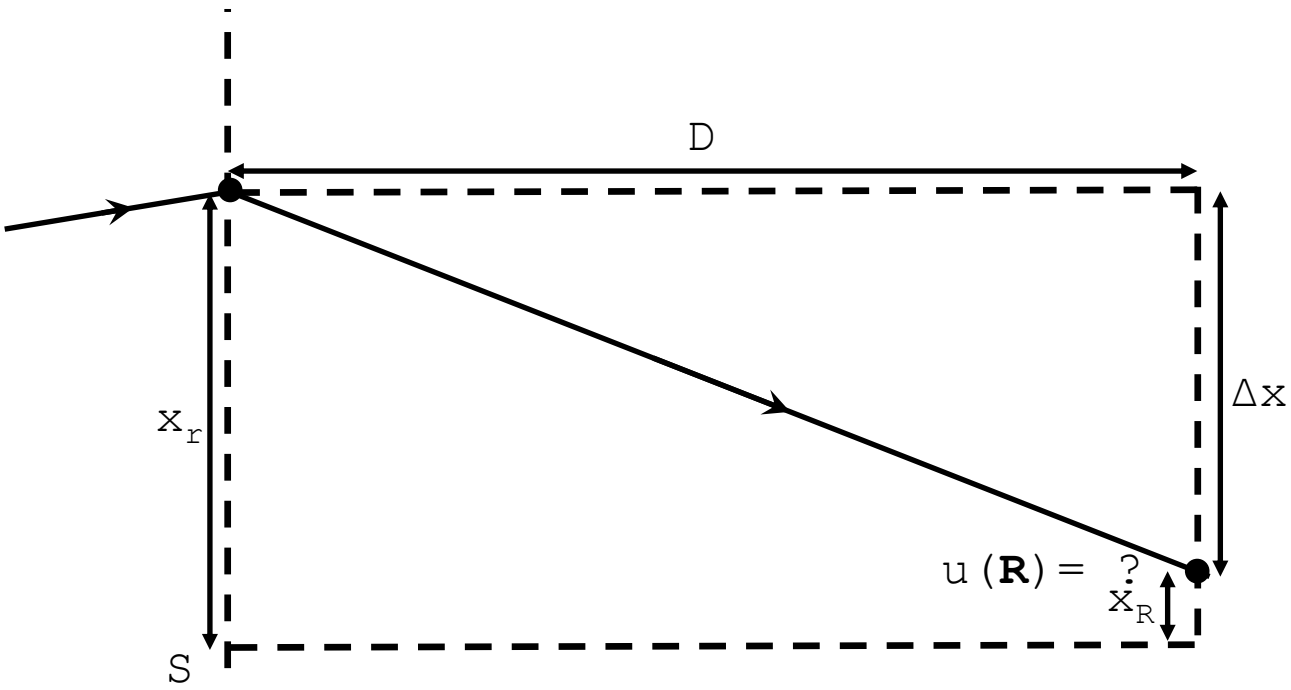

Figure 13.1: The optical setup where the image is formed at a very long distance away from the source plane.

2. The Fresnel-Kirchhoff integral reads

$$u(\mathbf{R}) \;=\; \frac{1}{\mathrm{i}\lambda}\iint_S u(\mathbf{r}) \times \frac{\eta(\theta_i,\theta_o)}{|\mathbf{R}-\mathbf{r}|} \times \mathrm{e}^{\mathrm{i}k|\mathbf{R}-\mathbf{r}|}\,\mathrm{d}a, \tag{12.4}$$

where the obliquity factor $\eta(\theta_i, \theta_o)$ is

$$\eta(\theta_i,\theta_o) \;=\; \frac{1}{2}(\cos\theta_i + \cos\theta_o). \tag{12.5}$$

§13. Fresnel Number

Far-Field Limit

We shall note that most of the optical setups that we will be looking into would have the image formed at a very long distance away from the source plane compared to the off-axis distance from the image. Such a setup is shown in figure 13.1, with $D \gg \Delta x$. Therefore we would like to modify the Fresnel-Kirchhoff integral, equation 12.2, so that it is suited for this situation. The main objects in equation 13.1 that we are able to simplify for this scenario are as follows.

- We assume that the light is mostly propagated through the principal axis, and therefore we take the obliquity factor to be 1.

- We exploit the relation $D \gg \Delta x$, so that we can write

$$|\mathbf{R}-\mathbf{r}| \;=\; \sqrt{D^2 + (\Delta x)^2} \;=\; D + \frac{(\Delta x)^2}{2D}. \tag{13.1}$$

We shall further note that, if we set a zero point for x, then the Fresnel-Kirchhoff integral is over x_r defined in figure 13.1 with x_R fixed. This means that utilising $\Delta x = x_R - x_r$, we can take equation 13.1 apart further, which then gives four terms in the expansion

$$|\mathbf{R} - \mathbf{r}| = D + \frac{x_R^2}{2D} - \frac{x_R x_r}{D} + \frac{x_r^2}{2D}. \tag{13.2}$$

When this is exponentiated to give a phase, the first two terms contributes to an overall phase since it is not dependent on x_r, and therefore can be ignored; the third term is linear on x_r, and the fourth term is quadratic on x_r. We shall then look at the consequences of this expansion.

Fresnel Number

If we only shine light through a very small area of S with a diameter a, then the integration limits of x_r are 0 and a, and therefore the maximum phase difference accumulated from the source plane to the image is

$$\Delta\delta = -\frac{kx_R a}{D} + \frac{ka^2}{2D} = -\frac{kx_R a}{D} + \pi F, \qquad F = \frac{a^2}{\lambda D}, \tag{13.3}$$

after removing the overall phase. F is called the **Fresnel number**, and depending on the value of F we can classify our problem into three different categories.

- If $F \ll 1$, then the phase shift with respect to x_r is linear on x_r, which we call the **Fraunhofer condition**, which field of optics is called the **far field regime**.
- If $F \approx 1$, then we are in the **near field regime**, which means that the quadratic term is able to change the phase by $\approx \pi$ i. e. shift the interference fringes from constructive to destructive.
- If $F \gg 1$, then the slit is simply too large for any interference effects to be observed — we have geometric optics, and what we see is just the shadow of the area.

We shall then move on to briefly look at the near field effects, and then move onto the far field effects, which we will spend a lot of time on, as they are applicable to practically any optical system in optics experiments.

Summary

1. We are usually interested in the far field, which we may assume by expanding the distance $|\mathbf{R} - \mathbf{r}|$ via a binomial expansion assuming $D \gg \Delta x$.
2. Whether we have near field or far field effects is determined by the Fresnel number $F = a^2/(\lambda D)$, where when $F \ll 1$ we have the far field regime, when $F \approx 1$ we have the near field regime, and when $F \gg 1$ we have geometric optics. The criterion for the far field regime is called the Fraunhofer condition, which suggests that if we are in the far field, then the phase accumulated is linear on the integration constant x_r.

§14. Talbot-Lau Effect

Talbot-Lau effect describes the near-field interference effect of a parallel bundle of light through a diffractive mask containing a large number of very thin slits, where these slits are equally spaced with spacing d, which is an optical device called a **transmission grating**. This is

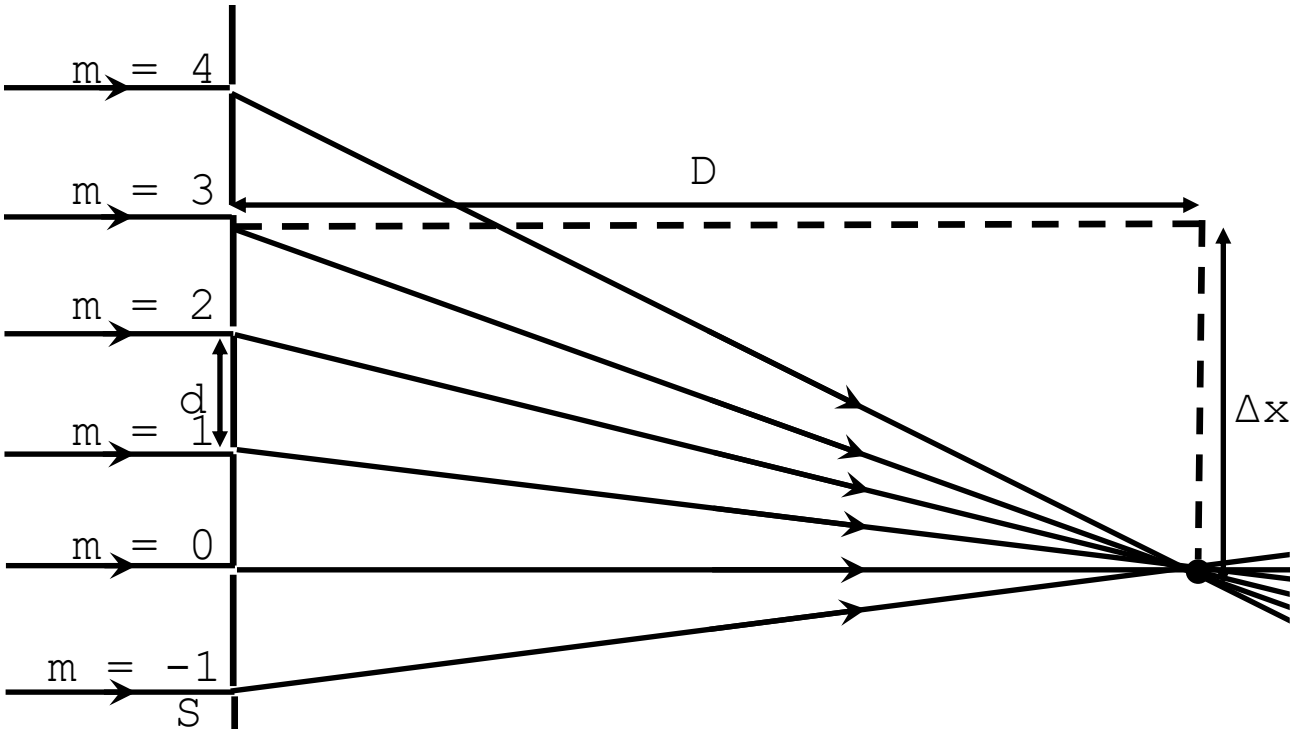

Figure 14.1: The setup that investigates the near-field effects of interference through a transmission grating, or the Talbot-Lau effect.

illustrated in figure 14.1, where we set the distance from the grating to the point in question along the principal axis as D. We then set the point we are interested in to point directly at a slit which we refer to as the 0^{th} slit, and set the transverse distance of that point to any other slits as Δx therefore the m^{th} slit will have $\Delta x = md$. Then, according to equation 13.1, the phase of the wave caused by the m^{th} slit from slit 0 is given by

$$\delta_m = \frac{k(\Delta x)^2}{2D} = \frac{\pi(\Delta x)^2}{\lambda D} = \frac{\pi m^2 d^2}{\lambda D}, \tag{14.1}$$

neglecting the overall phase. Therefore, the phase difference of the $(m+1)^{\text{th}}$ slit and the m^{th} slit is given by, using equation 14.1,

$$\delta_{m+1} - \delta_m = \frac{\pi d^2}{\lambda D}(2m+1). \tag{14.2}$$

Now let us select the point in question to be in the middle of two slits, such that the distance Δx to the m^{th} slit is $\Delta x = \left(m + \frac{1}{2}\right)d$, which then leads to

$$\delta_m = \frac{\pi\left(m+\frac{1}{2}\right)^2 d^2}{\lambda D} \quad \Rightarrow \quad \delta_{m+1} - \delta_m = \frac{\pi d^2}{\lambda D}(2m+2). \tag{14.3}$$

We note that interesting interference effects happens if we select D to be comparable to $d^2/(2\lambda)$ corresponding to the Fresnel number $F \approx 1$: if $D = d^2/(2\lambda)$, then we have the difference in phases to be $(2m+1) \times 2\pi$ and $(2m+2) \times 2\pi$, which, since m is an integer, means that we have light between consecutive slits in both of the two cases interfering constructively — therefore we are effectively *halving the grating spacing*. If we then select $D = z_{\text{T}} = 2d^2/\lambda$, we have the image identical to the grating, which is known as **self-imaging**. Selecting more

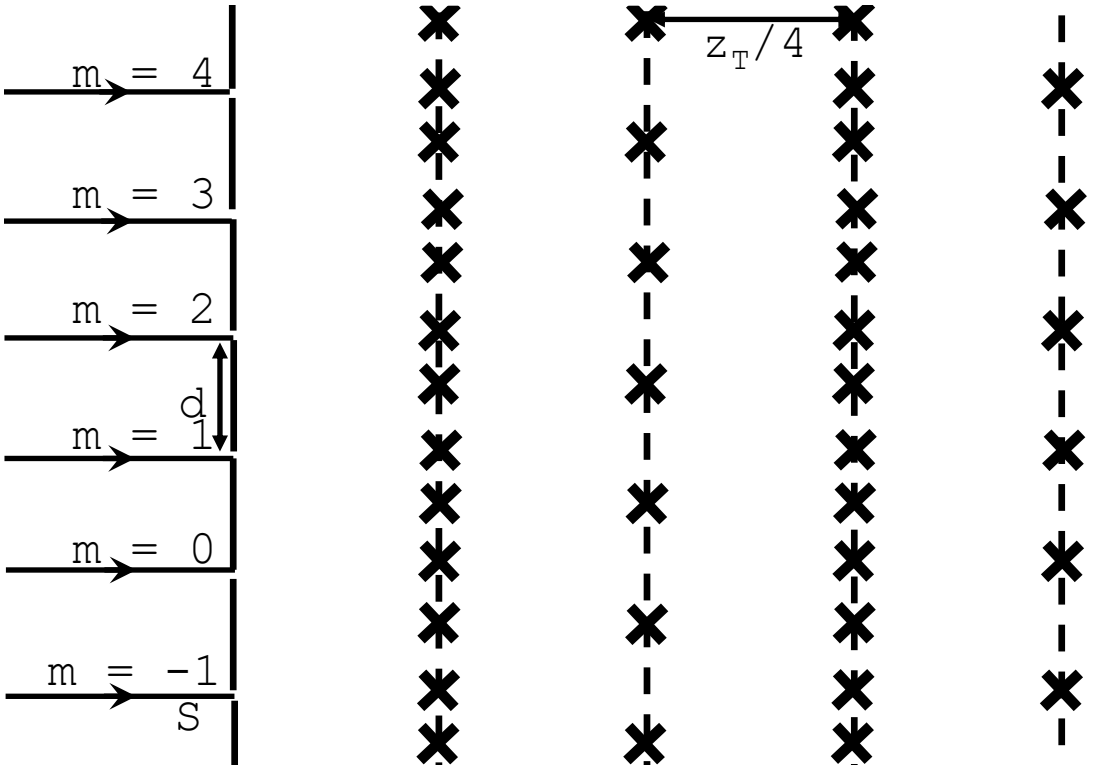

Figure 14.2: Interference of a parallel bundle of rays through a transmission grating at different values of D. The crosses represents points where light interferes constructively.

different values of D leads to more interesting effects, which, a selection of them is shown in figure 14.2.

Summary

1. The Talbot-Lau effect describes the interference between light sources that emerges out of a transmission grating in the near field. At a distance $z_{\mathrm{T}} = 2d^2/\lambda$ away from the grating we have the effect of self-imaging, where the image is exactly identical as the grating. At a distance $x_{\mathrm{T}}/4$ the image is the grating itself, but with the slit separation halved.

§15. Wave Propagation

Propagation of Light Rays Through Space

Now we shall give a brief account of how Huygens' principle and the Fresnel-Kirchhoff integral describes the rays travelling through space. Let us consider a plane wave hitting a surface S which then propagates to the point P, which we place on the principal axis, shown by figure 15.1. The off-axis rays travel further than the on-axis rays by an extra distance

$$\sqrt{z^2+\rho^2} - z \;=\; \rho^2/(2z). \tag{15.1}$$

Therefore, to find the maxima and minima, we equate the phase shift to an integer multiple p of π, and hence we write

$$\frac{\rho_p{}^2\pi}{z\lambda} \;=\; p\pi \quad\Rightarrow\quad \rho_p \;=\; \sqrt{p\lambda z}, \tag{15.2}$$

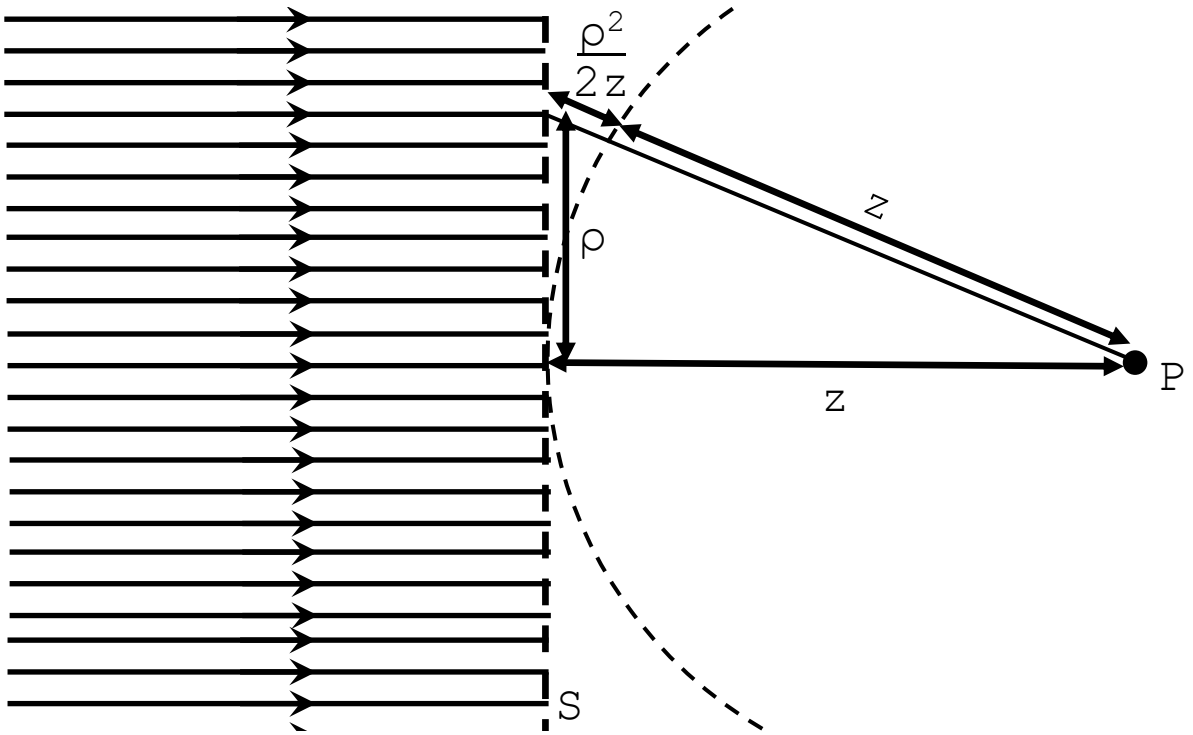

Figure 15.1: A plane wave travelling towards a point P.

which will be handy later. Substituting equation 15.1 this into the Fresnel-Kirchhoff integral, equation 12.2, with obliquity factor unity then gives

$$u_P(z) \;=\; \int_0^\infty \frac{1}{\mathrm{i}} \times \frac{u_0}{\sqrt{z^2+\rho^2}} \mathrm{e}^{\mathrm{i}k\rho^2/(2z)} \times \frac{2\pi}{\lambda}\rho \,\mathrm{d}\rho. \tag{15.3}$$

The best method of doing this integral is to break the integration into an infinite number of annuli, where each annulus has boundaries ρ_p and ρ_{p+1}, and assume that in the section of the integration where $\rho_{p-1} < \rho < \rho_p$ we are allowed to approximate $\rho = \rho_p = p\lambda z$. Then, the integral with a lower bound ρ_p gives

$$u_{P,p}(z) \;=\; -\frac{u_0 z}{\sqrt{z^2+p\lambda z}} \Big[\mathrm{e}^{\mathrm{i}k\rho^2/(2z)}\Big]_{\rho_{p-1}}^{\rho_p}. \tag{15.4}$$

Note that since the square bracket contains the complex exponential where the phase is an integer multiple of π, the exponential itself is ± 1, and therefore

$$u_{P,p}(z) \;=\; \frac{u_0 z}{\sqrt{z^2+p\lambda z}} \times 2 \times (-1)^{p+1}, \tag{15.5}$$

where the first term denotes a decreasing amplitude of the wave as p increases, and the final part of the equation represents an alternating phase for the corresponding annulus. Therefore,

$$u_P(z) \;=\; \sum_{p\,=\,1}^{\infty} u_{P,p}(z) \;=\; u_0, \tag{15.6}$$

and therefore the light propagates straight through with the intensity unchanged, agreeing with the plane wave solution of the Helmholtz equation.

Phasor Representation of Light Propagation

We shall note that now since we are summing over the annuli from the centre of the light to the edge, each annulus will give a phasor that is pointing along the imaginary axis, but as we are going off from the centre to the edge, the length of the phasor will decrease and the direction of the phasor will alternate. Or, we can think about each annulus as a resultant of phasors that originates from integrating continuously over the annuli with an ever-increasing phase, and therefore forms semi-circles on the complex plane. An illustration of this is shown in figure 15.2. Adding all these phasors together eventually gives us the scalar amplitude of the input wave.

We shall note that, by giving all the annuli which leads to the phasor pointing *towards* the origin a π phase change by, for example, directing the wavefront into a piece of glass with the annuli that requires a phase change travelling through a thicker layer of glass, we may significantly extend the length of the resultant phasor, as there will no longer be any phasors that reduces the length of the resultant phasor. This apparatus is called the **Fresnel lens**, which focuses all the light onto the principal axis. However the advantage of this lens is focussing, not imaging — the aberrations of the Fresnel lens is much larger than a convex lens for off-axis rays. As a result, such a device is usually used in lighthouses, such that light is perfectly focused onto a sharp beam. A more advanced technique is to introduce curved surfaces on the glass plate, which leads to straightening-out of the semi-spheres in figure 15.2, giving an even larger resultant scalar amplitude.

Poisson v Fresnel, 1818

Now let us consider a propagating plane wave and mask out the centre of the wave. This is equivalent to removing the outermost circle in the phasor diagram in figure 15.2, and still gives rise to a non-negligible overall scalar amplitude. As a result, the Fresnel-Kirchhoff theory predicts a bright spot can be observed *right behind* the mask in the near-field, although intuitively no light should arrive there at all. Poisson was supportive of the theory of light as particles, and he used this argument to show how ridiculous the wave theory of light, proposed by Fresnel, was. However, Arago verified the existence of the spot experimentally which was supportive of the wave theory of light, even though this phenomena was anti-intuitive. Nevertheless, the spot is now called the **Poisson spot**.

Summary

1. We artificially sub-divide the wavefront into rings, with the boundary of the rings matching the surfaces of constructive and destructive interference. Summing over these rings, we find that the scalar amplitude is invariant as the wave propagates.

2. We can introduce a π phase shift to alternate rings, which then leads to a larger scalar amplitude of the output, which is the underlying principle of a **Fresnel Lens**.

3. If we mask out the centre of the wavefront, then this leads to a bright spot in the near-field just behind the spot, called the **Poisson spot**.

§16. Fraunhofer Integral

We now look at the far-field regime, which we shall recall, is the case where the Fresnel number $F = a^2/(\lambda D)$ to be far less than unity, where D is the distance from the diffracting mask

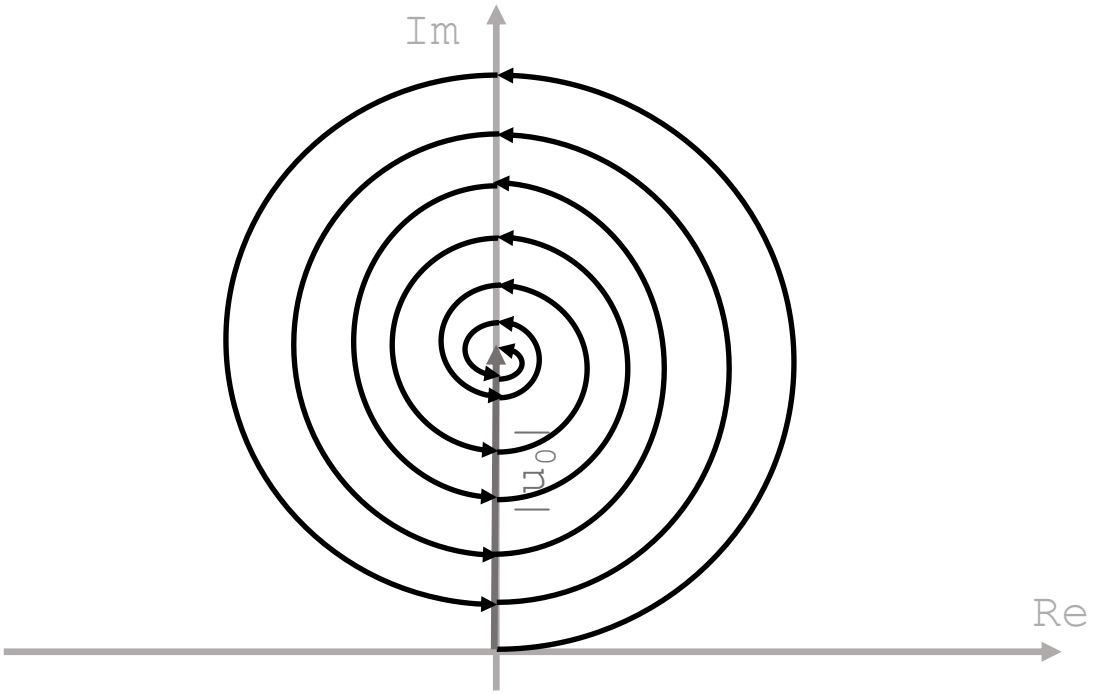

Figure 15.2: The phasor diagram of wave propagation.

to the screen, and a is be the size of the aperture. Recall that this limitation on the Fresnel number is equivalent to the condition that the phase shift on the screen to be a linear function of the distance travelled in the plane of the diffracting mask. These limitations are given the name **Fraunhofer condition** or **Fraunhofer (far-field) limit**. To study this condition, we shall investigate how the Fresnel-Kirchhoff integral, equation 12.2, simplifies. Recall that the Fresnel-Kirchhoff integral reads

$$u(\mathbf{R}) \;=\; \frac{1}{\mathrm{i}\lambda}\iint_S u(\mathbf{r}) \times \frac{\eta(\theta_i,\theta_o)}{|\mathbf{R}-\mathbf{r}|} \times \mathrm{e}^{\mathrm{i}k|\mathbf{R}-\mathbf{r}|}\,\mathrm{d}a. \tag{16.1}$$

The first modification is the fact that light travels mostly along the principal axis, and therefore we set the obliquity factor η to unity. Then, let us set the coordinates on the plane of the diffraction mask as $\mathbf{r} \;=\; (x,y,0)^{\mathrm{T}}$ and the coordinates on the plane of the screen as $\mathbf{R} \;=\; (X,Y,Z)^{\mathrm{T}}$. Then, the Fresnel-Kirchhoff integral reads, neglecting the $1/(\mathrm{i}\lambda)$ factor at the front,

$$u(\mathbf{R}) \;=\; \frac{1}{D}\int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\, u_0(x,y)\mathrm{e}^{\mathrm{i}k\sqrt{(X-x)^2+(Y-y)^2+Z^2}}. \tag{16.2}$$

Next we expand the square root, and the Fresnel number being small means that we only need to expand to first order for both x and y. This gives

$$u(\mathbf{R}) \;=\; \int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\, u_0(x,y)\mathrm{e}^{\mathrm{i}k\left(\frac{X}{|\mathbf{R}|}x+\frac{Y}{|\mathbf{R}|}y\right)}, \tag{16.3}$$

dropping an overall magnitude and phase $\mathrm{e}^{\mathrm{i}k|\mathbf{R}|}/D$. We also note that, strictly speaking, the exponent of the above equation should be negative if we follow the mathematics, but since we are always only detecting u^*u, the sign of this phase is not detectable anyway, so we can select this freely. Next, if we express X and Y using angles θ and φ, such that

$$X \;=\; |\mathbf{R}|\sin\theta, \quad Y \;=\; |\mathbf{R}|\sin\varphi, \tag{16.4}$$

then we end up with

$$\mathscr{U}(\theta,\varphi) \;=\; \mathscr{U}_x(\theta)\mathscr{U}_y(\varphi) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x\, u_x(x)\mathrm{e}^{\mathrm{i}xk\sin\theta} \times \int_{-\infty}^{\infty} \mathrm{d}y\, u_y(y)\mathrm{e}^{\mathrm{i}yk\sin\varphi}, \tag{16.5}$$

given that it is able to separate $u_0(x,y)$ into a product $u_x(x)\times u_y(y)$. It is very common to label the **spatial frequencies**

$$\beta_x \;=\; k\sin\theta, \quad \beta_y \;=\; k\sin\varphi, \tag{16.6}$$

which transforms the integral to

$$U(\beta_x,\beta_y) \;=\; U_x(\beta_x)\times U_y(\beta_y) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x\, u_x(x)\mathrm{e}^{\mathrm{i}\beta_x x} \times \int_{-\infty}^{\infty} \mathrm{d}y\, u_y(y)\mathrm{e}^{\mathrm{i}\beta_y y}, \tag{16.7}$$

the **Fraunhofer diffraction integral**. This is effectively a Fourier transform

$$U(\beta_x,\beta_y) \;=\; \mathscr{T}_F^{(x,y)}[u(x,y)](\beta_x,\beta_y) \;=\; \mathscr{T}_F^{(x)}[u(x)](\beta_x)\mathscr{T}_F^{(y)}[u(y)](\beta_y), \tag{16.8}$$

where

$$\mathscr{T}_F^{(\xi)}[f(\xi)](\beta_\xi) \;=\; \int_{-\infty}^{\infty} \mathrm{d}\xi\, f(\xi)\mathrm{e}^{\mathrm{i}\beta_\xi \xi}. \tag{16.9}$$

The Fourier transform is mathematically very easy to do, and we shall discuss the methods to exploit this Mathematical convenience in §20.

Summary

1. Modifying the Fresnel-Kirchhoff integral in the far-field (Fraunhofer) limit gives the Fraunhofer diffraction integral

$$U(\beta_x,\beta_y) \;=\; U_x(\beta_x)U_y(\beta_y) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x\, u_x(x)\mathrm{e}^{\mathrm{i}\beta_x x} \times \int_{-\infty}^{\infty} \mathrm{d}y\, u_y(y)\mathrm{e}^{\mathrm{i}\beta_y y}, \tag{16.10}$$

   which is a Fourier transform

$$U(\beta_x,\beta_y) \;=\; \mathscr{T}_F^{(x,y)}[u(x,y)](\beta_x,\beta_y) \;=\; \mathscr{T}_F^{(x)}[u(x)](\beta_x) \times \mathscr{T}_F^{(y)}[u(y)](\beta_y), \tag{16.11}$$

§17. Rectangular Aperture

$u(x,y)$ for the Rectangular Aperture

We shall now look into some concrete examples of diffraction in the far-field. One of such examples is the diffraction through a rectangular aperture with dimensions $a\times b$. To do this, we shall first write down $u(x,y)$, which we shall formulate as

$$u(x,y) \;=\; u_0\,\mathrm{tophat}_a\,(x)\times \mathrm{tophat}_b\,(y), \tag{17.1}$$

where we define

$$\mathrm{tophat}_d\,(\xi) \;=\; \begin{cases} 1, & \xi \;\in\; (-d/2, d/2) \\ 0, & \text{otherwise.} \end{cases} \tag{17.2}$$

It is rather clear that $u_0\,\mathrm{tophat}_a\,(x)$ describes a slit with width a along the x-direction with a constant scalar amplitude u_0 across the slit, and as a result $u_0\,\mathrm{tophat}_a\,(x)\times\mathrm{tophat}_b\,(y)$ describes a rectangular slit with dimensions $a\times b$, which justifies equation 17.1.

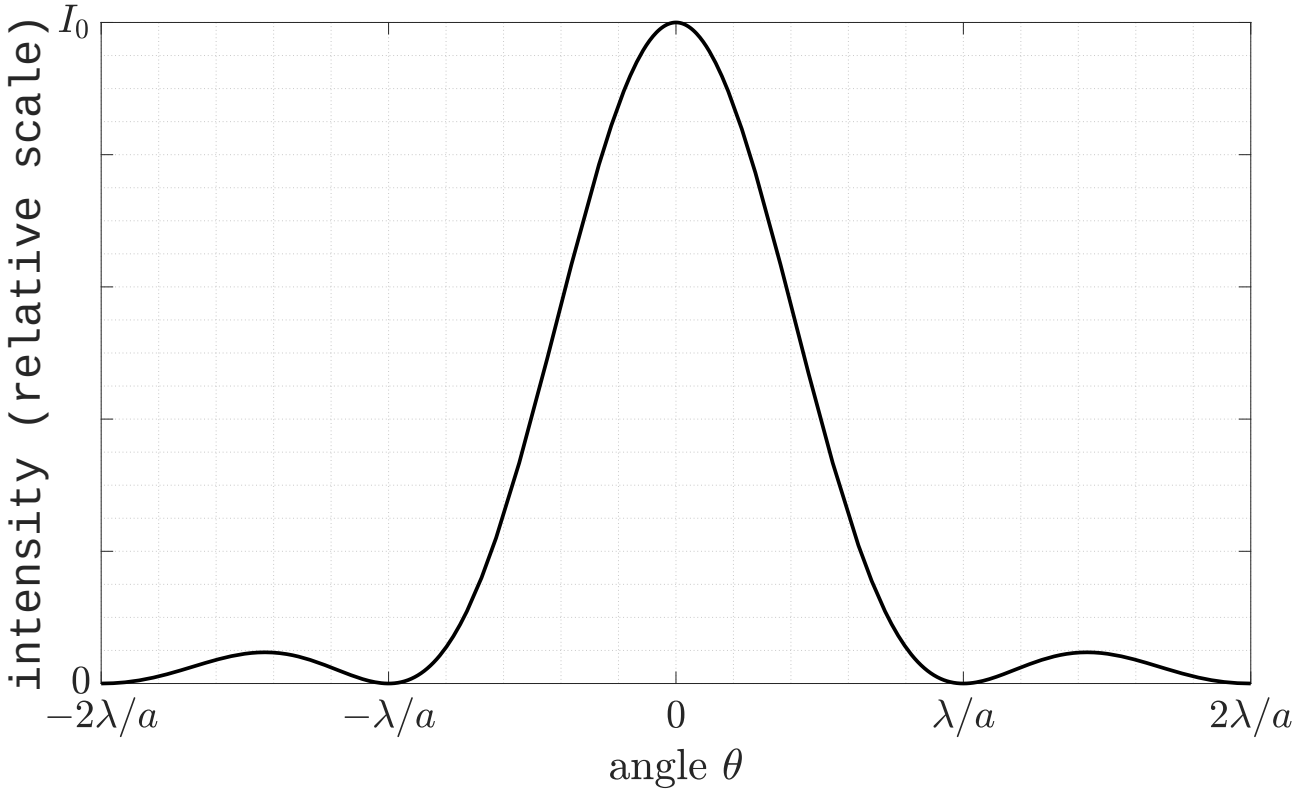

Figure 17.1: The intensity of a rectangular slit at $y = 0$, under the limit $\sin\theta \approx \theta$.

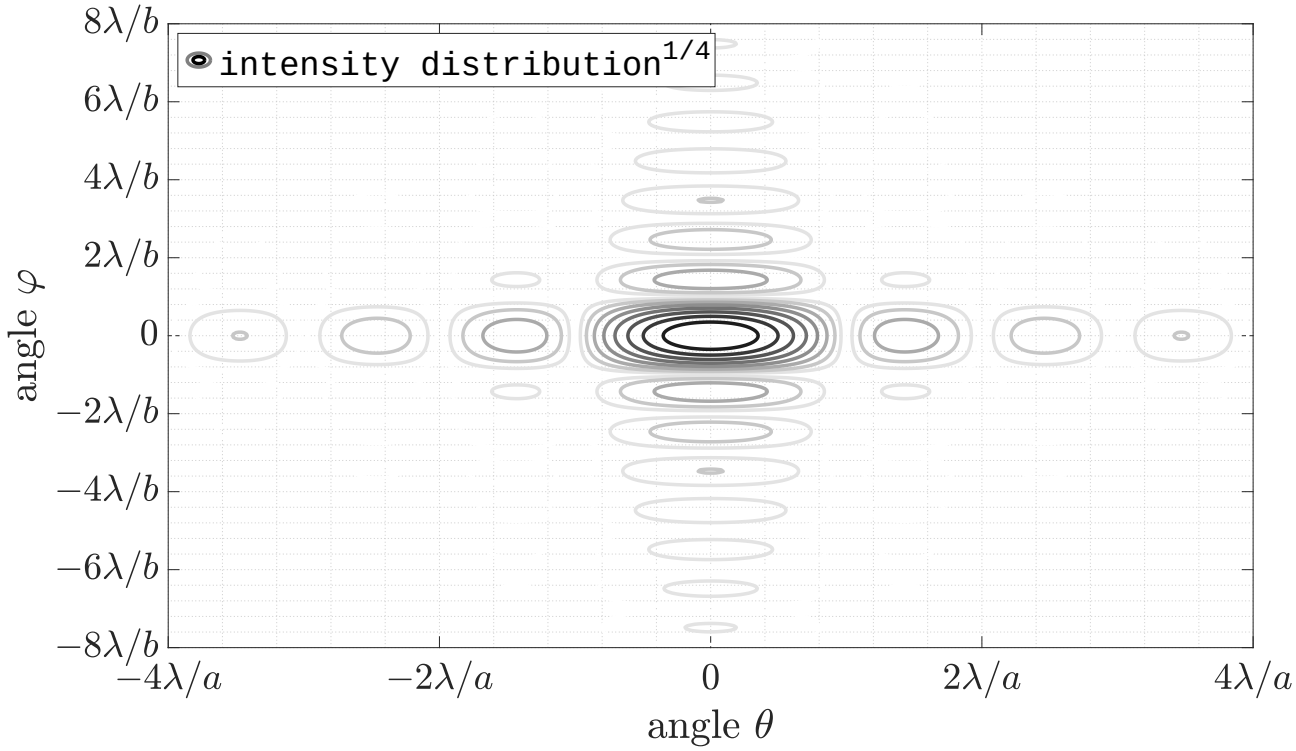

Figure 17.2: The contour plot of the 4^{th} root of the intensity of a rectangular slit, in the limit that $\sin\theta = \theta$ and $\sin\varphi = \varphi$. We take the 4^{th} root for better contrast of the contour plot.

Carrying out the Fraunhofer Integral

Let us now carry out the Fraunhofer integral for the tophat function. This is given by

$$U(\beta_\xi) \;=\; \int_{-\infty}^{\infty} \mathrm{d}\xi \;\mathrm{tophat}_d\,(\xi)\mathrm{e}^{\mathrm{i}\beta_\xi\xi} \;=\; \int_{-d/2}^{d/2} \mathrm{d}\xi\,\mathrm{e}^{\mathrm{i}\beta_\xi\xi} \;=\; d\,\mathrm{sinc}\left(\frac{\beta_\xi d}{2}\right), \tag{17.3}$$

where the sinc function is defined by

$$\mathrm{sinc}\,\zeta \;=\; \begin{cases} 1, & \zeta \;=\; 0 \\ \sin\zeta/\zeta, & \text{otherwise.} \end{cases} \tag{17.4}$$

Using the above result, the Fraunhofer integral of the rectangular aperture is

$$U(\beta_x,\beta_y) \;=\; u_0 ab\,\mathrm{sinc}\left(\frac{\beta_x a}{2}\right)\mathrm{sinc}\left(\frac{\beta_y b}{2}\right). \tag{17.5}$$

and therefore the intensity distribution with respect to the spatial frequencies is given as

$$\mathrm{I}(\beta_x,\beta_y) \;=\; U^*U \;=\; I_0\,\mathrm{sinc}^2\left(\frac{a}{2}\beta_x\right)\mathrm{sinc}^2\left(\frac{b}{2}\beta_y\right), \tag{17.6}$$

where I_0 is the maximum intensity of the distribution. If we just look along the x-direction with $y \;=\; 0$, then we note that the intensity distribution is

$$\mathrm{I}_x(\beta_x) \;=\; I_0\,\mathrm{sinc}^2\left(\frac{a}{2}\beta_x\right) \;=\; I_0\,\mathrm{sinc}^2\left(\frac{\delta_{\max}}{2}\right), \tag{17.7}$$

where $\delta_{\max}$ is the the maximum phase difference between a selection of any of the two rays, which is the phase difference between light at $x \;=\; \pm a/2$, i. e.

$$\delta_{\max} \;=\; \beta_x x \;=\; ka\sin\theta. \tag{17.8}$$

Therefore, the location of minima is given by

$$2\pi p \;=\; ka\sin\theta_p \quad\Rightarrow\quad \sin\theta_p \;=\; 2\pi p/(ka) \;=\; p\lambda/a, \tag{17.9}$$

where p is a non-zero integer. Note that this is exactly the same as just equating the optical path length difference the ray at the edge of the slit and the central ray to an integer number of wavelengths

$$\Delta\mathrm{OPL} \;=\; a\sin\theta_p \;=\; p\lambda \quad\Rightarrow\quad \sin\theta_p \;=\; p\lambda/a, \tag{17.10}$$

i. e. the condition that we are able to partition the slit into an even number of sub-divisions, where each sub-division on the top half of the slit being able to find a sub-division on the bottom half of the slit that interferes destructively with the corresponding sub-division on the top (the argument for the location of minima of the single slit in any standard A-level course).

The intensity distribution of the single slit is shown in figure 17.1 and the intensity distribution of the rectangular slit is shown in figure 17.2. There are two remarks for the shape of the intensity distribution.

- The position of the minimum is closer up if the slit width is wider — this is representative of a more general principle in diffraction: the closer-up the feature is on the diffraction pattern, the larger the feature is on the diffraction mask. Therefore, if the θ- and φ-axes in figure 17.2 are scaled equally, then we can say that the slit has $a \;<\; b$.

- The **width** of the diffraction pattern is usually defined as the full-width at half-maximum, which, in this case, is approximately half the minimum-to-minimum width of the central peak, or the maximum-to-minimum width for the central maximum. Therefore, the width of the peak in θ for the slice through $y = 0$ is given as

$$\Delta\theta \approx \arcsin(\lambda/a). \tag{17.11}$$

We shall take an extra note that usually optical apparatuses are circular — for example lenses for telescopes and microscopes, which means that we need to take into account of extra corrections caused by this. We shall take a look of this in the next section. After that, we shall look at the phasor representation of the single slit diffraction problem.

Summary

1. The rectangular slit can be described as

$$u(x,y) = u_0 \,\mathrm{tophat}_a(x) \times \mathrm{tophat}_b(y). \tag{17.12}$$

2. After carrying out the Fraunhofer integral, we have the intensity pattern as

$$\mathrm{I}(\beta_x,\beta_y) = I_0 \,\mathrm{sinc}^2\left(\frac{a}{2}\beta_x\right)\mathrm{sinc}^2\left(\frac{b}{2}\beta_y\right), \tag{17.13}$$

with minima located at $\sin\theta_p = p\lambda/a$ and width $\Delta\theta \approx \arcsin(\lambda/a)$.

§18. Circular Aperture

Intensity Distribution of a Circular Aperture

The derivation of the intensity distribution of the circular aperture is not required for this course, although the principle is the same as the rectangular aperture i. e. working out the Fraunhofer integral. The only difference is that the integral can only be expressed with special functions, which is an extra complication. Since this derivation contains no physical significance, we shall quote the result without proof.

The intensity pattern for a circular aperture with radius $\mathscr{R}$ is

$$I(\theta) = I_0 \times \left[\frac{2J_1(k\mathscr{R}\sin\theta)}{k\mathscr{R}\sin\theta}\right]^2 = I_0 \times \left[\frac{2J_1(\mathscr{R}\beta_r)}{\mathscr{R}\beta_r}\right]^2, \qquad \beta_r = k\sin\theta, \tag{18.1}$$

where, again, we are taking a slice through any radial axis. Here $J_1(\xi)$ is a Bessel function of the first kind $J_\nu(\xi)$ with $\nu = 1$. Since Bessel functions of the first kind are oscillatory and decaying for large ξs, $[J_1(\mathscr{R}\beta_r)/(\mathscr{R}\beta_r)]^2$ and $\mathrm{sinc}^2(a\beta_x)$ have roughly the same shape. Recall that the Bessel function of the first kind with $\nu = 1$ have zeros

$$z_{1k} = 3.832,\ 7.016,\ 10.173,\ \cdots, \tag{18.2}$$

which means that the first minimum is located at

$$k\mathscr{R}\sin\theta = \frac{2\pi}{\lambda}\times\frac{a}{2}\times\sin\theta = 3.832 \quad\Rightarrow\quad \theta = \arcsin\left(1.220\frac{\lambda}{a}\right), \tag{18.3}$$

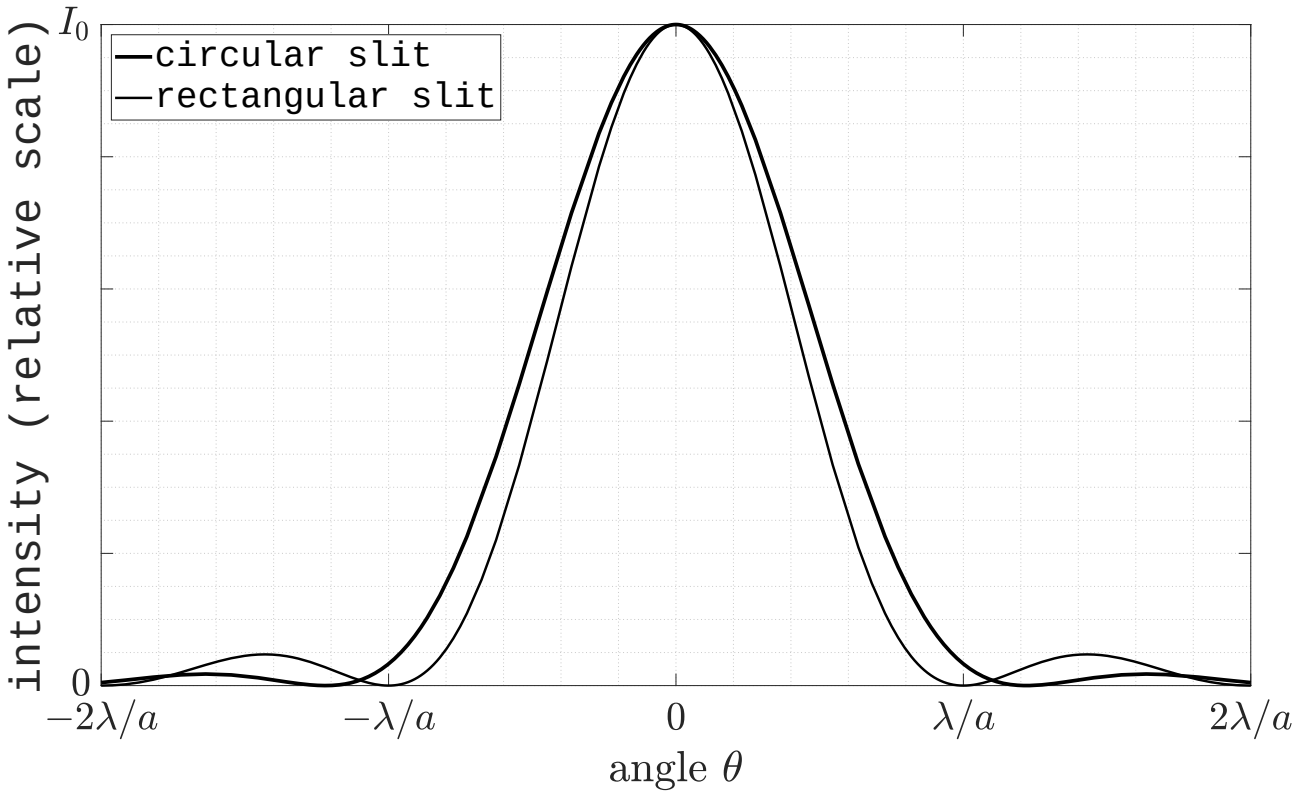

Figure 18.1: Comparison between the rectangular and circular slit diffraction patterns, under the limit $\sin\theta \approx \theta$.

where a is the diameter of the slit. Again the full-width at half-maximum is approximately the distance from the origin to the first minimum, and therefore the angular diameter is approximately

$$\Delta\theta \approx \arcsin(1.220\lambda/a), \tag{18.4}$$

very similar for the rectangular slit but with an addition of a constant. The two slit patterns are illustrated in figure 18.1.

Phasor Representation of a Single Slit

We now think about the shape of the phasors on the complex plane given by the single slit — let us say that the slit that we are looking at is just one-dimensional i. e. a rectangular slit with $b \gg 1$, although the statement is generalisable to circular apertures. Note that again we are integrating over infinitesimal sub-divisions, and as a result the phasor representation on the complex plane will still be a continuous curve, somewhat like the case where the phasors add-up in the near field, demonstrated in figure 15.2. However the main difference is that the complex amplitudes across the whole slit is constant, unlike the case in equation 15.4, where the amplitude has a decreasing amplitude with respect to p, which is representative of the transverse distance across the wavefront. This means that instead of circles with an ever-decreasing radii, each "loop" of the phasors have exactly *the same radius*.

As a result, if we view the slit at a small value of θ, then the phase change across the slit is small, therefore giving a large radius of the phasors, like the outermost loop in figure 18.2, and hence a relatively large resultant scalar amplitude and therefore a relatively large intensity. When θ increases further, the radius becomes smaller as the phase change per unit length across

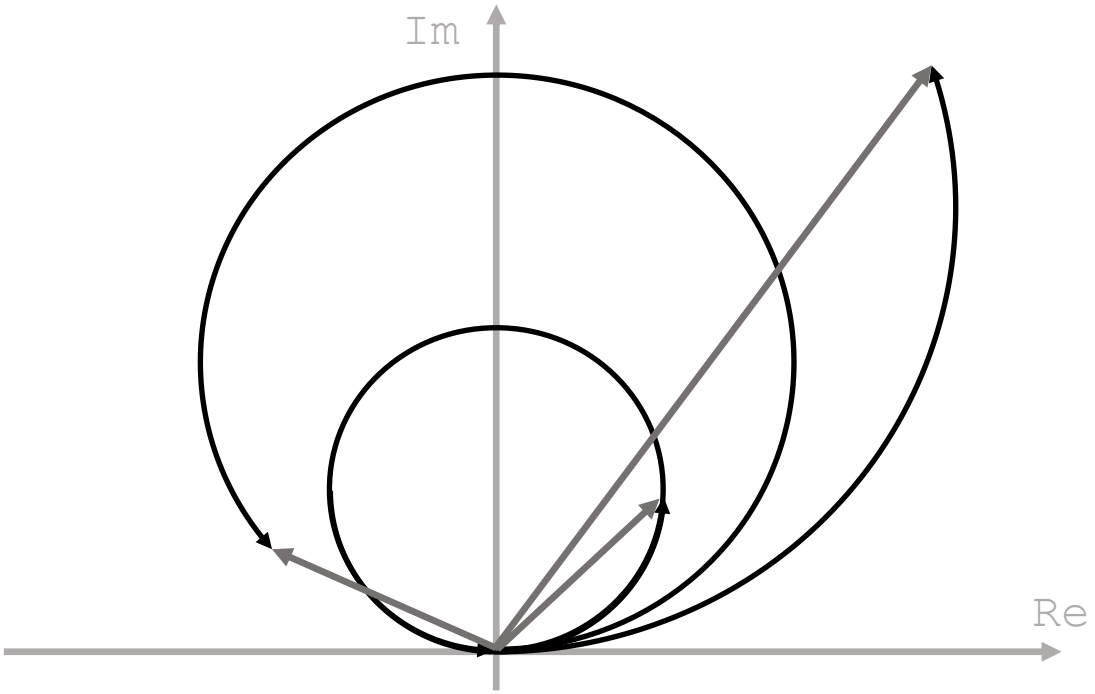

Figure 18.2: The phasor representation of the single slit.

the slit becomes larger. Then the loops start to wind-back to one-another, giving rise to smaller and smaller resultant scalar amplitudes, as θ keeps increasing to very large angles, which is why the higher maxima are far less bright than the central maxima. This can be demonstrated by the smallest loop in figure 18.2.

The phasor representation also illustrates the appearance of the central maximum and the minima. The central maximum means that we are viewing the slit at an angle $\theta = 0$, therefore all the phasors line up along the real axis, giving the maximum scalar amplitude achievable; and the minima means that the θ we are viewing at is exactly when the phasors form one exact circle and therefore has no resultant scalar amplitude.

Summary

1. The intensity distribution of the circular aperture along the radial axis has a similar shape with the intensity pattern of a rectangular slit, but with a constant addition such that the full-width at half-maximum reads

$$\Delta\theta = \arcsin\left(1.220\lambda/a\right). \tag{18.5}$$

2. The single slit diffraction pattern can be illustrated using the phasor method, which traves out circles on the complex plane. This demonstrates the appearances of the central maximum, the reduced amplitudes of the maxima with higher order, and the appearances of minima.

§19. Distinguishability of Two Sources

Fraunhofer, but not Far Field

Since Fraunhofer diffraction is mathematically simpler than the Fresnel-Kirchhoff diffraction,

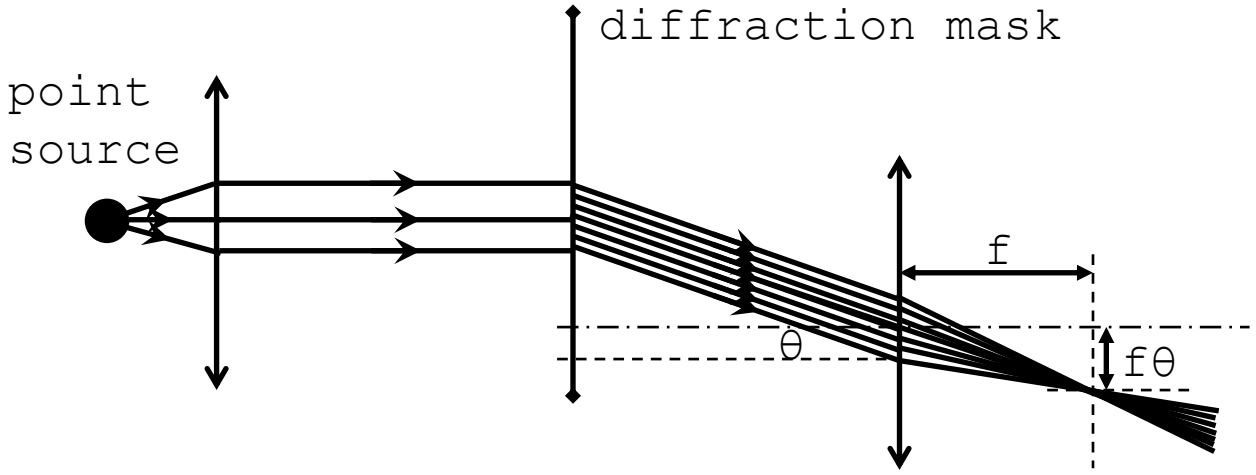

Figure 19.1: The addition of the thin lens after the diffraction mask, to achieve the Fraunhofer condition without going to the far field. Note that this is a simplification when θ is small. When we are looking at large values of θ, then we usually would like to mount the receiver unit on top of a turntable, which then allows us to determine θ more easily, and without causing an extra phase shift after passing the lens.

we often exploit this convenience in designing physics experiments. However we cannot afford to place everything in the far field: when we perform experiments in a laboratory, the size of the apparatus is limited by the size of the laboratory; if each experiment requires using apparatuses that is over 20 m long in length for relatively accurate measurements, then we will have far less experimental setups in a laboratory than we would ideally want. As a result, we would like to find a way to achieve the Fraunhofer condition but without being in the far field.

The method to go for would be to use a **thin lens**. After the light has gone out of the diffraction mask, we equip the exiting light with a thin lens, then the light can be projected onto a screen that is relatively closed-up, and now we have transformed the angle θ, which is representative of the relative phase, into the distance on the screen, $Y = f\theta$. However, the phase difference is still linearly dependent on the transverse distance on the slit, and therefore we have achieved the Fraunhofer diffraction limit, without going to the far-field. This is demonstrated in figure 19.1.

Distinguishability of Two Sources

So far we have looked at passing one coherent collimated beam through the a rectangular slit. However we can then think about projecting two incoherent light sources through the same slit, for example, imaging two stars. In this case, we need to take into account effects caused by both geometric and wave optics as follows.

- Geometric optics gives rise to the fact that after passing through the slit, the two stars are focussed on different parts of the detector.

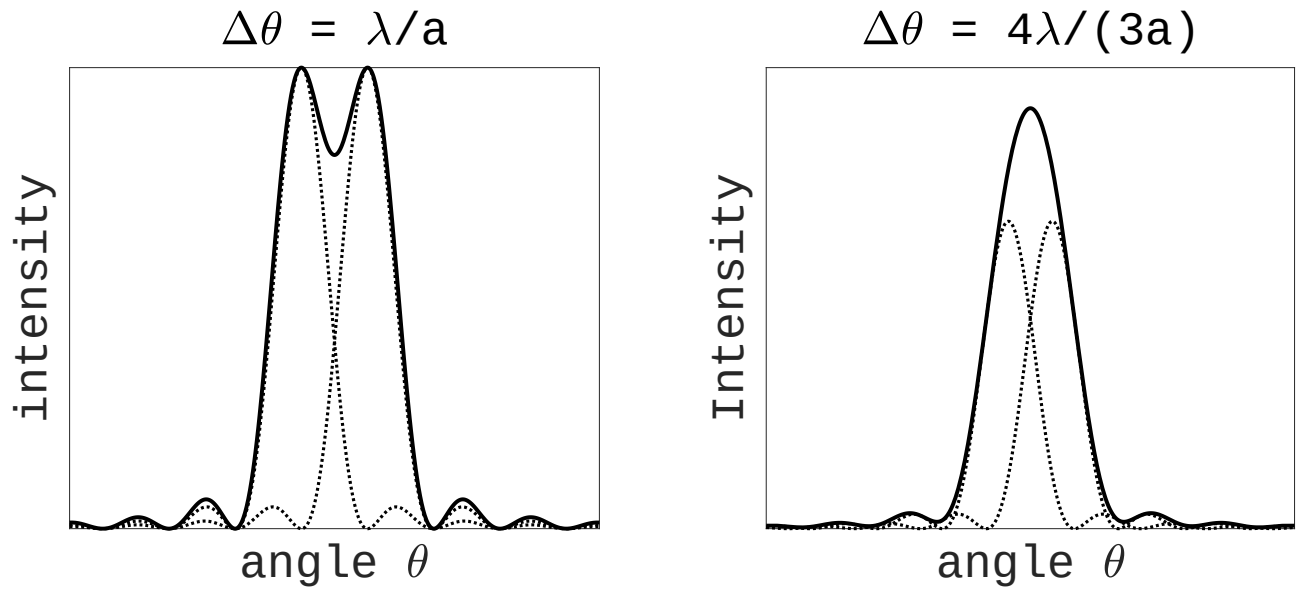

Figure 19.2: A demonstration of whether the two sources are distinguished. The dotted lines represent the intensity distribution of the light coming from an individual source diffracted through a slit with the matching width and position of the central maximum.

- Wave optics gives rise to the fact that, each star through the slit, even if the stars are point sources, forms patches of light on the detector of a finite angular size $\Delta\theta = 1.220\lambda/a$ (or, $\Delta\theta = \lambda/a$ for a rectangular slit with size $a \times a$) where we assume $\sin\theta = \theta$, and a is the diameter of the aperture.

This suggests that, if the geometrical angular separation of the two stars is smaller than the diffracted angular separation of the two stars, then the two stars cannot be distinguished i. e. we have no idea whether the large patch of light we are looking at is one star or two different stars. A demonstration of this effect for the rectangular aperture case is shown in figure 19.2. The left hand diagram is what we suggest as the limit of the two sources being distinguishable. However theoretically the two can be moved just a little bit closer and we can still tell them apart, hence strictly speaking the limit of distinguishability should lie at where the combined intensity does not have a local minima in the middle of the two diffraction patterns at all, which is a condition that requires more mathematical analysis. However, the more simplified condition $\Delta\theta = 1.220\lambda/a$ or λ/a is practically good enough and is called the **Rayleigh criterion**, with the stricter and more mathematically demanding condition called the **Sparrow criterion**. When these limits are exceeded, we cannot distinguish the two stars apart, which gives rise to an overall interference pattern similar to the right hand diagram of figure 19.2.

Summary

1. To achieve the Fraunhofer condition without going to the far-field, we put a thin lens before the detector unit. This way we achieve the desired linear phase shift and limit the apparatus size in the laboratory at the same time.

2. To consider two sources satisfying $\Delta\theta > 1.220\lambda/a$ for a circular aperture or $\Delta\theta > \lambda/a$ for a rectangular aperture as distinguishable sources, where $\Delta\theta$ is the angular separation of the two sources by the laws of geometric optics, and a is the size of the aperture.

§20. Fraunhofer Integral as a Fourier Transform

Fourier Transforms in Optics

We have previously suggested that the Fraunhofer integral is just a Fourier transform, i. e.

$$U(\beta) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x\, u(x) \mathrm{e}^{\mathrm{i}\beta x} \;=\; \mathscr{T}_F[u(x)](\beta). \tag{20.1}$$

Note that this is different from the Fourier transform used in quantum mechanics for relating the wavefunctions in the position and momentum representations

$$\phi(k) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\mathrm{i}kx} \psi(x) \;=\; \mathcal{F}[\psi(x)](k), \qquad k \;=\; \frac{p}{\hbar} \tag{20.2}$$

by a factor of $1/\sqrt{2\pi}$ and a sign difference on the phase. Since they are not quite the same, let us go through the similarities and differences in the properties associated with these differences.

- The Fourier transform and its inverse must share a factor of $1/(2\pi)$ for normalisation; since the Fourier transform in optics does not have the $1/\sqrt{2\pi}$ factor, the whole $1/(2\pi)$ factor must be dumped on the inverse Fourier transform. Furthermore, the sign on the phase of the inverse Fourier transform must be opposite to that of the Fourier transform. These effects combined means that the inverse Fourier Transforms in the two conventions read

$$u(x) \;=\; \mathscr{T}_F^{-1}[U(\beta)](x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\beta\, U(\beta) \mathrm{e}^{-\mathrm{i}\beta x}, \tag{20.3}$$

$$\psi(x) \;=\; \tilde{\mathcal{F}}[\phi(k)](x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, \phi(k) \mathrm{e}^{\mathrm{i}kx}. \tag{20.4}$$

- The Fourier transform under both conventions are linear.

- The relation between the Fourier transform, the translation operator T_a (shifts the function towards the *right* by a distance a), and the modulation operator E_b (multiplies the function by a phase $\mathrm{e}^{\mathrm{i}kb}$ or $\mathrm{e}^{\mathrm{i}\beta b}$) is the same, i. e.

$$\mathscr{T}_F T_a \;=\; E_{-a} \mathscr{T}_F, \qquad \mathscr{T}_F E_b \;=\; T_b \mathscr{T}_F. \tag{20.5}$$

  The first relationship in quantum mechanics means that the state in the momentum space is multiplied by an overall phase if the whole quantum system is translated. In optics this means that the scalar amplitude on the screen will experience a phase shift if the diffraction mask is translated, i. e. there is no observable change to the scalar amplitude on the screen whatsoever. The second relationship in quantum mechanics means that if the position space wavefunction is modulated with a phase, then in the momentum space, this is the same as an overall shift in momentum for the quantum system. For example, a quantum particle with wavefunction

$$\psi(x) \;=\; \frac{1}{\pi^{1/4}\sqrt{d}} \exp\!\left(\mathrm{i}k_0 x - \frac{x^2}{2d^2}\right) \tag{20.6}$$

  can be thought as a Gaussian wavepacket (which has momentum expectation value 0) modulated by k_0, and therefore the expectation value of the momentum of the particle is $\langle p \rangle \;=\; \hbar k_0$. Sometimes we use filters to modulate light rays and this will in turn cause a translation on the image plane, completely in analogy with the case in quantum mechanics.

- Convolution is defined analogously

$$u_1(x) \otimes u_2(x) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x'\, u_1(x') u_2(x - x'), \tag{20.7}$$

$$(\psi_1 \star \psi_2)(x) \;=\; \int_{-\infty}^{\infty} \mathrm{d}y\, \psi_1(y) \psi_2(x - y), \tag{20.8}$$

  however it is important to note that there is a difference in the convolution theorem — the convolution theorem in optics has the $\sqrt{2\pi}$ factor stripped off:

$$\mathscr{T}_F[u_1 \otimes u_2] \;=\; \mathscr{T}_F[u_1] \times \mathscr{T}_F[u_2], \qquad \mathcal{F}[\psi_1 \star \psi_2] \;=\; \sqrt{2\pi} \times \mathcal{F}[\psi_1] \times \mathcal{F}[\psi_2]. \tag{20.9}$$

  We shall also note that the convolution theorem, under the optics convention, can "work backwards":

$$\mathscr{T}_F[u_1 \times u_2] \;=\; \mathscr{T}_F[u_1] \otimes \mathscr{T}_F[u_2]. \tag{20.10}$$

  To show this, first note that since $\mathscr{T}_F^{-1} \;=\; \mathcal{F}/\sqrt{2\pi}$, it is also a Fourier transform hence must satisfy the convolution theorem itself, i. e.

$$\mathscr{T}_F^{-1}[U_1 \otimes U_2] \;=\; \mathscr{T}_F^{-1}[U_1] \times \mathscr{T}_F^{-1}[U_2], \tag{20.11}$$

  then Fourier transforming both sides and using $\mathscr{T}_F \mathscr{T}_F^{-1} \;=\; \mathrm{id}$, the identity map, we have

$$U_1 \otimes U_2 \;=\; \mathscr{T}_F\big[\mathscr{T}_F^{-1}[U_1] \times \mathscr{T}_F^{-1}[U_2]\big], \tag{20.12}$$

  which we may then proceed to get equation 20.10 by setting $U_1 \;=\; \mathscr{T}_F[u_1]$ and $U_2 \;=\; \mathscr{T}_F[u_2]$.

- In both cases a Dirac-delta centred at 0 and a constant are Fourier transforms of one another. In quantum mechanics, this means that if I have the particle in a state with definite position at $x \;=\; 0$, then its momentum wavefunction would be uniform across all space. In optics, this means that if I have an infinitesimally thin slit at the origin, then I shall see a constant intensity as the image. However, since the momentum wavefunction needs to be normalised and an uniform distribution across an infinite distance is non-normalisable, and the total amount of light through the infinitesimally thin slit is zero, what we observe in practice is nothing at all for both of these two cases. If we translate the Dirac-delta, then its Fourier transform is modulated, i. e. we see a monochromatic plane wave across the momentum space (or just a phase in the image plane in the optics case, which means that the intensity is still constant across the whole plane i. e. there is no change in the image).

Now let us look at the Fourier transforms of basic functions.

Some Fourier Transform Primitives

Here we tabulate some basic principles that can help us in interpreting Fourier Transform between functions.

- We have discussed this just now: the Fourier transform of a point source that is not located at 0 gives an oscillatory image, i. e.

$$\mathscr{T}_F[u_0 \delta(x - x_0)] \;=\; u_0 \mathrm{e}^{\mathrm{i} x_0 \beta}. \tag{20.13}$$

- We have discussed this in the previous section: the Fourier transform of a tophat function is a sinc function, i. e.

$$\mathscr{T}_F[\text{tophat}_a\,(x)] \;=\; \text{sinc}\,(a\beta/2). \tag{20.14}$$

- This has been discussed in the quantum mechanics course: the Fourier transform of a Gaussian wavefunction is still a Gaussian wavefunction:

$$\mathscr{T}_F\left[\frac{u_0}{\sigma_x\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2\right]\right\}\right] \;=\; u_0\mathrm{e}^{\mathrm{i}kx_0}\mathrm{e}^{-\frac{1}{2}\beta^2{\sigma_x}^2}, \tag{20.15}$$

where we have the standard deviation for the spatial frequency β

$$\sigma_\beta \;=\; 1/\sigma_x \quad\Rightarrow\quad \sigma_x\sigma_\beta \;=\; 1. \tag{20.16}$$

These scalar amplitudes gives the intensity distribution functions as

$$I(x) \;=\; \mathrm{e}^{-[(x-x_0)/\sigma_x]^2} \quad\text{and}\quad \mathrm{I}(\beta) \;=\; \mathrm{e}^{-\beta{\sigma_x}^2}, \tag{20.17}$$

giving variances in the *intensity* distributions as

$$\text{Var}\,[I(x)] \;=\; \frac{1}{2{\sigma_x}^2}, \qquad \text{Var}\,[\mathrm{I}(\beta)] \;=\; \frac{{\sigma_x}^2}{2}, \quad\Rightarrow\quad \text{Var}\,[I(x)]\times\text{Var}\,[\mathrm{I}(\beta)] \;=\; \frac{1}{4}, \tag{20.18}$$

same as the **uncertainty relation** in quantum mechanics:

$$\Big\langle(\Delta x)^2\Big\rangle\Big\langle(\Delta p)^2\Big\rangle \;=\; \langle\psi|\,(\Delta x)^2\,|\psi\rangle\times\langle\psi|\,\hbar^2(\Delta k)^2\,|\psi\rangle \;=\; \frac{\hbar^2}{4}. \tag{20.19}$$

This is another illustration of the fact that: the smaller the feature is on the diffraction mask, the more extended is the feature on the screen.

- We consider a plane wave through a wide open plane i. e. the scalar amplitude $u_{\text{plane}}(\mathbf{x})$ is assumed to be uniform throughout the wide open plane, but there is an obstruction in the middle. We compare the intensity pattern of this situation to the intensity pattern where the whole plane is obstructed, but we open a slit in the obstruction such that some light goes through, which the scalar amplitude is given by $u_{\text{slit}}(\mathbf{x})$. We shall note that the scalar amplitude of the light in the former case, i. e. light through the obstructed plane, is

$$u(\mathbf{x}) \;=\; u_{\text{plane}}(\mathbf{x}) - u_{\text{slit}}(\mathbf{x}), \tag{20.20}$$

and therefore the scalar amplitude on the screen is given as

$$U(\boldsymbol{\beta}) \;=\; \delta^{(2)}(\boldsymbol{\beta}) - U_{\text{slit}}(\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \;=\; (\beta_x,\beta_y)^{\mathrm{T}}, \tag{20.21}$$

exploiting the linearity of the Fourier transform, and note that we approximate the Fourier transform of the wide feature $u_{\text{plane}}(\mathbf{x})$ as a narrow spot on the image plane $\delta^{(2)}(\boldsymbol{\beta})$. As a result, the intensity pattern with the obstruction would be

$$\mathrm{I}(\boldsymbol{\beta}) \;=\; \begin{cases}\text{high intensity} & \beta \;=\; 0;\\ \mathrm{I}_{\text{slit}}(\boldsymbol{\beta}) & \text{otherwise},\end{cases} \tag{20.22}$$

i. e. the same as the intensity pattern with the slit, but with a very bright peak on the principal axis. This is called **Babinet's principle**.

Now we are well-acquainted with the theory of interference, we are equipped to investigate into diffraction patterns through more complicated diffraction masks.

Summary

1. The Fraunhofer integral is a Fourier transform, with minor differences from that used in quantum mechanics. The properties are mostly carried through, although some of the properties require minor modifications.

2. We collect a number of primitives of Fourier transforms, which will then equip us better for more complicated diffraction masks.

4 Grating

§21. Transmission Grating

Image of a Grating

We shall now march onto the optics of diffraction gratings (or, just gratings), which are useful in separating colours, or, wavelengths. As an example, this is important in distinguishing between spectra of different elements, which is a crucial part of finding the chemical composition of stars.

The simplest grating is called a transmission grating, which is just a large obstruction with thin and long slits that are evenly spaced, which we model as

$$u(x) \;=\; u_0\,\mathrm{grating}_d{}^{(N)}(x) \;=\; u_0\sum_{m\,=\,1}^{N}\delta(x-md) \tag{21.1}$$

for a grating with N slits and consecutive slits separated by a distance d. For this simple treatment we shall ignore the slit width of the grating. Then, the scalar amplitude on the screen is given as

$$\begin{aligned} U(\beta) \;&=\; u_0\mathscr{T}_F\big[\mathrm{grating}_d{}^{(N)}(x)\big] \;=\; u_0\sum_{m\,=\,1}^{N}\mathscr{T}_F[\delta(x-md)] \;=\; u_0\sum_{m\,=\,1}^{N}\left(\mathrm{e}^{\mathrm{i}\delta}\right)^m \\ &=\; u_0\times\text{common phase}\times\frac{\sin\left(\frac{1}{2}N\delta\right)}{\sin\left(\frac{1}{2}\delta\right)}, \end{aligned} \tag{21.2}$$

where $\delta \;=\; \beta d \;=\; kd\sin\theta$. Note that we have utilised the equation of the summation of a geometric series when working out the sum over the exponentials. This then gives the intensity distribution as

$$\mathrm{I}(\beta) \;=\; I_0\frac{\sin^2\left(\frac{1}{2}N\delta\right)}{\sin^2\left(\frac{1}{2}\delta\right)} \;=\; I_0\frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)}, \tag{21.3}$$

where I_0 is the maximum intensity of the distribution of *each individual slit*. Noting that, by analysing the intensity distribution function, the maximum intensity is given by I_0N^2. The shapes of the intensity distribution pattern is shown in figure 21.1.

Analysis of the Image

We shall first locate the maxima and the minima. Naively, the maxima should be located when every pair of consecutive slits constructively interfere, i. e.

$$\mathrm{OPL} \;=\; d\sin\theta \;\overset{!}{=}\; p\lambda \quad\Rightarrow\quad \sin\theta \;=\; p\lambda/d \tag{21.4}$$

where p is an integer. This is verified by setting the denominator of equation 21.3 to zero, i. e.

$$\frac{1}{2}\beta d \;=\; \frac{1}{2}kd\sin\theta \;=\; \frac{1}{2}\times\frac{2\pi}{\lambda}\times d\sin\theta \;=\; p\pi \quad\Rightarrow\quad \sin\theta \;=\; \frac{p\lambda}{d}. \tag{21.5}$$

These maxima are sometimes called **bright fringes**. Note that every time the denominator of equation 21.3 is zero, the numerator will always be zero at the same time, which means that the intensity will always be finite. We shall also note that there are small-scaled periodic minima when the numerator is zero but the denominator is non-zero at equally-spaced angles that has a spacing $\Delta(\sin\theta) \;=\; \lambda/(Nd)$, i. e. every consecutive pair of maxima has $N-1$ small-scaled

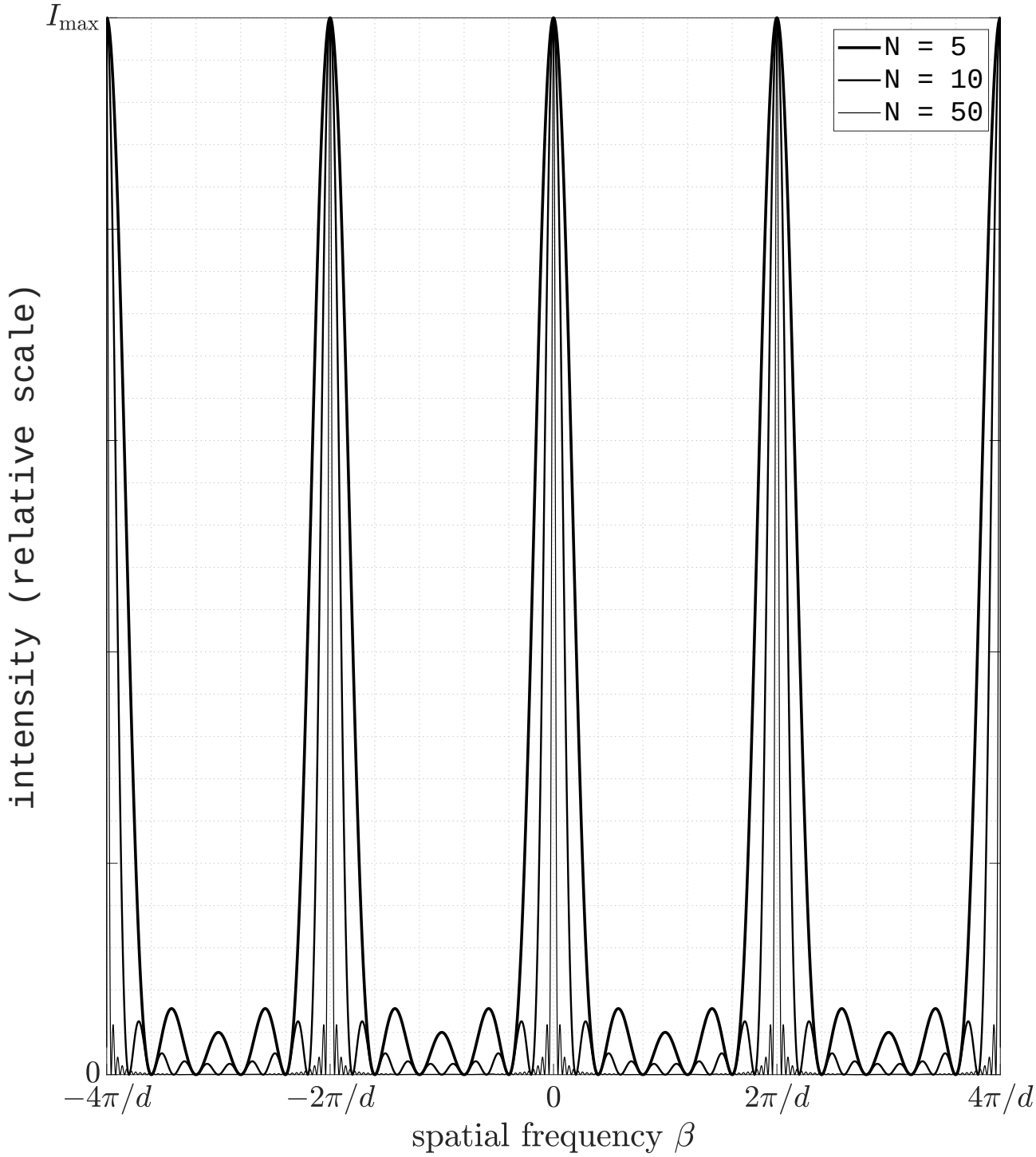

Figure 21.1: The image of a diffraction grating. Note that the image with large N has a much higher value of maximum intensity proportional to N^2, but in this image it is suppressed so that we can compare the shapes of the different images.

minima in between.

We shall also take a special note that if the grating has an infinite number of slits, then the scalar amplitude will be

$$U(\beta) \;=\; u_0 \lim_{N\to\infty} \frac{\sin\left(\frac{1}{2}N\delta\right)}{\sin\left(\frac{1}{2}\delta\right)} \;=\; u_0 \lim_{N\to\infty} \mathrm{grating}_{2\pi/d}{}^{(N)}(\beta), \tag{21.6}$$

i. e. Dirac-deltas that are evenly spaced on the spatial-frequency domain with $\Delta\beta = 2\pi/d$ extending all the way to infinity. Note that this is equivalent to the shape of the distribution in Figure 21.1 in the $N \to \infty$ limit.

Finally we shall investigate into the phasor representation of the grating. This time we are adding them as the number of slits are finite, and thus we do vector addition on the complex plane. We shall illustrate this with the $N = 3$ and $N = 5$ case, with the intensity pattern and the corresponding phasor representation of certain values of βs shown in figures 21.2 and 21.3.

Summary

1. The distribution of scalar amplitude of a grating on the source plane is

$$u(x) \;=\; u_0\,\mathrm{grating}_d{}^{(N)}(x) \;=\; u_0 \sum_{m\,=\,1}^{N} \delta(x - md), \tag{21.7}$$

   and intensity distribution of the image of the grating is given as

$$\mathrm{I}(\beta) \;=\; I_0 \frac{\sin\left(\frac{1}{2}N\delta\right)}{\sin\left(\frac{1}{2}\delta\right)} \;=\; I_0 \frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)}. \tag{21.8}$$

2. The location of the maxima is $\sin\theta = p\lambda/d$ and the location of the minima is $\sin\theta = p\lambda/(Nd)$: the separation of the maxima is N times the separation of the small-scaled maxima. If the number of slits is infinite, then the image is an infinite "comb" of Dirac-deltas on the domain of spatial frequencies, with separation $\Delta\beta = 2\pi/d$.

§22. Grating as a Spectrometer

Finding the Wavelength of an Unknown Source

Now we have a colour made up of several unknown wavelengths and we want to find these wavelengths by sending the light through a grating. At the other side of the grating we see the colours separate into different angles, which, if we then place a lens after the colours are separated, become distances. If we capture the image on the other side by an eyepiece, then the image will be similar to figure 22.1, where we are looking into how a source with a continuous spectrum separates into. If the wavelengths are discrete, then it will fall into lines that are separated. The master equation that we shall apply is equation 21.5, which after slight algebraic manipulations we have

$$\theta \;=\; \arcsin\left(p\frac{\lambda}{d}\right), \qquad \lambda \;=\; \frac{d}{p}\sin\theta, \tag{22.1}$$

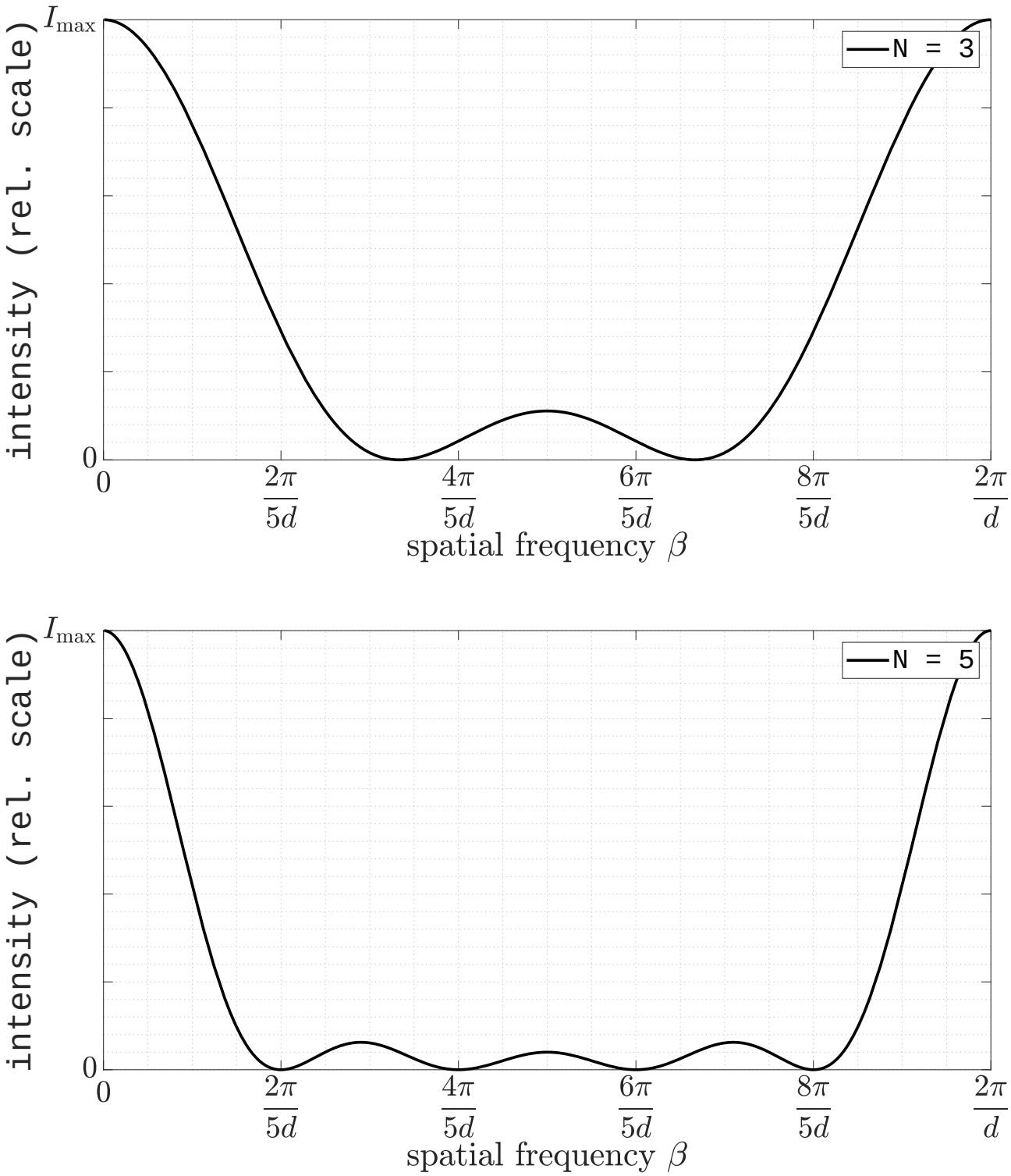

Figure 21.2: The image of a diffraction grating with $N = 3$ and $N = 5$.

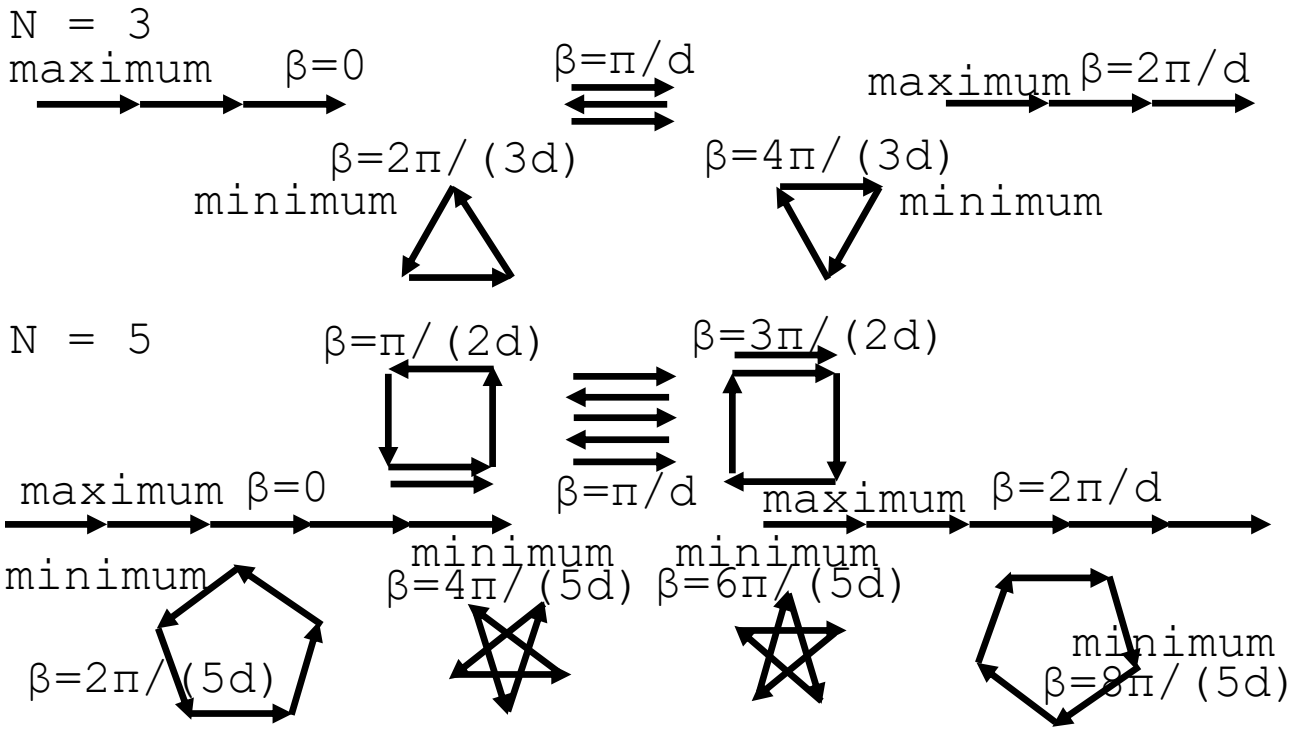

Figure 21.3: The phasor representation for the maxima and minima of the intensity patterns for the diffraction grating $N = 3$ and $N = 5$ case.

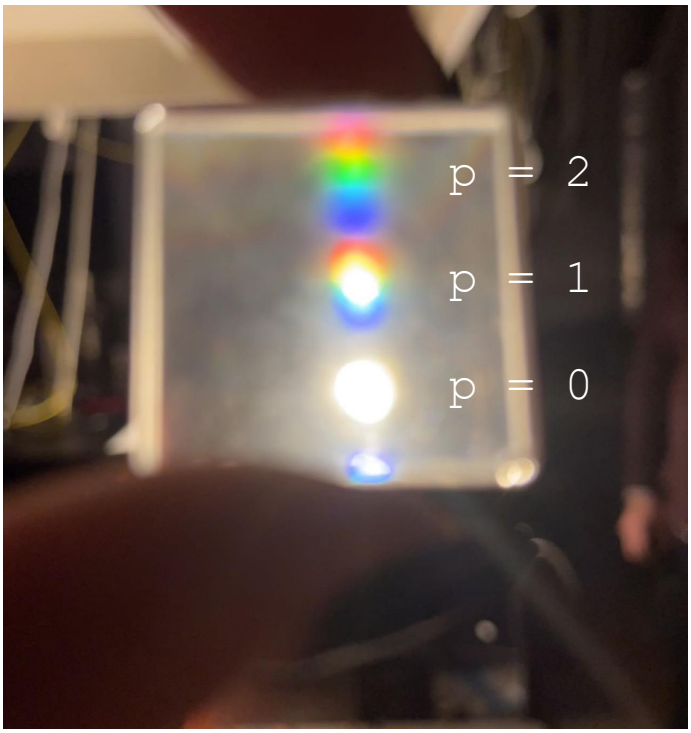

Figure 22.1: The appearance of light with a continuous spectrum after sent through a transmission grating. The photograph is captured by a phone camera, which contains a lens which transfers the angles onto distances on the CCD sensor of the phone.

a link between λ and θ. Here the integer p is called the **order**, and the maxima corresponding to the p^{th} order is sometimes labelled as θ_p. p can be determined by counting the number of maxima from the central maximum (where $p = 0$). The slit separation d usually comes with the grating — when one purchases a diffraction grating we are able to tell d from the manufacturer's instrumentation, usually given in lines per centimetres, with typical values of d across 10^3 to 10^4 lines per centimetre. With all these three pieces of information, we may then work out λ using equation 22.1.

Instrumental Range and Width

We now use the grating to resolve two different colours (or, wavelengths of the light) from the same source. To check whether we can do this, we check two things.

- The two wavelengths are not too far apart, such that the light coming from different orders do not overlap. For example, in figure 22.1, the red light from the first order and the blue light of the second order is challenging this limit. This maximum difference in wavelength that the apparatus can separate is called the **instrumental range**, or the **free spectral range**. To calculate the free spectral range for light at order p, if the light at order $p+1$ has wavelength λ and the light at order p has wavelength $\lambda + \text{FSR}_\lambda$, then using equation 22.1, the free spectral range is given as

$$\text{FSR}_\lambda = \frac{d}{p}\sin\theta - \frac{d}{p+1}\sin\theta = \frac{d}{p}\times\frac{(p+1)\lambda}{d} - \frac{d}{p+1}\times\frac{(p+1)\lambda}{d} = \frac{\lambda}{p}. \tag{22.2}$$

- The two wavelengths are not too close together, so the image of the two different colours of the same order do not overlap. This minimum difference in wavelength that the apparatus can separate is called the **instrumental width**. To find this, we compare the separation of the centre of the image of the two wavelengths that we want to resolve with the width of the minima — applying the Rayleigh criterion. We shall use our usual approximation that the width of the maxima is the distance from the centre of the maxima to the first minima, which, we note is N times smaller than the separation between two orders of the same colour $\Delta(\sin\theta) = \lambda/d$. This gives the instrumental width as

$$\text{INST}_\lambda = \frac{d}{p}\times\frac{\Delta(\sin\theta)}{N} = \frac{\lambda}{Np}, \tag{22.3}$$

  or, sometimes expressed in wavenumbers, which can be calculated just by treating this width as an "experimental error" and propagating this error through using the error correction formula,

$$\text{INST}_{\bar\nu} = \left|\frac{\mathrm{d}\bar\nu}{\mathrm{d}\lambda}\right| \times \text{INST}_\lambda = \frac{1}{\lambda^2}\times\text{INST}_\lambda = \frac{\bar\nu}{Np}. \tag{22.4}$$

The instrumental range FSR_λ and the instrumental width INST_λ are two labels for the ability of resolving different wavelengths for a spectrometer, which we will discuss for other types of interferometric apparatuses later.

Chromatic Dispersion and the Chromatic Resolving Power

From the instrumental width, there are some more definitions that are used to describe the grating,

- We shall note that the limit in angular separation that we can resolve is given by

$$\text{INST}_\theta \;=\; \frac{\mathrm{d}\theta}{\mathrm{d}\lambda} \times \text{INST}_\lambda \;=\; \frac{p}{d\cos\theta} \times \text{INST}_\lambda \;=\; \frac{p\lambda}{Nd\cos\theta}, \tag{22.5}$$

  where

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} \;=\; \frac{1}{\mathrm{d}\lambda/\mathrm{d}\theta} \;=\; \frac{p}{d\cos\theta} \tag{22.6}$$

  is sometimes called the **chromatic dispersion**.

- The **chromatic resolving power** is usually used to describe the ability of an optical apparatus in separating similar wavelengths. This is defined by

$$\mathscr{P} \;=\; \frac{\text{operating wavelength}}{\text{resolvable difference in } \lambda} \;=\; \frac{\lambda}{\text{INST}_\lambda} \;=\; \frac{\lambda^2 \times \bar{\nu}}{\lambda^2 \times \text{INST}_{\bar{\nu}}} \;=\; \frac{\bar{\nu}}{\text{INST}_{\bar{\nu}}}, \tag{22.7}$$

  which for a grating we have

$$\mathscr{P} \;=\; Np, \tag{22.8}$$

  where p is the order we are operating on.

Previously we have modelled each slit in the grating as a Dirac-delta, which is usually not the case. A better model is to treat each slit as a source with a finite extent i. e. a single slit described by a tophat function, which can be calculated from the convolution theorem, which we shall look at next.

Summary

1. To find the wavelength of a light source, we pass it through a grating, which we calculate the wavelength from

$$\lambda \;=\; \frac{d}{p}\sin\theta. \tag{22.9}$$

2. To use the grating to resolve two different colours, we want to make sure that the separation in wavelength is smaller than the instrumental range

$$\text{FSR}_\lambda \;=\; \lambda/p \tag{22.10}$$

   where λ is the smaller wavelength, and is larger than the instrumental width

$$\text{INST}_\lambda \;=\; \lambda/(Np). \tag{22.11}$$

3. The angular resolution of the grating is given as

$$\text{INST}_\theta \;=\; \frac{p}{Nd\cos\theta} \tag{22.12}$$

   and the chromatic resolving power of the grating is given as

$$\mathscr{P} \;=\; \frac{\lambda}{\text{INST}_\lambda} \;=\; \frac{\bar{\nu}}{\text{INST}_{\bar{\nu}}} \;=\; Np. \tag{22.13}$$

§23. Convolution Theorem

Practicalities in Using the Convolution Theorem

Recall that convolution is defined as

$$u_1(x) \otimes u_2(x) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x'\, u_1(x') u_2(x - x'), \tag{23.1}$$

and the convolution theorem reads

$$\mathscr{T}_F[u_1 \otimes u_2] \;=\; \mathscr{T}_F[u_1] \times \mathscr{T}_F[u_2] \tag{23.2}$$

forwards and

$$\mathscr{T}_F[u_1 \times u_2] \;=\; \mathscr{T}_F[u_1] \otimes \mathscr{T}_F[u_2] \tag{23.3}$$

backwards. Practically in optics the convolution theorem is used to find the Fourier transform of more complicated functions. In order to do this, we need to know how to separate a more complicated function into convolutions of two functions with the Fourier transform that we know how to do. Two practical separation methods are given as follows.

- A grating with finite slit width has a scalar amplitude proportional to the convolution of a regular array of Dirac-deltas (or, the scalar amplitude proportional to grating with negligible width) and a tophat function (or, the scalar amplitude proportional to a single slit). This can be shown explicitly as

$$\begin{aligned} \mathrm{grating}_d{}^{(N)}(x) \otimes \mathrm{tophat}_a(x) \;&=\; \int_{-\infty}^{\infty} \mathrm{d}x' \sum_{m=1}^{N} \delta(x' - md) \times \mathrm{tophat}_a(x - x') \\ &=\; \sum_{m=1}^{N} \mathrm{tophat}_a(x - md), \end{aligned} \tag{23.4}$$

  exactly a grating with finite width.

- A symmetric triangle function, for example $f(x) \;=\; \max(2a - |x|, 0)$, can be expressed as a convolution of two tophats with half the width of the triangle.

Now let us try to use the convolution theorem with a simple setup.

Two-Slit Interference with Finite Slit Widths

To test the power of the convolution theorem, we can consider a two-slit interference problem with slit separation d and finite slit widths a, with scalar amplitude on the source plane as

$$u(x) \;=\; u_0 \times \big\{[\delta(x) + \delta(x - d)] \otimes \mathrm{tophat}_a(x)\big\}. \tag{23.5}$$

From here we can work out

$$\mathscr{T}_F[\delta(x) + \delta(x - d)] \;=\; \cos\left(\frac{\delta}{2}\right) \quad \text{and} \quad \mathscr{T}_F[\mathrm{tophat}_a(x)] \;=\; \mathrm{sinc}\left(\frac{1}{2}\beta a\right), \tag{23.6}$$

up to an overall phase. Using the convolution theorem, we have

$$\begin{aligned} U(\beta) \;&=\; \mathscr{T}_F\Big[u_0 \times \big\{[\delta(x) + \delta(x - d)] \otimes \mathrm{tophat}_a(x)\big\}\Big] \\ &=\; u_0 \times \mathscr{T}_F[\delta(x) + \delta(x - d)] \times \mathscr{T}_F[\mathrm{tophat}_a(x)] \\ &=\; u_0 \cos\left(\frac{1}{2}\beta d\right) \mathrm{sinc}\left(\frac{1}{2}\beta a\right), \end{aligned} \tag{23.7}$$

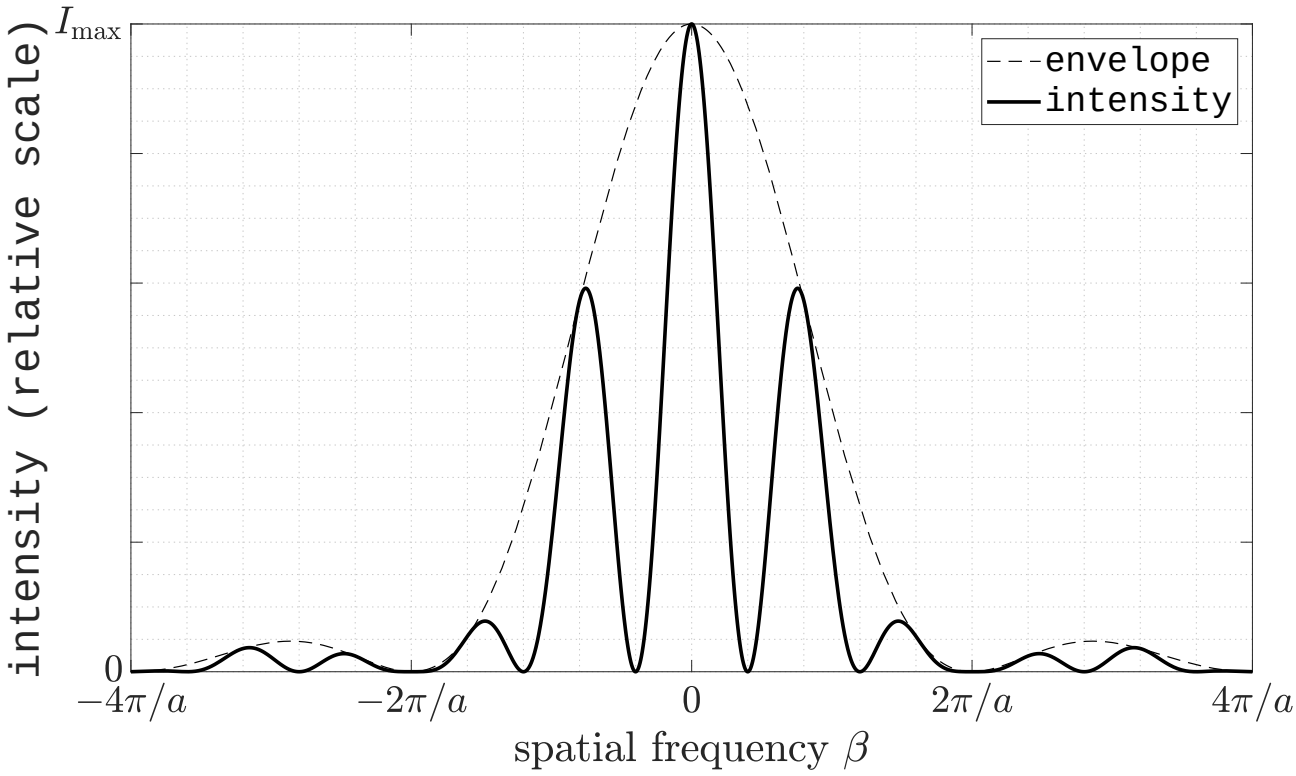

Figure 23.1: The two-slit interference pattern with slit separation d and slit width a.

up to an overall phase, and hence the intensity is

$$\mathrm{I}(\beta) \;=\; I_{\max}\cos^2\left(\frac{1}{2}\beta d\right)\mathrm{sinc}^2\left(\frac{1}{2}\beta a\right). \tag{23.8}$$

We note that by the geometry, the slit separation cannot be smaller than the slit width. As a result, the sinc function gives maxima that are wider apart. Therefore the sinc function will form an envelope of the cosine function, shown in figure 23.1. Note that again we have the smallest feature on the source plane, the slit width, gives rise to the largest feature, the envelope, on the image plane.

Summary

1. The convolution theorem allows us to do Fourier transforms for complicated function by claiming that the Fourier transform of it is the product of the Fourier transform of the functions that convolutes it. Two common tricks is to tore a periodic array of tophat functions (describing a grating with finite slit width) apart into a convolution of a regularly spaced Dirac-deltas and a tophat function; and to tore a symmetric triangle function apart into a convolution of two tophats.

2. Equipping a pair of slits with a finite slit width gives an envelope on top of the fringes of the slits treated as two Dirac-deltas.

§24. Grating with Finite Slit Width

Convolution Theorem, Forwards

The use of convolution theorem forwards to find the intensity distribution of a grating with finite slit width is a straightforward extension from the two-slit case. We simply change the

two Dirac-deltas into N evenly spaced Dirac-deltas and convolute that with the tophat instead. This gives

$$u(x) \;=\; u_0 \times \text{grating}_d{}^{(N)}(x) \otimes \text{tophat}_a\,(x). \tag{24.1}$$

The convolution theorem then gives

$$\begin{aligned} U(\beta) \;&=\; \mathscr{T}_F\big[u_0 \times \text{grating}_d{}^{(N)}(x) \otimes \text{tophat}_a\,(x)\big] \\ &=\; u_0 \times \mathscr{T}_F\big[\text{grating}_d{}^{(N)}(x)\big] \times \mathscr{T}_F[\text{tophat}_a\,(x)] \\ &=\; u_0 \frac{\sin\left(\frac{1}{2}N\beta d\right)}{\sin\left(\frac{1}{2}\beta d\right)} \times \text{sinc}\left(\frac{1}{2}\beta a\right) \end{aligned} \tag{24.2}$$

up to an overall phase. The intensity distribution of the image is given as

$$\text{I}(\beta) \;=\; I_0 \times \frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)} \times \text{sinc}^2\left(\frac{1}{2}\beta a\right). \tag{24.3}$$

This is plotted in figure 24.1. Note that the effect of the slit width is that it focuses on the order $p \;=\; 0$ and reduces the intensity of higher orders, extinguishing intensities in certain orders, for example in figure 24.1, the slit separation is three times the slit width, and as a result the order $p \;=\; 3$ has no intensity at all, so if we would like to use it as a spectrometer, then we simply cannot use that order at all.

This effect of intensity suppression at high orders is disastrous if we want to use the grating as a spectrometer, as recall that the instrumental width is

$$\text{INST}_{\bar{\nu}} \;=\; \bar{\nu}/(Np) \tag{24.4}$$

so to distinguish between finer wavelengths we want to have a small instrumental width, and therefore we want to go to as high order as possible (but of course the order is limited by $\sin\theta \;<\; 1$), however the intensity at that order is greatly reduced. Most of the light is at the order $p \;=\; 0$ but all wavelengths emerges at the same angle, so this is completely useless in distinguishing different wavelengths. In the next section we shall introduce methods to reduce this effect.

Convolution Theorem, Backwards

Note that we can also formulate the grating with an finite number of slits as a product of a grating with an infinite number of slits with a finite slit width and a tophat function with exactly the width as the number of slits, so the slits outside the tophat function are scaled to zero. This gives the scalar amplitude at the source as

$$u(x) \;=\; u_0\Big\{[\text{tophat}_a\,(x)] \otimes \Big[\lim_{M\,\to\,\infty} \text{grating}_d{}^{(M)}(x)\Big]\Big\} \times [\text{tophat}_{Nd}\,(x)]. \tag{24.5}$$

This then gives, using the convolution theorem,

$$U(\beta) \;=\; u_0\left\{\text{sinc}\left(\frac{1}{2}\beta a\right) \times \Big[\lim_{M\,\to\,\infty} \text{grating}_{2\pi/d}{}^{(N)}(\beta)\Big]\right\} \otimes \text{sinc}\left(\frac{1}{2}Nd\beta\right). \tag{24.6}$$

From here onwards it is difficult to push forward any strict mathematical reasoning, and therefore we have to argue just using logic. The first two terms now become a product, which gives regularly spaced Dirac-deltas with period $\Delta\beta \;=\; 2\pi/d$, with an envelope with a large-scale

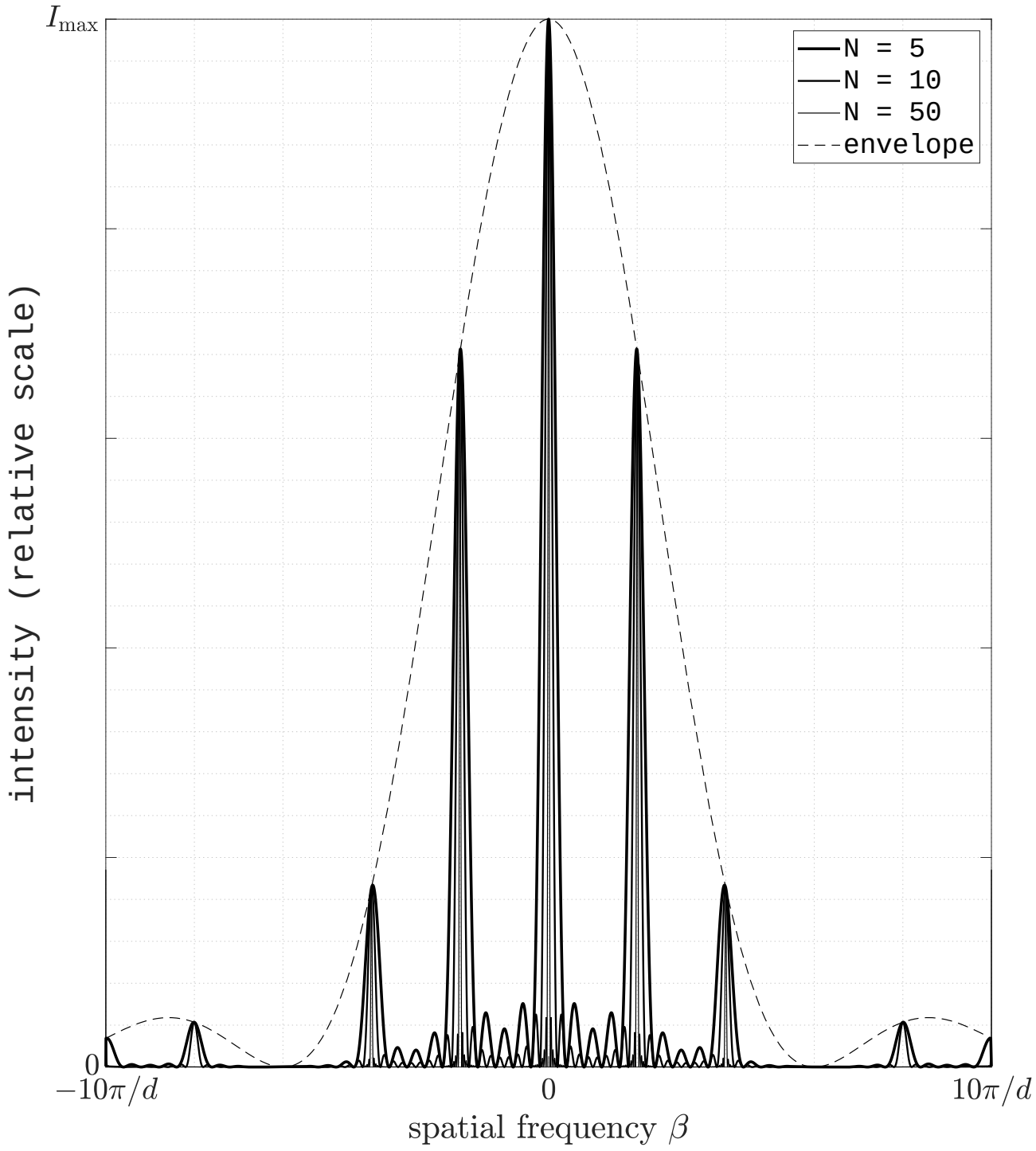

Figure 24.1: A grating with slit separation d and slit width a.

variation with period inversely proportional to the slit width a. Finally we have the sinc function, which convolutes with the Dirac-delta, which changes them into sinc-like functions with the period inversely proportional to Nd. However by looking at the exact form of the scalar amplitude, sinc is not quite the right description for the shapes caused by the individual slits.

Again, we note that the largest feature at the source, the large tophat with size Nd, gives the smallest feature on the image, which is the individual peaks. The smallest feature, which is the slit width a, givers the the larges feature on the image, which is the large-scale envelope.

Summary

1. The intensity pattern of the diffraction grating with separation d and slit width a has an intensity distribution

$$\mathrm{I}(\beta) \;=\; I_0 \times \frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)} \times \mathrm{sinc}^2\left(\frac{1}{2}\beta a\right). \tag{24.7}$$

2. The effect of the addition of the slit width gives rise to a large-scale envelope with minima located at $\beta \;=\; 2p\pi/a$ to the image of the grating with negligible slit width, and a reduced intensity for higher orders of diffraction.

§25. Reflection Grating

Reflection Grating

A common method to overcome the problem of intensity suppression at higher orders is to use a **blazed reflection grating**. Before introducing the blaze, we shall first look at just a simple reflection grating. This is just made of N strips of mirrors with size a separated by d, and one shines light on the grating at an oblique angle. Note that the physics at this sort of reflection grating is exactly the same as a transmission grating, but with the input at an oblique angle α. The link is shown in figure 25.1.

To analyse the transmission grating with a non-zero angle of incidence, simply note that there is an extra phase shift of $kd\sin\alpha$ between consecutive slits *before* the light hits the grating. As a result, the shape of the intensity distribution function is exactly the same, i. e. ,

$$\mathrm{I}(\beta) \;=\; I_0 \times \frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)} \times \mathrm{sinc}^2\left(\frac{1}{2}\beta a\right), \tag{25.1}$$

but now with β defined alternatively as

$$\beta \;=\; k(\sin\theta - \sin\alpha). \tag{25.2}$$

This means that the zeroth order diffraction pattern is no longer at $\theta \;=\; 0$ but at $\theta \;=\; \alpha$. The reflection grating has exactly the same intensity pattern. However, even if the zeroth order is now oblique to the grating, it still does not separate wavelengths and is still where most of the light ends up at, and therefore is equally useless compared to the transmission grating. The real trick that makes this all works out is by blazing the grating.

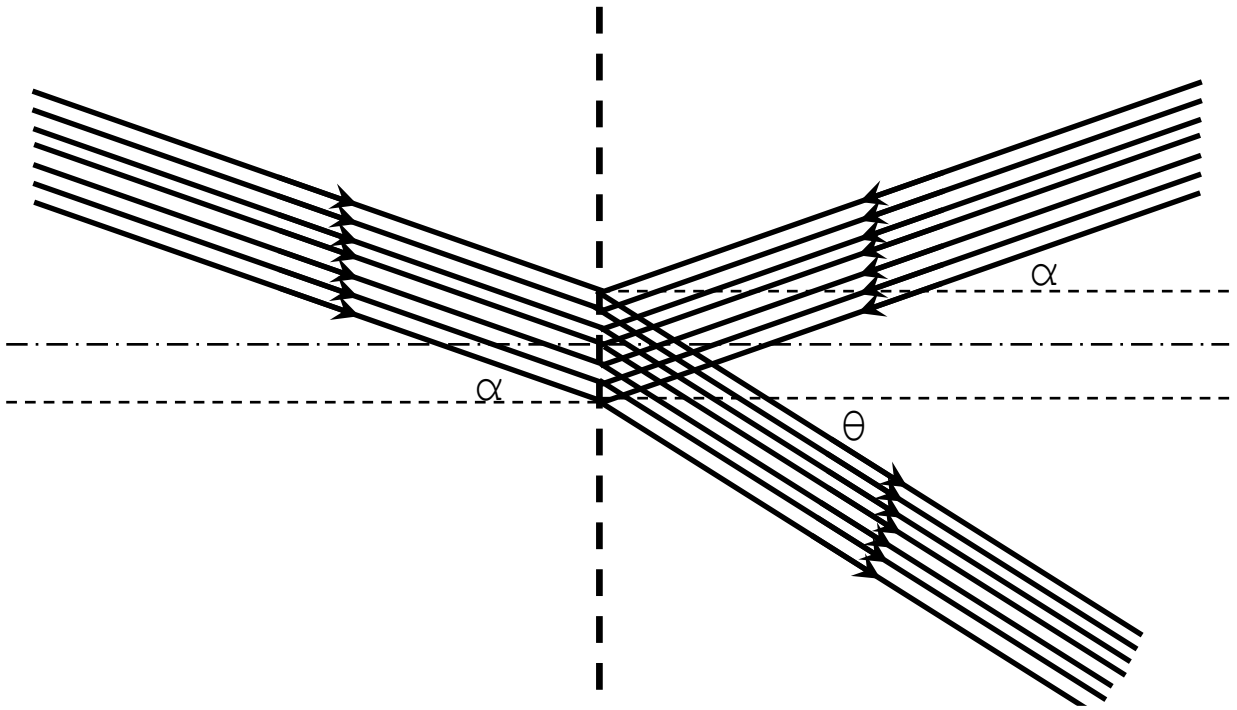

Figure 25.1: The equivalence between a transmission grating with an oblique angle of incidence and a reflection grating. The dashed line represents the position of the grating.

Blazed Gratings

The idea of blazing a grating is to introduce a blazing angle φ_{bl} that the mirrors are positioned with respect to the normal. This is shown in figure 25.2. With respect to the normal of the principal axis of the grating, the angle of entrance and exit is α and θ respectively, and when α and θ are equal, we have no phase difference i. e. we have the zeroth order reflection. However with respect to the normal of the mirrors, the angle of entrance and exit is φ_i and φ_r respectively, which suggests that the central maximum of the single slit will be at $\sin\varphi_i = \sin\varphi_r$ (equating the phase shifts *before* and *after* entering the single slit) i. e. the maximum intensity is *not* at $p = 0$, but when $\varphi_i = \varphi_r$. In terms of the intensity distribution, this is equivalent to preserving the fringes but shifting the envelope, i. e.

$$\mathrm{I}(\beta) = I_0 \times \frac{\sin^2\left(\frac{1}{2}N\beta d\right)}{\sin^2\left(\frac{1}{2}\beta d\right)} \times \mathrm{sinc}^2\left(\frac{1}{2}\beta_\varphi a\right), \quad \beta = k(\sin\theta - \sin\alpha), \quad \beta_\varphi = k(\sin\varphi_r - \sin\varphi_i) \tag{25.3}$$

so now the maximum intensity is at the order that we want to look at more carefully, instead of at $p = 0$: for example we might want to blaze at first order, which then gives an intensity pattern as figure 25.3.

Note that, using the geometry of the setup, at maximum intensity,

$$\varphi_i = \varphi_r = (\theta + \alpha)/2, \qquad \varphi_{\text{bl}} = (\theta - \alpha)/2. \tag{25.4}$$

In practice if we would like to buy a reflection grating to distinguish two very close wavelengths, then of course the manufacturer does not have all the different blaze angles we would like, and therefore we shall select the angles of θ and α that roughly matches the blaze angles available at the manufacturer.

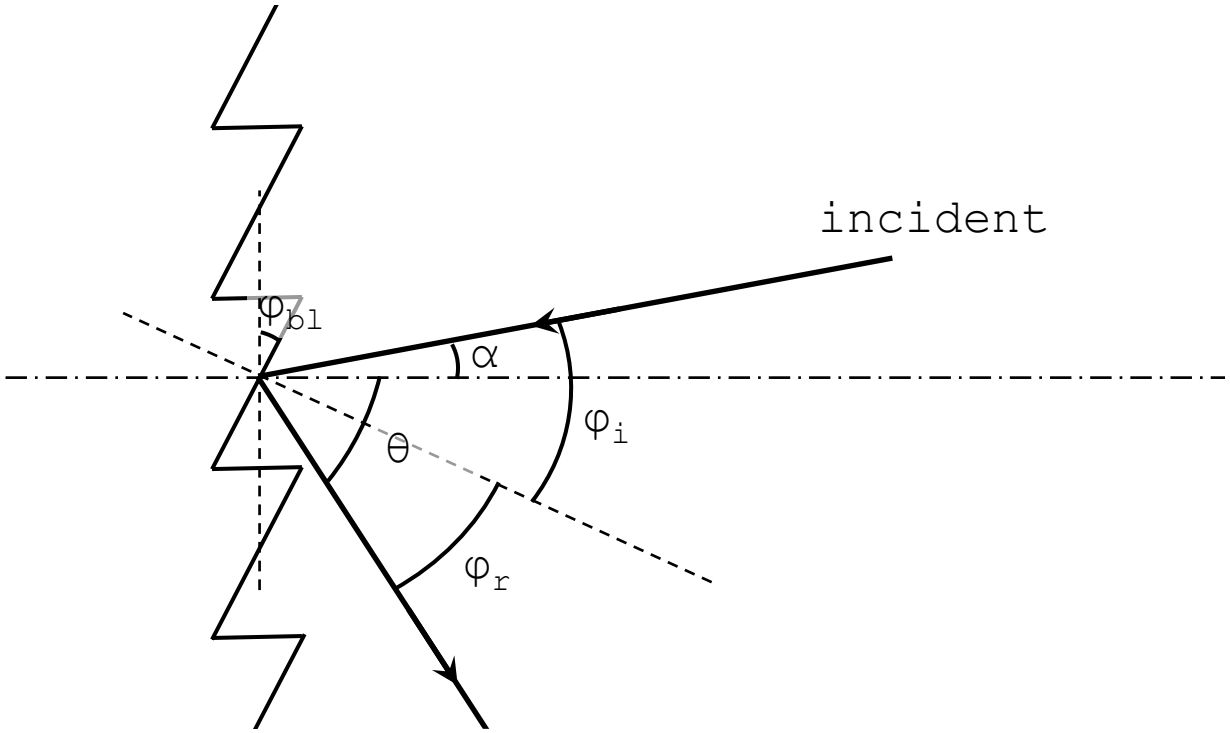

Figure 25.2: A blazed reflection grating.

Summary

1. An unblazed reflection grating has an intensity pattern that is the same as the transmission grating where the light is incident on the grating obliquely. The spatial frequency therefore needs to be modified to

$$\beta \;=\; k(\sin\theta - \sin\alpha). \tag{25.5}$$

2. The intensity of the light out of the diffraction grating decreases as we go to higher and higher orders. To solve this issue we can blaze the grating, which means that we have the maximum intensity be at a higher order, instead of at $p = 0$ where the wavelengths cannot be resolved at all.

§26. Spectrometer Designs

Czerny-Turner Configuration

In this section we shall introduce two of the interferometer setups using the blazed reflection grating.

Obviously the easiest method of building the spectrometer is to just use a lens to collimate the light such that they are input into the grating at right angles, and use a lens after the light passes the grating to focus it such that it can be detected by eye or by a camera. However this setup suffers from spherical and coma aberrations of the lens, plus only a specific range of wavelengths can transmit through glass, therefore is not the most ideal design.

To avoid these aberrations, one of the methods to change the collimating and focusing lenses

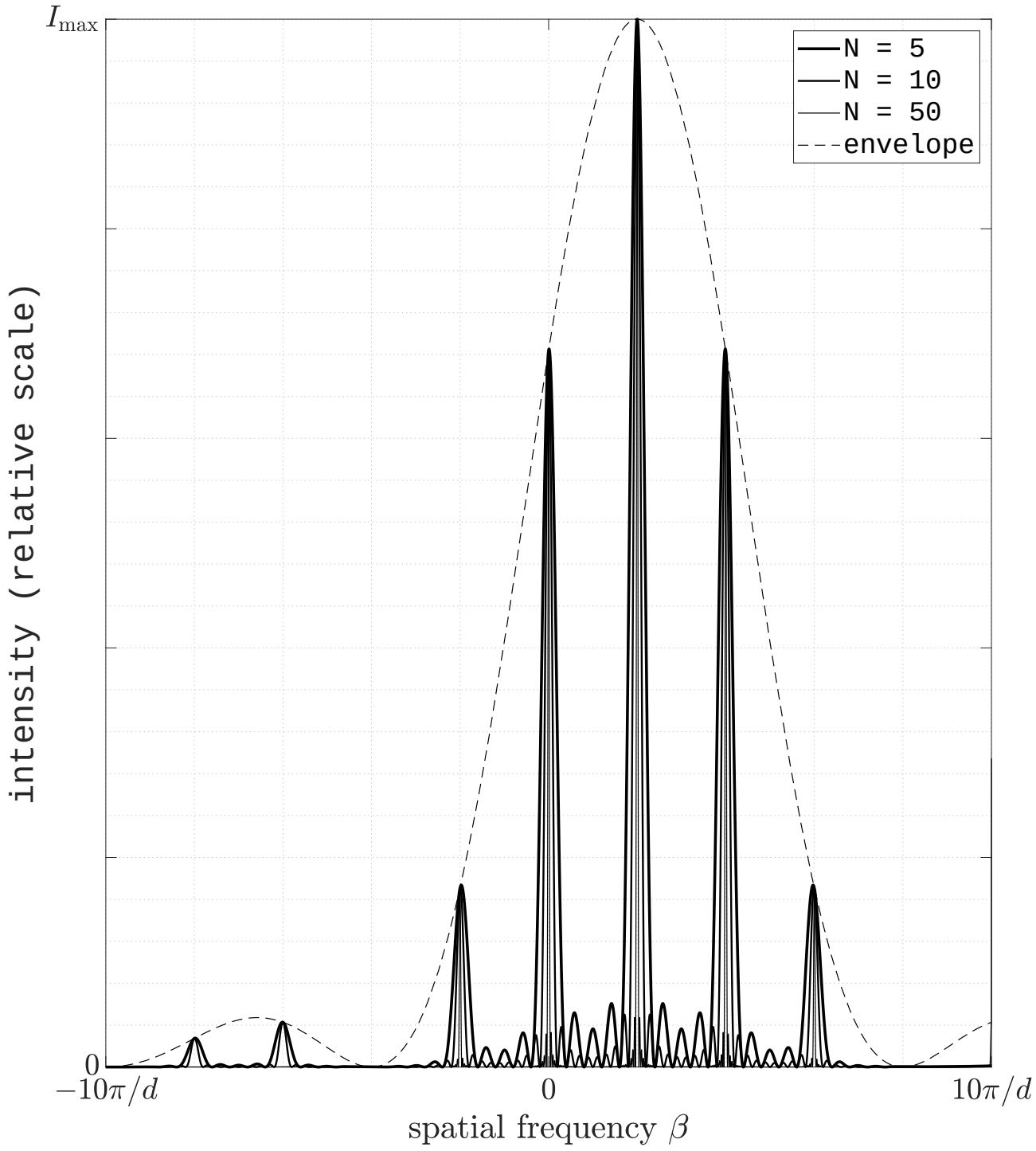

Figure 25.3: Intensity of a blazed reflection grating to first order.

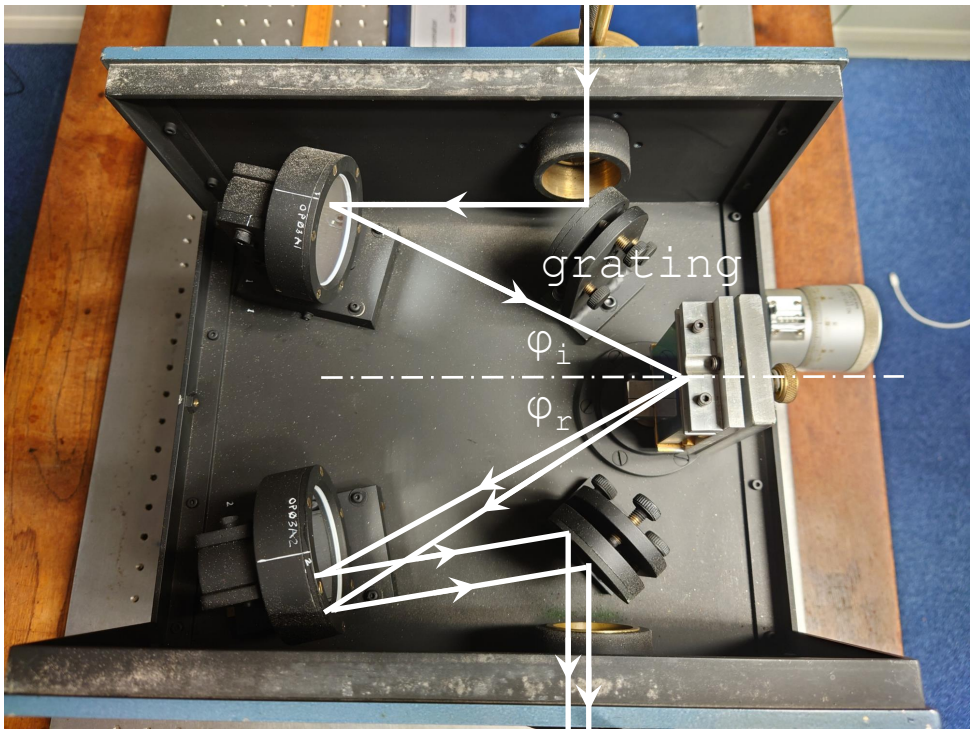

Figure 26.1: A grating spectrometer in the Czerny-Turner configuration.

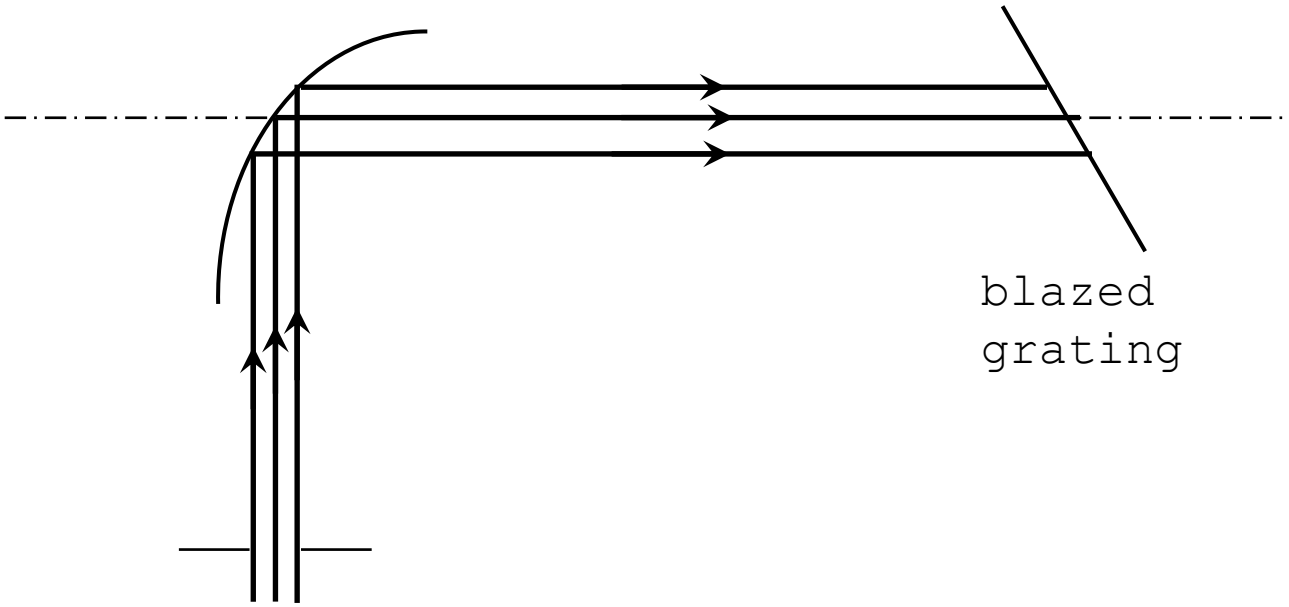

Figure 26.2: A grating spectrometer in the Littrow configuration.

into mirrors that does the exact same jobs. This is called the **Czerny-Turner configuration**, illustrated in figure 26.1. The rays on the figure 26.1 is composed of two different wavelengths, and they are clearly separated from the grating and emerges at two different angles, which is then focussed onto two different distances by the focussing mirror.

However, we shall note that the if we replace a single thin lens by, for example, a good camera lens, then the designer of the lens has enough pieces of glass to construct a lens with good image quality with the effect of the aberrations reduced, and therefore we do see the use of lenses in more sophisticated optical interferometers.

Littrow Configuration

Alternatively we are able to blaze the grating such that the first order diffracted beam of one selected wavelength is turned back towards the light source. This is called the **Littrow** configuration, and it is used to select light of a given frequency. This setup is shown in figure 26.2. Practically it can be used as a feedback setup, where the light comes out of a laser beam goes back towards the direction of the laser beam, and therefore acts as a frequency selector.

Summary

1. We can use mirrors to collimate and focus the light ray that enters and exits a reflection grating, which gives a spectrometer in the Czerny-Turner configuration. The two wavelengths are then focussed to different distances on the eyepiece after passing through the focussing lens, and therefore separated.

2. We can blaze the diffraction grating to first order such that the grating turns back the light in the angle which the light is incident in. This is called the Littrow configuration, which is used to select the frequency that we are looking for.

5 Imaging

§27. Scalar Amplitudes on the Focal and Image Planes of the Thin Lens

Revisiting the Single Thin Lens

This chapter investigates how to process coherent light beams to form an image. For example, the light beams we are referring here could be light coming from a diffraction mask. The simplest way of imaging, as we have learnt in previous chapters, is to image by a thin lens. So far we have already learnt two simple methods of analysing how light beams with scalar amplitude $u_o(x_o)$ at a distance u away from a thin lens with a focal length f transforms after passing through a lens. These are listed below. For this simple discussion let us assume that the light from the object is coherent and collimated, and the lens is infinitely big (we shall discuss how to relax this final assumption later).

- We may use lens maker's equation from geometric optics. From this we know that there is an image formed in the image plane at a distance v from the lens, where

$$\frac{1}{u}+\frac{1}{v} \;=\; \frac{1}{f}, \tag{27.1}$$

  and by the theory of similar triangles,

$$u_i(x_i) \;=\; u_o(-ux_i/v), \tag{27.2}$$

  that is, the scalar amplitude of the light on the image plane is a scaled version of the input.

- Previously we have suggested that a lens changes angles to distances. Assuming that the light from the source $u_o(x_o)$ is well collimated, different spatial frequencies β then travel towards different angles, which then turns into different positions on the focal plane of the lens where the angle θ and the position x_f is linked by $\tan\theta \;=\; x_f/f$. Formulating this mathematically, using the approximation $\sin\theta \;=\; \tan\theta$ and therefore $\beta \;=\; k\sin\theta \;=\; kx_f/f$, we have

$$U_i(\beta) \;=\; U_i(kx_f/f) \;=\; u_f(x_f), \tag{27.3}$$

  or, since $U_i(\beta)$ is the Fourier transform of $u_i(x_i)$,

$$u_f(x_f) \;=\; \mathscr{T}_F[u_o(x_o)](kx_f/f). \tag{27.4}$$

The results of these two approaches are summarised in figure 27.1.

Now consider the following experiment. If we wiggle the object up and down, according to equation 20.5, then this leads to a phase change in the Fourier transformed scalar amplitude. Since the intensity is only dependent on the modulus of the scalar amplitude, we should see no change on the intensity distribution on the focal plane. However of course we do see a change on the image plane; yet in some sense, the light on the image plane is propagated by light from the focal plane. How come that the same intensity distribution of light gives rise to two different sets of intensity distribution of light on the image plane? We must be missing something and we shall attempt to solve this problem next.

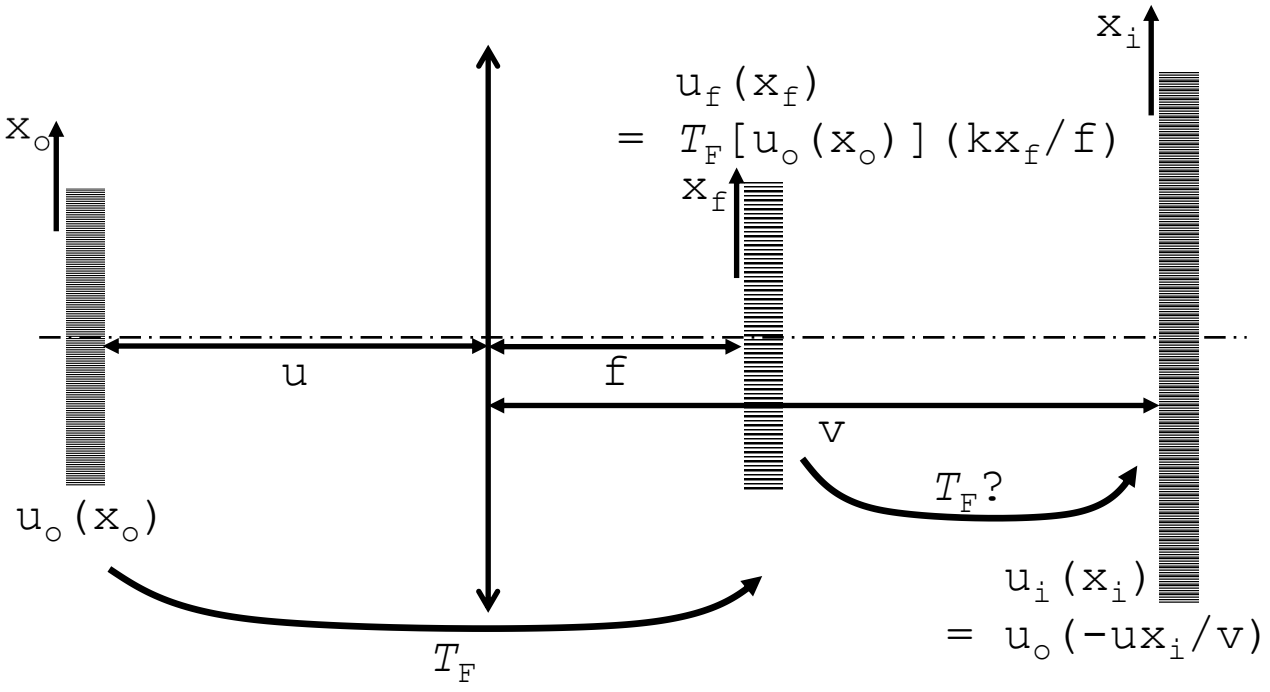

Figure 27.1: Imaging a coherent beam of light with scalar amplitude $u_o(x_o)$.

Series of Fourier Transforms

The only solution for the above problem is that, to travel from the focal plane to the image plane, we must have another operation that takes into account of the phases of the scalar amplitude of the intermediate focal plane. The simplest guess of what is happening is that light Fourier transforms itself from the focal plane to the image plane. Justifying this mathematically is lengthy and is discussed in §28, however we could justify this by a simple thought experiment.

Consider figure 27.2 where we have a pair of thin concave and convex lenses which have shapes complementary to one another (so light through this combination will neither be focussed nor be de-focussed) sandwiching the focal plane. Since the lenses are thin, the scalar amplitude in the image plane $u_i(x_i)$ is unaltered by this insertion of the pair of lenses. Since the focal plane and the concave lens are back-to-back, light does not have physical space to be de-focussed by the concave lens and therefore will also be un-altered. Then, it is clear that when light travels from focal plane to the image plane, it has to pass through this convex lens, justifying this Fourier transform between the focal and image planes.

Now we have some preliminary justification, we shall put this theory that the scalar amplitude at the image and focal planes are linked by a Fourier transform on firm mathematical footing. We shall do that in §28.

Summary

1. In this chapter we investigate into how coherent light can be processed, and the simplest method is to use a single thin lens. After some analysis we find that the scalar amplitude of light on the focal plane is the Fourier transform of the input and that on the image plane is a scaled version of the input.

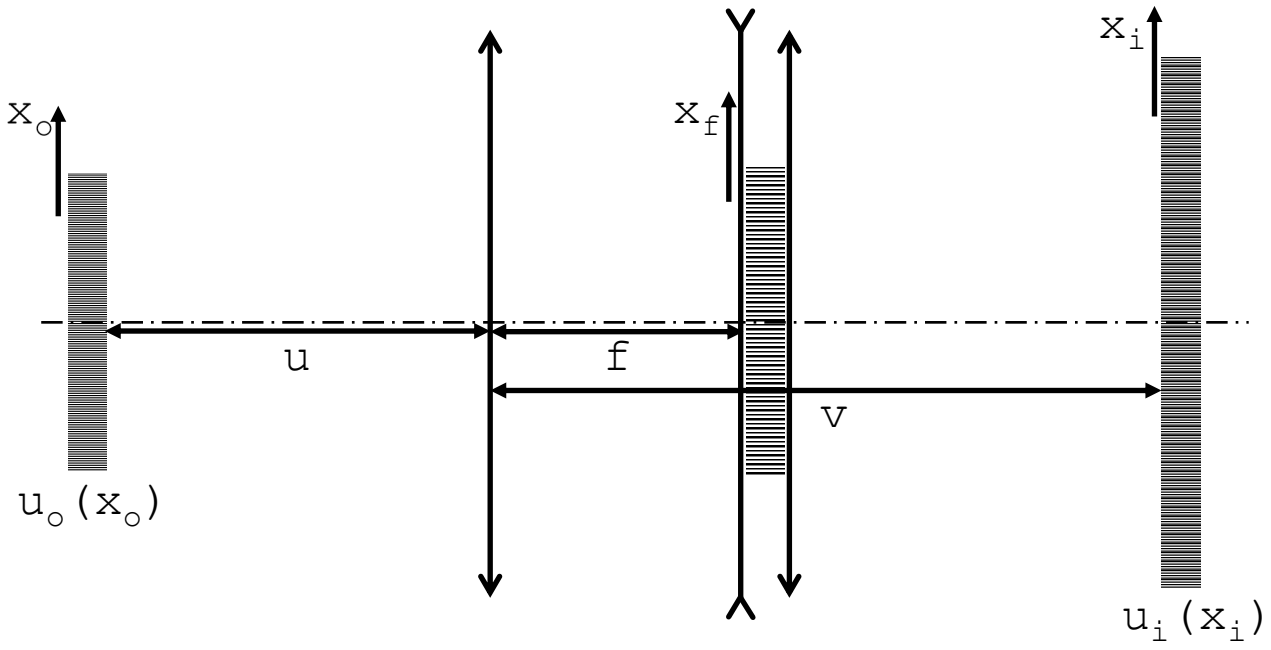

Figure 27.2: Analysing the problem by placing virtual lenses sandwiching the focal plane.

2. By the argument of the insertion of a pair of thin concave and convex lenses, we conclude that the scalar amplitude on the image plane is a Fourier transform of that in the focal plane.

§28. Off-Axis Image Formation by a Thin Lens

Contributions to the Optical Path Length

In order to fully understand what is happening mathematically, including the phase of the light, we must take into account of a full mathematical analysis of a light beam from *any* point from the input, that is, a general off-axis point. Figure 28.1 is constructed for us to analyse this problem and the optical path length that we are interested in is the light from X_o to X_i.

We now have two tasks in hand, that is, to consider how light goes through the lens and reach the focal plane, and how light goes from the focal plane to a point on the image plane. Let us tackle these one by one.

- The difficulty in finding the ray that reaches the focal plane is that we are unsure which way light takes from point X_o. The method that we shall take is to decompose the spherical wavefront originating from X_o into a superposition of many plane wavefronts, including, for example, the wavefront X_oA on figure 28.1. We may consider any ray on the plane wavefront, as light from the entire wavefront has the same phase. Of course, the easiest would be the chief ray since the chief ray passes straight through the lens without bending. This gives us two contributions to the optical path length, that are (the distances refers to the distances labelled on figure 28.1)

$$\Delta_1 \;=\; b\cos\theta, \qquad \Delta_2 \;=\; f/\cos\theta. \tag{28.1}$$

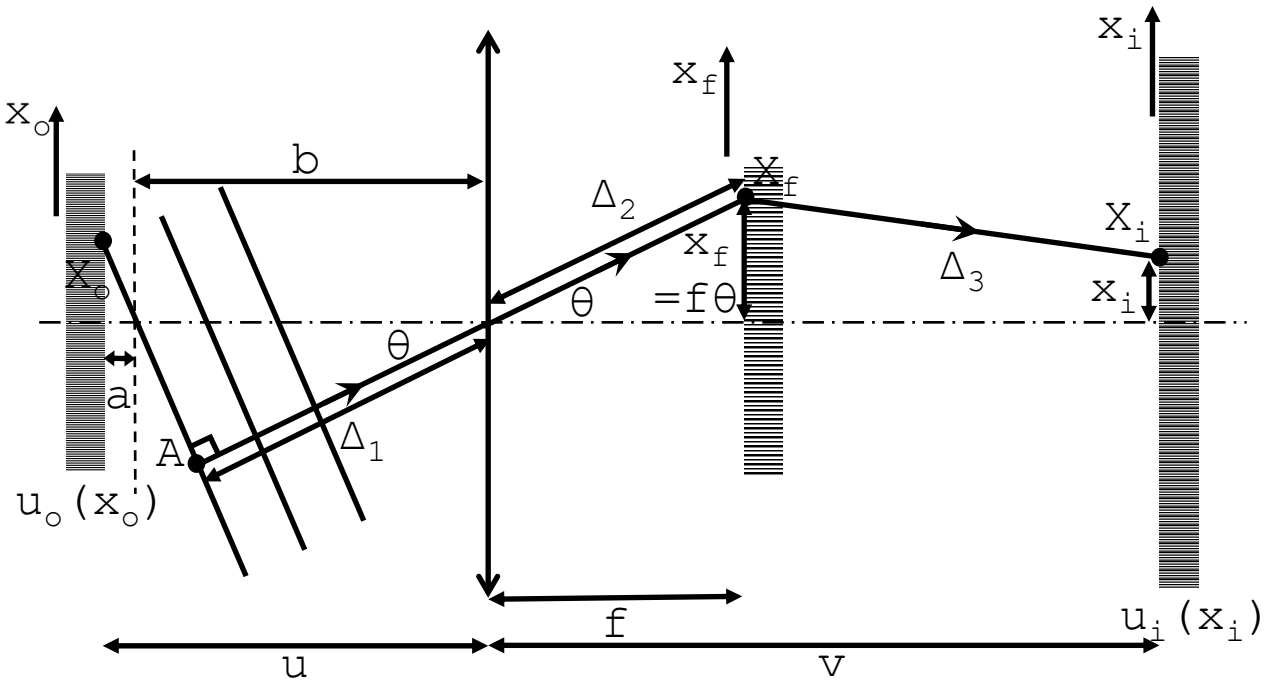

Figure 28.1: An diagram that assists us to set up the problem for an analysis of the off-axis rays towards the image plane.

- Now all the plane waves arrives at X_f, we would like to guide these waves to the image plane. The problem is that, the ray that we were considering previously, AX_f, is only one of many rays that arrives at X_f, which, since there are no real obstructions or lenses on the focal plane, all of these rays travel straight on. Therefore we can only focus on a general ray that travels to X_i, and set x_i as a free parameter. The contribution of this leg to the total optical path length is

$$\Delta_3 \;=\; \sqrt{(v-f)^2+(x_i-x_f)^2} \;=\; (v-f)\left[1+\frac{(x_i-x_f)^2}{(v-f)^2}\right]^{\frac{1}{2}} \;=\; v-f+\frac{(x_i-x_f)^2}{2(v-f)}, \tag{28.2}$$

  where we have binomially expanded the square root under the usual approximation $(x_f - x_i) \ll (v-f)$.

Now, assuming that θ is small, we have the usual small angle approximations, along with $a = x_o\theta$ and therefore $b = u - x_o\theta$, and hence the total optical path length from X_o to X_i is

$$\begin{aligned}
\text{OPL} &= \Delta_1+\Delta_2+\Delta_3 \\
&= \left[(u-x_o\theta)\left(1-\frac{\theta^2}{2}\right)\right]+\left[f\left(1+\frac{\theta^2}{2}\right)\right]+\left[v-f+\frac{(x_i-x_f)^2}{2(v-f)}\right] \\
&= \left[\left(u-x_o\frac{x_f}{f}\right)\left(1-\frac{x_f^2}{2f}\right)\right]+\left[f\left(1+\frac{x_f^2}{2f^2}\right)\right]+\left[v-f+\frac{(x_i-x_f)^2}{2(v-f)}\right].
\end{aligned} \tag{28.3}$$

Note that the un-fixed parameters in this problem are x_f and x_i, therefore everything that is *only* dependent on x_o, u, v, and f are a common optical path for all rays. Removing all such

terms and only keeping terms up to x_*^2 (where $* = i, f$), we have

$$\text{OPL} = \frac{1}{2}\left(\frac{x_f}{f}\right)^2 (f-u) - x_o\left(\frac{x_f}{f}\right) + \frac{(x_i - x_f)^2}{2(v-f)}. \tag{28.4}$$

So far we have obtained the total optical path, and naturally we would like to write down the integral for all points on the object plane. However we note that from this expression of the optical path length, we are already able to recreate some of the results we have obtained using geometric optics. To do that, we use Fermat's principle to find the pathways where the optical path length OPL via finding the minimum of OPL with respect to x_f i. e. we differentiate OPL with respect to x_f and set this to 0, giving

$$\frac{\mathrm{dOPL}}{\mathrm{d}x_f} = \frac{x_f(f-u)}{f^2} - \frac{x_o}{f} - \frac{x_i - x_f}{v-f} = \left(\frac{1}{v-f} - \frac{u-f}{f^2}\right)x_f - \frac{x_o}{f} - \frac{x_i}{v-f} \stackrel{!}{=} 0, \tag{28.5}$$

or,

$$Ax_f - B \stackrel{!}{=} 0, \qquad A = \frac{1}{v-f} - \frac{u-f}{f^2}, \qquad B = \frac{x_o}{f} + \frac{x_i}{v-f}. \tag{28.6}$$

Now note that the focal plane, the image plane, and the input are related by the lens maker's equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \quad \Rightarrow \quad (u-f)(v-f) = f^2 \quad \Rightarrow \quad A = 0 \tag{28.7}$$

therefore for equation 28.6 to hold, we must have

$$B = 0 \quad \Rightarrow \quad x_i = -vx_o/u, \tag{28.8}$$

the geometrical link between x_i and x_o which we already know from the theory of geometric optics.

From the Input to the Focal Plane

Now we have already found the optical path length, we have everything we need to figure out what is happening between these three planes. To go from the the input to the focal plane, we have an optical path length (again eliminating high order terms)

$$\text{OPL}_{of} = \Delta_1 + \Delta_2 = \left[(u - x_o\theta)\left(1 - \frac{\theta^2}{2}\right)\right] + \left[f\left(1 + \frac{\theta^2}{2}\right)\right] = S + F + \Delta_f(\theta), \tag{28.9}$$

where

$$S = u + f, \qquad F = -x_o\theta, \qquad \Delta_f(\theta) = \frac{\theta^2}{2}(f-u). \tag{28.10}$$

Here, S is the contribution that is the same for all beams of light that propagates through the system which we drop as it contributes to an overall phase; F is the term that is linear on x_o therefore corresponds to a Fraunhofer integral; and $\Delta_f(\theta)$ is independent of x_o but dependent on θ, which we may take outside the integral. Then, using the relation $\beta = -k\theta$ (now changing the definition of β to get the usual Fraunhofer integral; recall that we have discussed previously that this sign is completely arbitrary as we effectively looking at the modulus of the scalar amplitude) for small values of θ, we have

$$U_f(\beta) = \mathrm{e}^{\mathrm{i}k\Delta_f(\beta/k)} \int_{-\infty}^{\infty} \mathrm{d}x_o\, u_o(x_o)\mathrm{e}^{\mathrm{i}\beta x} = \mathrm{e}^{\mathrm{i}\frac{k}{2}\left(\frac{x_f}{f}\right)^2(f-u)}\,\mathscr{T}_F[u_o(x_o)](-kx_f/f) \tag{28.11}$$

retrieving equation 27.4 but also retaining an extra phase factor. This extra phase tells us that light arriving at the focal plane does not have the same phase, or that, we have a **curved wavefront** arriving at the focal plane. Wiggling the object up and down will not change the intensity pattern in the focal plane, however it will affect the *shape* of the wavefront arriving at the focal plane, explaining why light hitting the image plane will then change in turn, solving the problem introduced in the previous section.

From the Focal Plane to the Image Plane

Now let us travel from the focal plane to the image plane, i. e. do the integral

$$u_i(x_i) \;=\; \int_{-\infty}^{\infty} \mathrm{d}x_f \, u_f(x_f) \mathrm{e}^{\mathrm{i}k\Delta_3}. \tag{28.12}$$

Now to unpack this, we insert equation 28.11, the relation $\beta = -k\theta = -kx_f/f$, and the expression for Δ_3, that is, equation 28.2. Putting these together, we have

$$u_i(x_i) \;=\; \mathrm{e}^{\mathrm{i}\frac{kx_i^2}{2(v-f)}} \int_{-\infty}^{\infty} \mathrm{d}x_f \, \mathrm{e}^{\mathrm{i}\frac{1}{2}kAx_f^2} \mathrm{e}^{-\mathrm{i}k\frac{x_i x_f}{v-f}} \mathscr{T}_f[u_o(x_o)](-kx_f/f) \tag{28.13}$$

where A is defined in equation 28.6. Noting that we have already shown that A must be 0 by the lens maker's equation, it is therefore possible for us to deduce

$$u_i(x_i) \;=\; u_o(-ux_i/v) \;\stackrel{!}{=}\; \int_{-\infty}^{\infty} \mathrm{d}x_f \, \mathrm{e}^{\mathrm{i}\beta' x_f} \mathscr{T}_F[u_o(x_o)](-kx_f/f) \;=\; \mathscr{T}'_F[\mathscr{T}_F[u_o]](\beta'). \tag{28.14}$$

where we have introduced $\beta' = -kx_i/(v-f)$ and dropped an overall phase. Now it is clear that $u_f(x_f)$ and $u_i(x_i)$ are also linked by a Fourier transform.

Summary

1. By writing down the total optical path length from the input to the image plane and applying Fermat's principle explicitly, we obtain the relation

$$x_i \;=\; -vx_o/u. \tag{28.15}$$

2. By considering the integral from the input plane to the focal plane of the length, we find that the light arriving at the focal plane forms a curved wavefront.

3. Taking into account of this curved wavefront and considering an integral from the focal plane to the image plane, we retrieve the fact that the light is also Fourier transformed travelling from the focal plane to the image plane.

§29. Abbé Theory of Imaging

Object with a Periodic Structure

Modern imaging theory, developed by Abbé and Zeiss, starts with analysing how to image input light with a periodic structure. (Of course, the next step would be to create any objects as considering these objects with periodic structures as a Fourier basis.) Let us consider an input beam with period d

$$u_o(x_o) \;=\; u_0 + u_1 \cos\left(\frac{2\pi}{d} x_o\right), \tag{29.1}$$

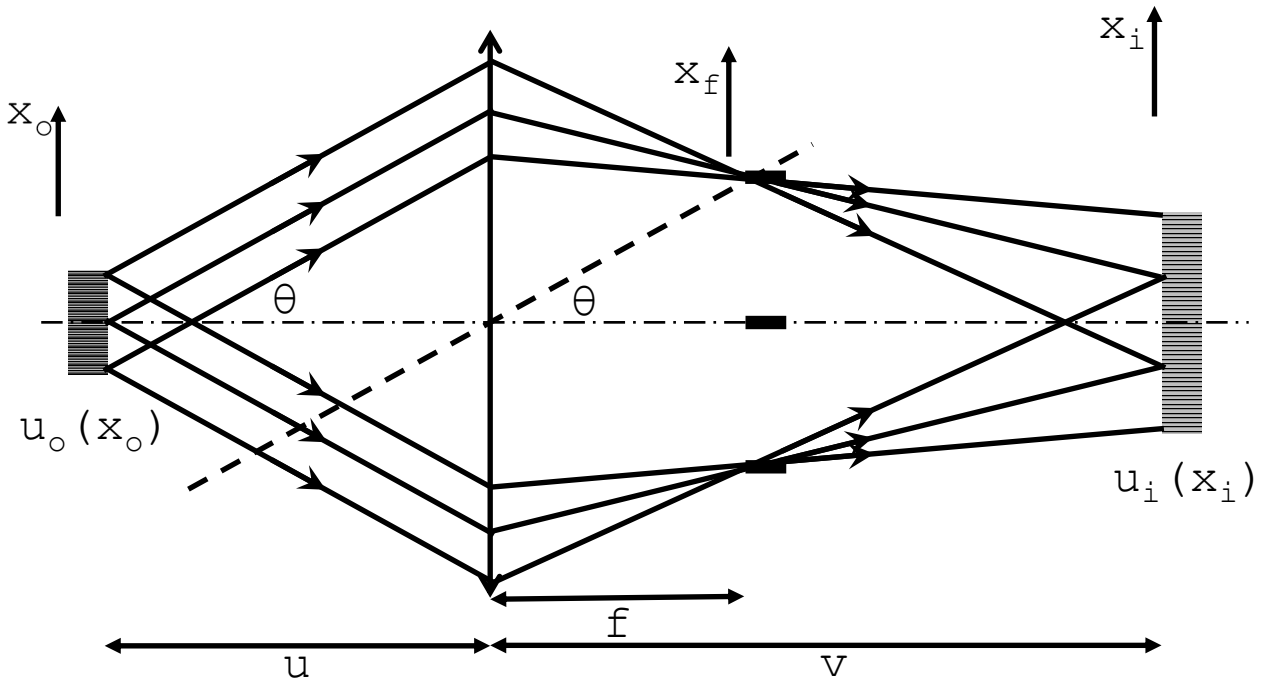

Figure 29.1: An object with a periodic structure forms bright spots at the focal plane.

where clearly the spatial frequency is given as $\beta = 2\pi/d$. The Fourier transform of this input beam is rather straightforward to write down, it is simply

$$U_f(\beta) = u_o\delta(\beta) + \frac{u_1}{2}\delta\left(\beta - \frac{2\pi}{d}\right) + \frac{u_1}{2}\delta\left(\beta + \frac{2\pi}{d}\right). \tag{29.2}$$

Now, since a lens changes angles to distances, using the relation

$$\beta = k\theta = \frac{2\pi\theta}{\lambda} = \frac{2\pi x_f}{\lambda f}, \tag{29.3}$$

we have

$$u_f(x_f) = u_o\delta\left(\frac{2\pi x_f}{\lambda f}\right) + \frac{u_1}{2}\delta\left(\frac{2\pi x_f}{\lambda f} - \frac{2\pi}{d}\right) + \frac{u_1}{2}\delta\left(\frac{2\pi x_f}{\lambda f} + \frac{2\pi}{d}\right), \tag{29.4}$$

i. e. we expect bright spots at $x_f = 0$ and $\pm\lambda f/d$. This is illustrated in figure 29.1.

Now we introduce the important thought that runs through this chapter, that is, *if we mask out these bright spots in the focal plane, then this spatial frequency will be missing on the image plane*. Sometimes these spots are masked out intentionally — that is how we process images; and sometimes there is a physical limitation of the apparatus that causes the spots to be masked out, for example the lens is not big enough. We shall treat the latter case now and treat the former case for the chapters that follows.

Finite Lens Size

The first object that modern imaging theory attempts to deal with is a microscope, where the radius of the lens is given as $\mathscr{R}$ and we are trying to resolve a periodic detail with a regular spacing d. Note that this finite radius $\mathscr{R}$ limits the angle θ i. e. $\theta < \mathscr{R}/u$. However, by

previous analysis, the bright spot at the focal plane is located at an angle

$$\frac{2\pi\theta}{\lambda} - \frac{2\pi}{d} \;=\; 0 \qquad \Rightarrow \qquad \theta \;=\; \frac{\lambda}{d}. \tag{29.5}$$

Therefore, equating these expressions, we have the smallest detail resolved being

$$d \;>\; \frac{\lambda}{(\mathscr{R}/u)} \;=\; \frac{\lambda_0}{(n\mathscr{R}/u)} \;=\; \frac{\lambda_0}{\mathrm{NA}}, \tag{29.6}$$

where λ_0 is the wavelength of the light in a vacuum, n is the refractive index of the medium, and $\mathrm{NA} \;=\; n\mathscr{R}/u$ is called the **numerical aperture** (recall that we have came across a similar, but slightly different definition in §2, in the context of an optical fibre). In real life the basic lens can be improved by adding a condenser lens, improving the resolution by about a factor 2, giving **Abbé's equation** for the smallest structure resolved

$$d \;=\; \frac{\lambda_0}{2\mathrm{NA}} \;=\; \frac{\lambda_0}{2n\sin\theta_{\max}}. \tag{29.7}$$

When we actually buy lenses, it is impossible to buy a lens with any numerical aperture, as this parameter is dependent on the focussing power of the lens. The largest numerical aperture of a lens that can be engineered is about 1.4 — which actually is rather amazing as the focal lens of the mirror is already shorter than the radius, giving a very large acceptance angle. Therefore, the best possible diffraction-limited resolution for any lens system is given as

$$d \;\approx\; \lambda/2.8. \tag{29.8}$$

A remark here is that Abbé theory is originally for microscopes and therefore sometimes one might see an alternative definition for the numerical aperture

$$\mathrm{NA} \;=\; \mathscr{R}/f \tag{29.9}$$

as the object is usually placed very near the focal length for a microscope.

Note that there is also a very common parameter used in photography, where the common convention for the ratio between $\mathscr{R}$ and f is the ƒ-number, defined as

$$ƒ/\# \;=\; \frac{f}{2\mathscr{R}} \;=\; \frac{1}{2\mathrm{NA}}, \tag{29.10}$$

where the latter equality holds in the microscope case where $u \;\approx\; f$. The upper bound on the numerical aperture gives a lower bound on the ƒ-number: the lower bound of the ƒ-number on any camera lens one can buy is 1.0, unless one would like to spend several thousands of pounds on it; then one can reduce it further to 0.8, but that is it — the absolute minimum. Recall that increasing the ƒ-number masks out higher frequency components on the focal plane, which then removes the higher frequencies. The effect of this is usually a more blurry background as the higher frequencies corresponds to the "details" of the input, which we shall discuss in the sections that follows.

Summary

1. An input with a periodic structure gives bright spots on the focal plane, i. e. the scalar amplitude on the focal plane is a sum of Dirac-deltas; which then interferes to form the image. Masking out these bright spots in the focal plane means that the corresponding spatial frequencies will be missing on the image plane.

2. A finite lens size automatically means that some spots on the focal plane will be missing and therefore this gives a limit on the highest spatial frequency resolvable. The corresponding parameter of the lens are the numerical aperture and the ƒ-number, defined as

$$\text{NA} \;=\; n\mathscr{R}/u \qquad \text{and} \qquad \text{ƒ}/\# \;=\; f/(2\mathscr{R}) \tag{29.11}$$

respectively. The largest possible numerical aperture is about 1.4 and the smallest possible ƒ-number is about 0.8, corresponding to the best possible diffraction-limited resolution

$$d \;\approx\; \lambda/2.8. \tag{29.12}$$

§30. Wavefront-Preserving Imaging

Shape of the Wavefront

Back in equation 28.10 we have suggested that the contribution

$$\Delta_f(\theta) \;=\; \frac{\theta^2}{2}(f-u) \tag{30.1}$$

to the total optical path gives a curved wavefront. This means that the wavefront at the image plane will also be curved, i. e. we are not preserving the wavefront that arrives at the image plane: we have distorted its shape. In this section we would like to suggest a method to give an image with a different size and have flat wavefronts at the output at the same time. But first let us try to understand the shape of this curved wavefront.

The first observation is that Δ_f is negative since $u < f$ for any real image formed. Therefore, equation 28.10 tells us that an off-axis point will reach the focal plane lagging the on-axis point, and therefore it needs to travel a further optical path length $-\Delta_f$ further after passing through the focal plane to make up this optical path length deficit. With this in mind, a rough sketch of the curved wavefront is demonstrated in figure 30.1.

The hypothesis that we shall verify is that the curved wavefront can be approximated as spherical for small angles towards the on-axis point on the image plane. If such a hypothesis is correct, then we must also have

$$-\Delta_f \;\stackrel{?}{=}\; (v-f)(\sec\varphi - 1) \;=\; \frac{\varphi^2}{2}(v-f) \;=\; \frac{x_f}{2(v-f)}, \tag{30.2}$$

where $\varphi = x_f/(v-f)$ is defined in figure 30.1, and also $\theta = x_f/f$ for an on-axis point. Indeed, with lens maker's equation

$$\frac{1}{u}+\frac{1}{v} \;=\; \frac{1}{f} \quad\Rightarrow\quad u-f \;=\; \frac{f^2}{v-f}, \tag{30.3}$$

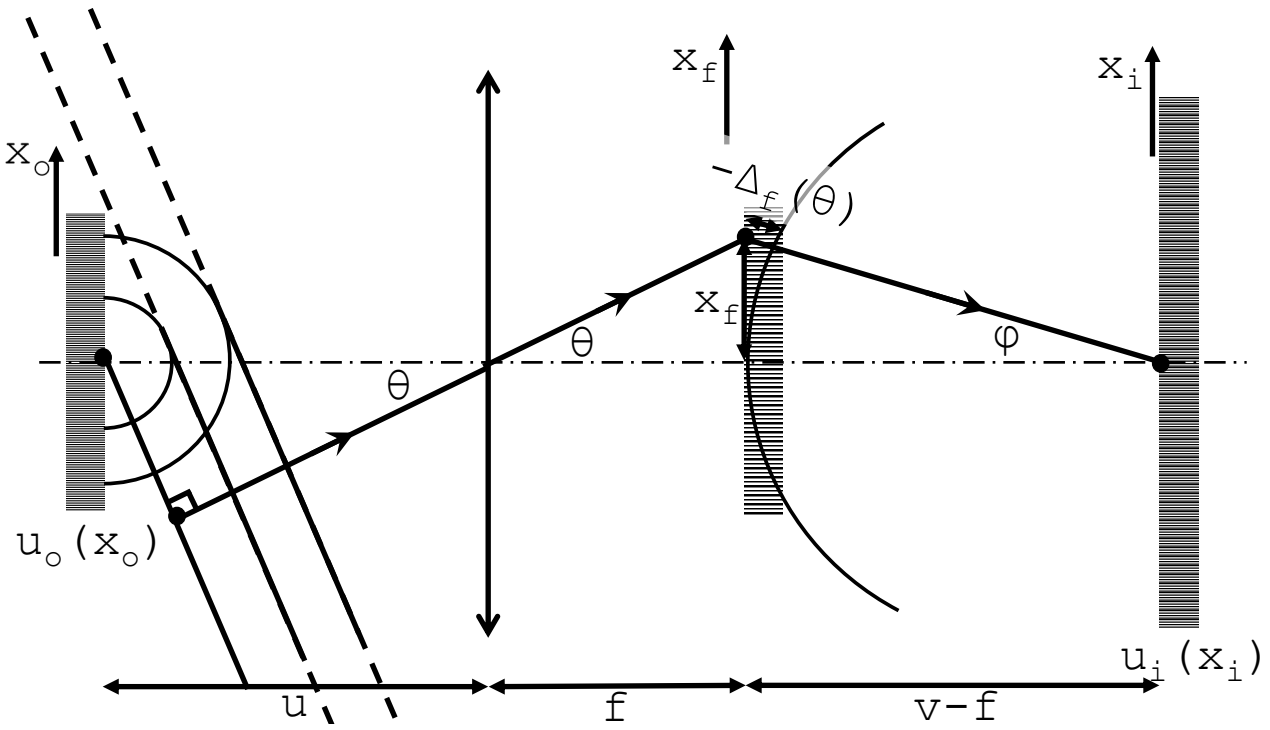

Figure 30.1: An illustration of the curved wavefront.

we have

$$\Delta_f \;=\; \frac{x_f}{2f^2}(f-u) \;=\; \frac{x_f}{2(f-v)}, \tag{30.4}$$

hence showing that our hypothesis is correct.

Wavefront Preservation

Now that we understand what the shape of the wavefront is, let us try to correct this, and this is done by adding another lens i. e. use the other lens to correct for these spherical distortions reaching the image plane.

Figure 30.2 shows the position of the setup: we have the input and the image at one focal length away from the two lenses with focal lengths f_1 and f_2, and space the two lenses at a distance $f_1 + f_2$ apart from each other. If we have plane wavefronts arriving at the first lens, then the extra phase difference will then be compensated by the addition of the second lens. Quantitatively, on the focal plane of the first lens (the lens with focal length f_1), we have

$$U_f(\beta_1) \;=\; \mathscr{T}_F[u_o(x_o)](\beta_1), \tag{30.5}$$

therefore, using $\beta_1 \;=\; kx_f/f_1$,

$$u_f(x_f) \;=\; \mathscr{T}_F[u_o(x_o)](kx_f/f_1); \tag{30.6}$$

this is then immediately Fourier transformed passing the second lens

$$U_i(\beta_2) \;=\; \mathscr{T}_F[u_f(x_f)](\beta_2), \tag{30.7}$$

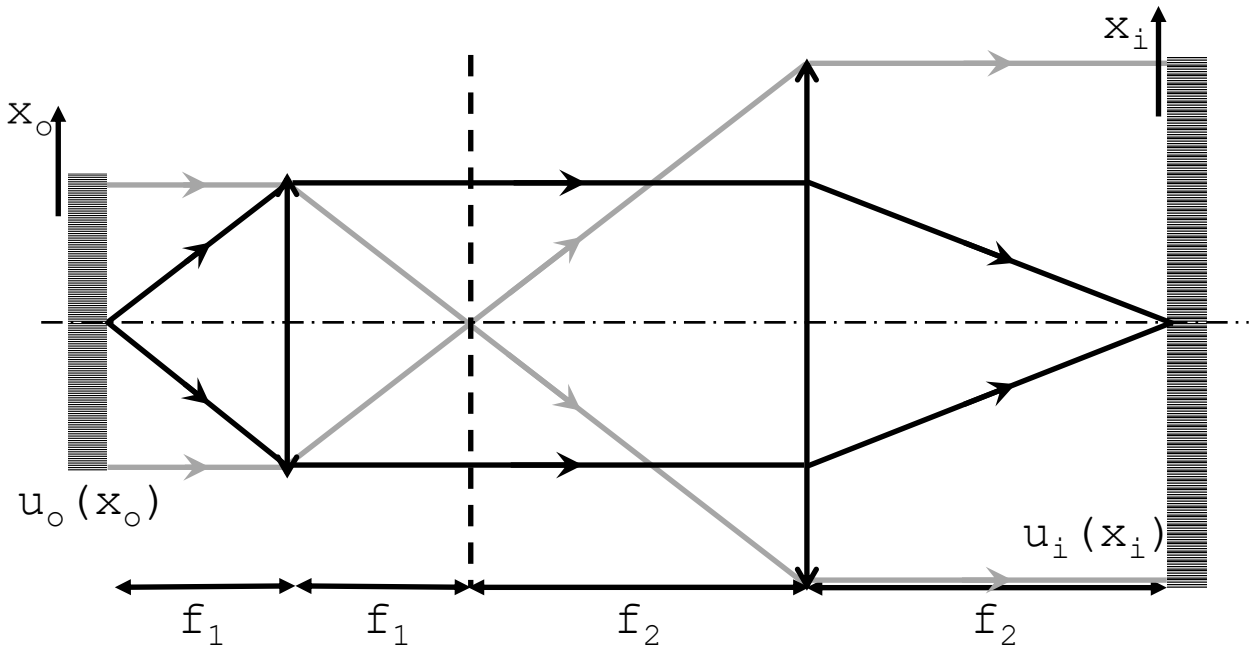

Figure 30.2: A setup to preserve the wavefronts arriving at the image plane.

which, then using $\beta_2 = kx_i/f_2$, gives

$$\begin{aligned} u_i(x_i) &= \mathscr{T}_F[\mathscr{T}_F[u_o(x_o)](kx_f/f_1)](kx_i/f_2) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x_f \int_{-\infty}^{\infty} \mathrm{d}x_o\, u_o(x_o) \mathrm{e}^{\mathrm{i}\frac{kx_f}{f_1}x_o} \mathrm{e}^{\mathrm{i}\frac{kx_i}{f_2}x_f} \\ &= \int_{-\infty}^{\infty} \mathrm{d}x_o\, J(x_o), \end{aligned} \tag{30.8}$$

where the integral $J(x_o)$ is defined as

$$J(x_o) = u_o(x_o) \int_{-\infty}^{\infty} \mathrm{d}x_f\, \mathrm{e}^{\mathrm{i}\frac{k}{f_1}x_o x_f} \mathrm{e}^{\mathrm{i}\frac{k}{f_1}\times\frac{f_1}{f_2}x_i x_f}, \tag{30.9}$$

a Fourier Transform of the exponential function $\mathrm{e}^{\mathrm{i}\frac{k}{f_1}\times\frac{f_1}{f_2}x_i x_f}$. However, since the this exponential can be obtained from the inverse Fourier transform of a Dirac-delta:

$$\frac{2\pi f_1}{k}\mathscr{T}_F^{-1}\left[\delta\left(x_o + \frac{f_1}{f_2}x_i\right)\right] = \frac{f_1}{k}\int_{-\infty}^{\infty} \mathrm{d}\left(\frac{k}{f_1}x_o\right) \mathrm{e}^{-\mathrm{i}\frac{k}{f_1}x_o x_f} \delta\left(x_o + \frac{f_1}{f_2}x_i\right) = \mathrm{e}^{\mathrm{i}\frac{k}{f_1}\times\frac{f_1}{f_2}x_i x_f}, \tag{30.10}$$

we must have

$$J(x_o) = u_o(x_o)\mathscr{T}_F\left[\mathrm{e}^{\mathrm{i}\frac{k}{f_1}\times\frac{f_1}{f_2}x_i x_f}\right] = \frac{k}{2\pi f_1}u_o(x_o)\delta\left(x_o + \frac{f_1}{f_2}x_i\right). \tag{30.11}$$

Hence,

$$u_i(x_i) = \frac{k}{2\pi f_1}\int_{-\infty}^{\infty} \mathrm{d}x_o\, u_o(x_o)\delta\left(x_o + \frac{f_1}{f_2}x_i\right) \propto u_o\left(-\frac{f_1}{f_2}x_i\right). \tag{30.12}$$

This way, we have achieved the same goal as using a single lens — giving an image of a different size, and also we have a flat wavefront as an output, i. e. we have successfully imaged the input

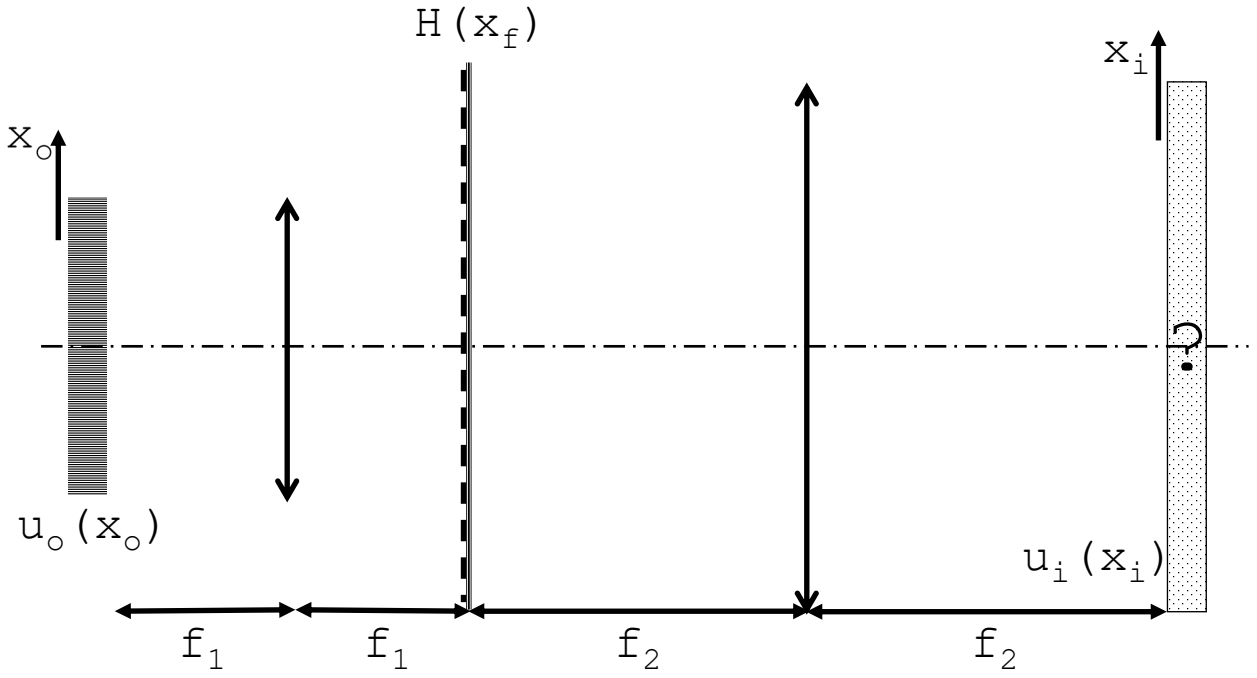

Figure 31.1: Filtering light on the focal plane by a filtering function $H(x_f)$.

beam while preserving the shape of the wavefront.

Summary

1. The curved wavefront upon arrival on the focal plane of a single lens can be seen as approximately circular.

2. Imaging with two lenses with focal lengths f_1 and f_2 at a distance $f_1 + f_2$ apart with the input light at a distance f_1 away from the first lens gives an image at a distance f_2 away from the second lens, enlarged with a scale factor f_2/f_1 and inverted, also preserving the shape of the wavefront.

§31. Filtering

Optical Point-Spread Function

At the start of this discussion, we have suggested that we may mask out certain points on the focal plane to obtain images with different spatial frequencies in the object plane. Now we have understood how to preserve our wavefront, let us do this with our new wavefront-preserving imaging system.

For example, let us consider figure 31.1 where the focal plane is masked out using a filtering function $H(x_f)$, defined as the fraction of the scalar amplitude of the light let through by the filter at x_f. This can either describe a filter placed by hand, or a filter due to the setup, for example, the finite size of the second lens. Then, using the convolution theorem backwards, we have

$$u_i(x_i) \;\propto\; \mathscr{T}_F[u_f(x_f) \times H(x_f)] \;=\; \mathscr{T}_F[u_f] \otimes h(x_i), \tag{31.1}$$

where

$$h(x_i) \;=\; \mathscr{T}_F[H(x_f)] \tag{31.2}$$

is called the **optical point-spread function**.

Coherent and Incoherent Illumination

Using the relation

$$\mathscr{T}_F[u_f] \;\propto\; u_o\!\left(-\frac{f_1}{f_2}x_i\right) \tag{31.3}$$

and the definition of the convolution, we have

$$u_i(x_i) \;\propto\; \int_{-\infty}^{\infty} \mathrm{d}x_i'\, u_o\!\left(-\frac{f_1}{f_2}x_i'\right) h(x_i - x_i'). \tag{31.4}$$

Then, the intensity of the image as a function of x_i is

$$I_i(x_i) \;\propto\; \int_{-\infty}^{\infty} \mathrm{d}x_i' \left[u_o\!\left(-\frac{f_1}{f_2}x_i'\right)\right]^* [h(x_i - x_i')]^* \int_{-\infty}^{\infty} \mathrm{d}x_i''\, u_o\!\left(-\frac{f_1}{f_2}x_i''\right) h(x_i - x_i''). \tag{31.5}$$

Note that, although in this chapter our main focus is to discuss imaging coherent light, with the above equation it is rather simple to write down the equation for the incoherent case. The trick is to argue that the cross term (or, interference term) must vanish for incoherent illumination, and therefore the integrals must merge (i. e. we artificially add a Dirac-delta $\delta(x_i' - x_i'')$ in the expression for $I_i(x_i)$); thus using such an argument, we have

$$I_i(x_i) \;\propto\; \int_{-\infty}^{\infty} \mathrm{d}x_i' \left|u_o\!\left(-\frac{f_1}{f_2}x_i'\right)\right|^2 |h(x_i - x_i')|^2 \tag{31.6}$$

for incoherent illumination.

Let us write down our results again. For coherent illumination, we have

$$I_i(x_i) \;\propto\; \left|u_o\!\left(-\frac{f_1}{f_2}x_i\right) \otimes h_i(x_i)\right|^2; \tag{31.7}$$

and for incoherent illumination, we have

$$I_i(x_i) \;\propto\; I_o\!\left(-\frac{f_1}{f_2}x_i'\right) \otimes |h_i(x_i)|^2. \tag{31.8}$$

Alternatively, for the incoherent illumination case, we may use the convolution theorem to write down the expression for the intensity in the following form

$$I_i(x_i) \;\propto\; \mathscr{T}_F{}^{-1}\!\left[\mathscr{T}_F\!\left[I_o\!\left(-\frac{f_1}{f_2}x_i'\right)\right] \times K(\beta)\right], \qquad K(\beta) \;=\; \mathscr{T}_F\big[|h_i(x_i)|^2\big], \tag{31.9}$$

where $K(\beta)$ is called the **optical transfer function**. Recasting it into such a form separates clearly the contribution of the filter and that of the intensity caused by the input beam in an ideal system.

Summary

1. We may filter light on the focal plane of the wavefront-preserving imaging system with a filtering function $H(x_f)$. Then, the output on the image plane is $\mathscr{T}_F[u_f] \otimes h(x_i)$, where $h(x_i) = \mathscr{T}_F[H(x_f)]$ is the optical point-spread function.

2. For incoherent illumination, we have the intensity on the image plane as

$$I_i(x_i) \propto \mathscr{T}_F^{-1}\left[\mathscr{T}_F\left[I_o\left(-\frac{f_1}{f_2}x_i'\right)\right] \times K(\beta)\right], \qquad K(\beta) = \mathscr{T}_F\left[|h_i(x_i)|^2\right], \tag{31.10}$$

   where $K(\beta)$ is called the optical transfer function.

§32. Manipulating the Image at the Focal Plane

Shapes at the Focal and Image Planes

So far we have discussed how different spatial frequencies can be masked out, either by the physical limitations of the setup (i. e. the size of the lens being finite), or using a filter. Or, alternatively, we may choose to detect the image at the focal plane instead of the image plane, filter this intensity pattern *on the focal plane* on a computer, then Fourier transform on the computer to obtain an image, usually used in X-ray diffraction. The thing that we have missed out is how masking out these spatial frequencies actually alter the image. This is best seen by seeing a number of images between an image and its corresponding Fourier transform (or, its shape at the focal plane of our wavefront-preserving imaging system).

The first of such examples is given by figure 32.1, where we have a number of equally spaced stripes as the input. This then gives equally spaced points as the shape on the focal plane. We then consider three masks as follows:

- the first mask that we apply leaves only the centre point on the focal plane and removes all other points, this then leads to an image with nearly the same spatial frequency across;
- the second mask that we discuss removes everything other than the five central points on the focal plane, this retrieves the equally spaced stripes but the image is not as sharp as the input;
- the third mask masks out points on the focal plane but in a manner that it only does so for the second other point, this halves the spacing of the stripes.

The second example, illustrated in figure 32.2, is to consider a mesh of two sets of stripes with different orientations. By masking out a set of points in the focal plane corresponding to one orientation, we have successfully deleted that set of stripes at the image plane.

The third example is to consider an artwork, given in figure 32.3. Masking out the high frequency components keeps the shape of the artwork, however the fine details are blurred out. However, masking out the low frequency components leaves the details of the artwork but removes the shape of the artwork from the image.

Schlieren Photography

Our wavefront-preserving imaging system cannot show any relative phases of the light arriving upon the image plane, merely demonstrating its amplitude. In order to "see" the phases of

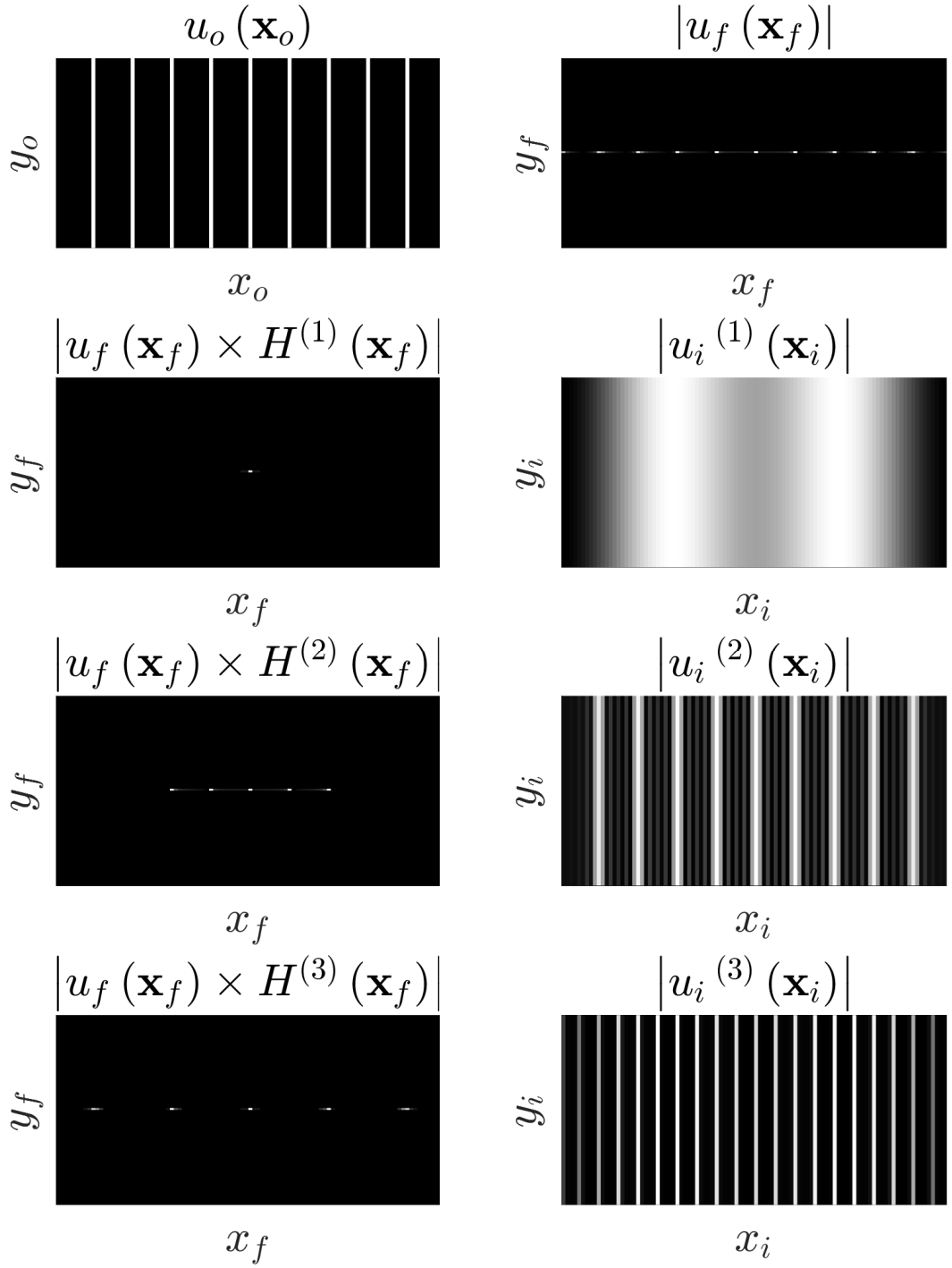

Figure 32.1: Using a pattern with thin stripes to illustrate the relation between the shapes at the focal and image planes.

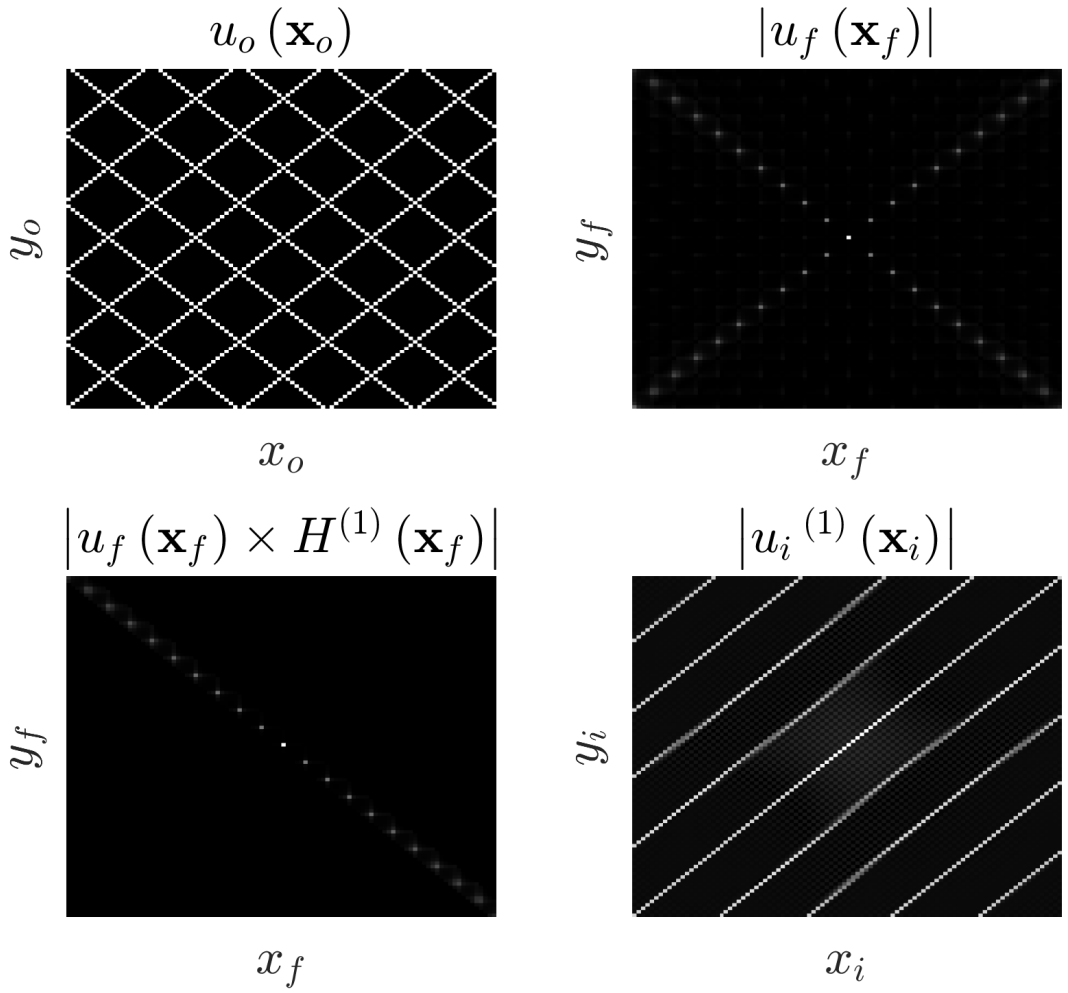

Figure 32.2: Filtering out a set of stripes with a certain orientation at the focal plane.

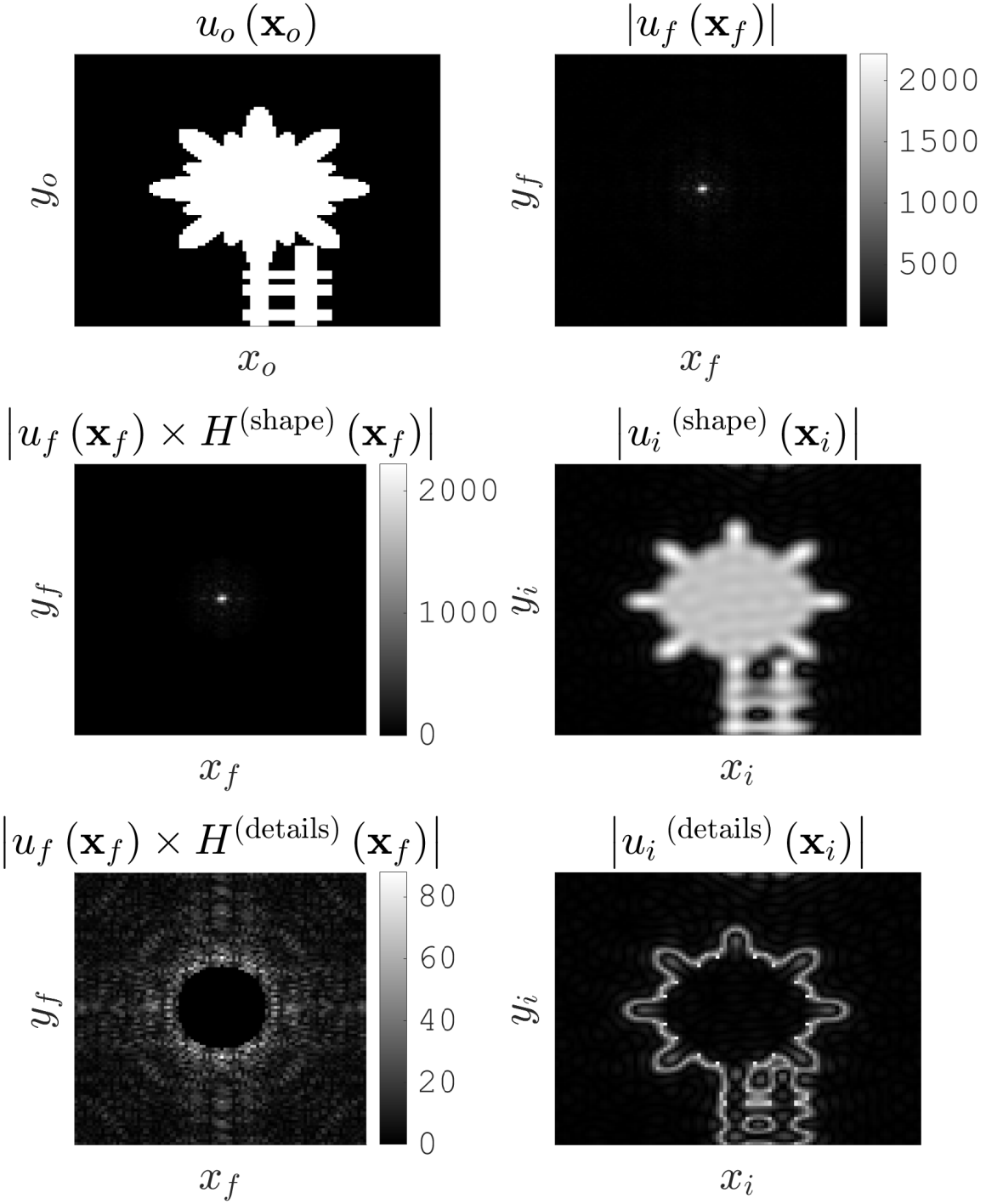

Figure 32.3: Manipulating an artwork at the focal plane to extract its shape and details.

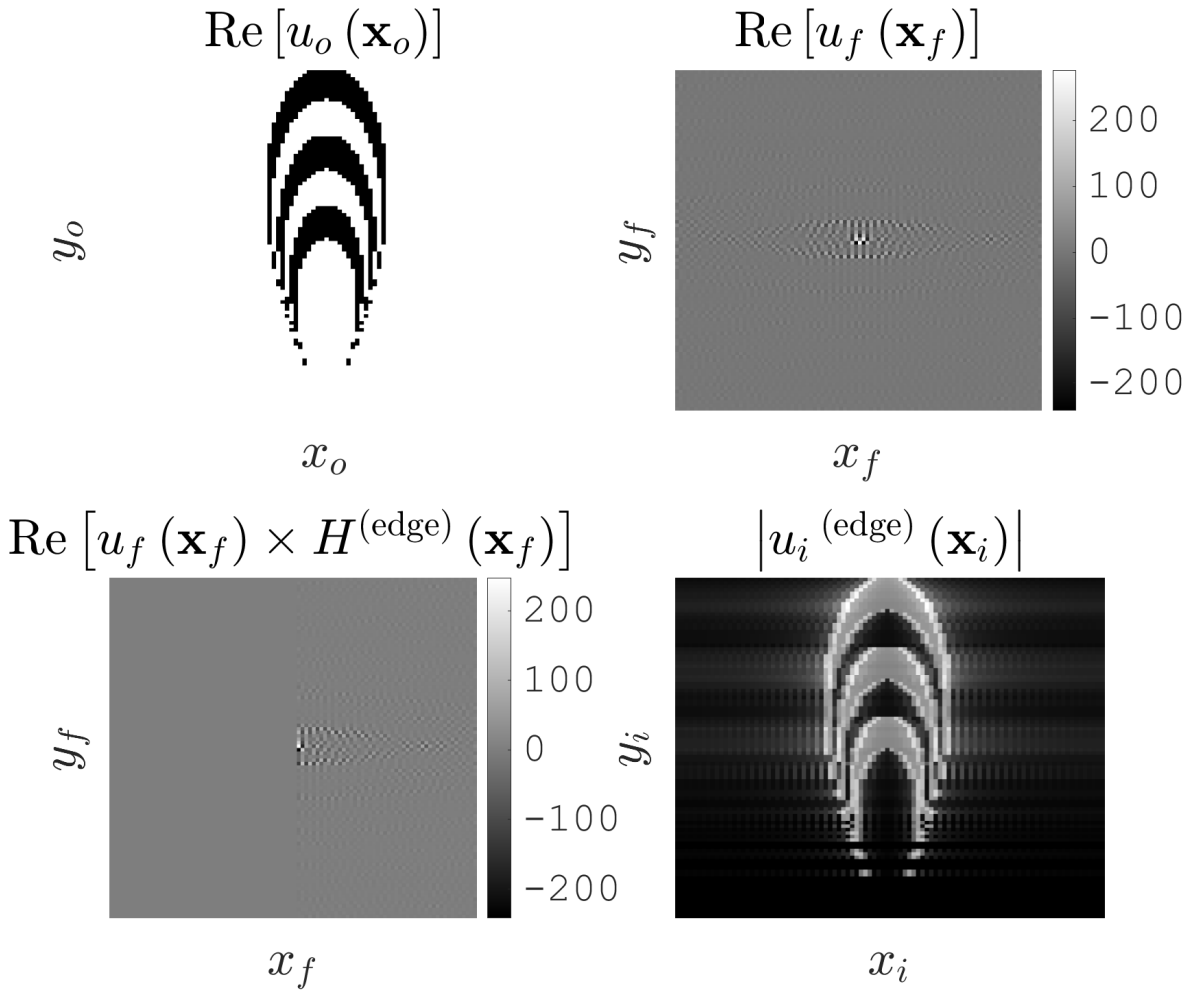

Figure 32.4: An illustration of creating a Schlieren photograph. Note that in the top-right graph, the zero frequency point (located at the middle) has such a large real part (9 542) removing the contrast of all the other frequencies, and hence is artificially removed from the figure.

the light, for example the phases originated from small refractive index variations of air, we may consider Schlieren photography, which is to mask out half the zero frequency point and a half-plane of frequencies at the focal plane, usually by a knife-edge (this sentence should be clear after a comparison between the top-right and bottom-left images of figure 32.4). After such a processing method, the phases will "turn into" different amplitudes.

A demonstration of this is figure 32.4, where our input is light with equal amplitude across, but we have introduced a $\pi/4$ phase change at some points at the input plane. If we were to use our wavefront-preserving imaging system at the output directly, then we would see a constant amplitude across. However, by introducing this knife-edge at the focal plane, we can clearly see the scalar amplitudes of the parts of the object with different phases are equipped with different moduli at the image plane, which is observable in the intensity variations of the light at the image plane.

Summary

1. By masking out certain points on the focal plane, we are able to alter the input light, giving structures of different shapes at the image plane. The low frequency points at the focal plane corresponds to the shape of the input, and the high frequency points at the focal plane corresponds to the details of the input.

2. By masking out the zero frequency point and a half-plane of frequencies, we are able to see the difference in phases in the input light (for example, that caused by refractive index variations of air) at the image plane.

6 Mach-Zehnder Interferometer

§33. Building an Interferometer

We shall now focus back on separating different wavelengths of light. There are three main methods for doing this, and we shall go through them one by one.

Dispersion
This is the effect that light with different wavelengths have different refractive indices in materials, and therefore we can use a prism-like apparatus to separate these wavelengths, just by using the dispersive properties of light. The resolution associated with this method is usually poor, and as a result this will not be the instrument that we look into in this course.

Division of Wavefront
This is how a grating works. We have a plane wave that travels towards the grating, and the grating splits the plane wave into many different smaller wavefronts, which then interferes to give fringes.

We note that in order for this to work, we need to have a plane wavefront travelling towards the grating, i. e. different parts of the plane wavefront must have a constant phase apart. This is a condition on the source, which is called **transverse coherence**.

Division of Amplitude
This is the method that we will look into next, which is to divide the amplitude of the wave into different parts, and give each one of the divided amplitude a phase, then send the divided amplitudes back together to interfere.

We note that in order for this method to work, we do not necessarily need a plane wavefront, as what we really need is the light to be coherent with itself delayed in time, i. e. **longitudinal coherence** or **temporal coherence**. Therefore practically, if one would like to resolve light with longitudinal but not transverse coherence, then the method would be to use a spectrometer with division of amplitude designs.

The three examples that exploits this effect that we shall look into are

- Mach-Zehnder interferometers;
- Michelson interferometers;
- Fabry-Perot etalons and interferometers,

which the next part of the course is to survey through these three different types of spectrometer designs.

Summary

1. Light with different wavelengths can be separated by passing it through a dispersive medium, which has a low resolution and therefore is not commonly used.
2. Light can be resolved by the division of wavefront, e. g. passing the light through a diffraction grating. This method requires transverse coherence.

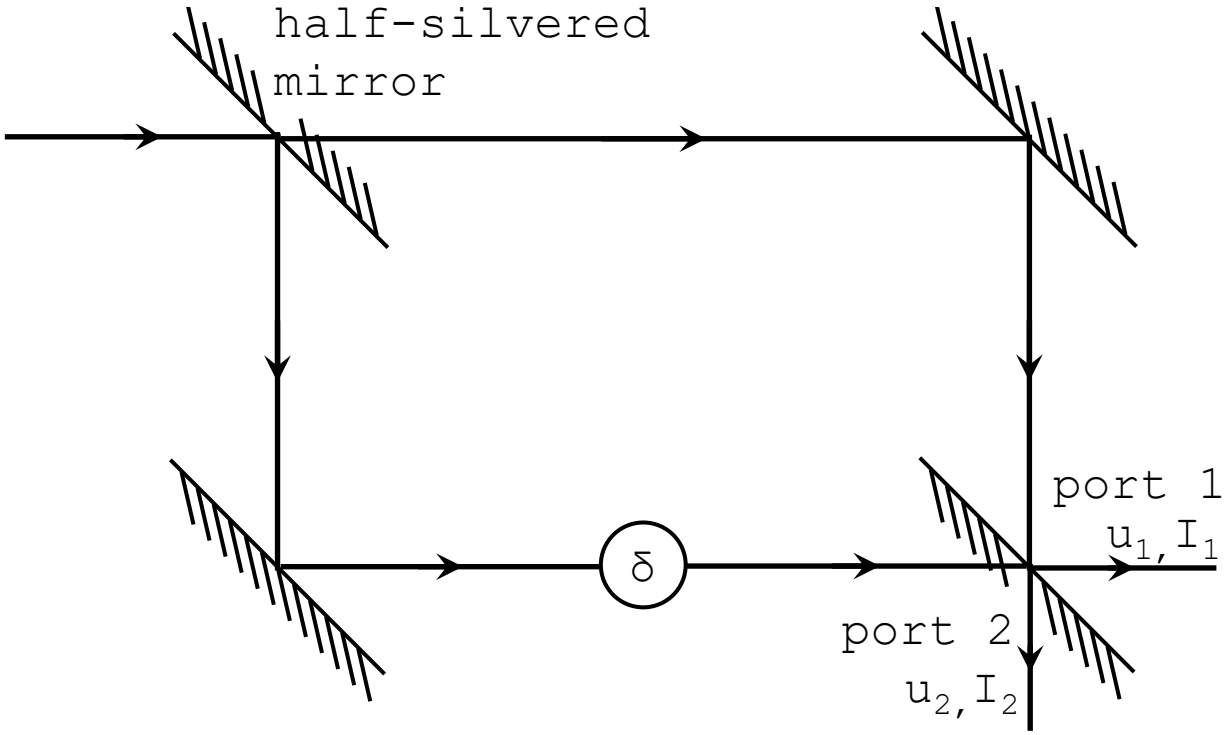

Figure 34.1: An illustration of the Mach-Zehnder Interferometer

3. Light can be resolved by the division of amplitude, i. e. breaking up the light into two directions, and interfere one direction with the other direction that is delayed in time.

§34. Mach-Zehnder Interferometer

The simplest interferometer by the division of amplitude is the Mach-Zehnder interferometer, illustrated in figure 34.1. Let us equip the input beam with intensity I_0 and scalar amplitude u_0, which is split by a half-silvered mirror (or beam splitter) into two equal intensities $I_0/2$, and hence with scalar amplitudes $u_0/\sqrt{2}$, where an extra π phase shift is to be added on the reflected beam.

Note that the beam splitter is constituted of a glass plate with a high refractive index and a dielectric coating, which means that only reflections from air to the coating will accumulate this phase shift; both reflection from the other side (i. e. entering the glass first) and transmission will not pick up any additional phase. In figure 34.1, a phase change is added every time the beam reflects off a beam-splitter or a mirror *with the oblique "bars" on the other side of the reflected beam*, as the bars represents the location of the glass.

After the division of amplitude by the beam splitter, the lower beam picks up an extra phase δ. The two beams then meets at the final half-silvered mirror, which then reduces the scalar amplitudes by a factor of $1/\sqrt{2}$ further, when the two beams are directed to the two different directions. The intensity at output port 1 is then

$$I_1 \;=\; |u_1|^2 \;=\; \left|\text{overall phase} \times \left(\frac{u_0}{\sqrt{2}} \times \frac{1}{\sqrt{2}} + \frac{u_0}{\sqrt{2}} \times \mathrm{e}^{\mathrm{i}\delta} \times \frac{1}{\sqrt{2}}\right)\right|^2 \;=\; I_0 \cos^2\left(\frac{\delta}{2}\right), \qquad (34.1)$$

where the overall phase includes both the phase change accumulated when the two beams travels to the final mirror and the phase change due to reflections off mirrors. Here we note

that both beams reflected twice upon arrival at port 1, and therefore the reflection off mirrors gives an overall phase, however for port 2, we note that the top beam reflected once off a mirror and the bottom beam reflected three times but only two of them accumulates a phase change (the final reflection is *in the glass*, and hence there is no phase shift), causing a *relative* phase shift of $e^{i\pi} = -1$ due to the reflection. This gives

$$I_2 = |u_2|^2 = \left|\text{overall phase} \times \left(\frac{u_0}{\sqrt{2}} \times \frac{1}{\sqrt{2}} - \frac{u_0}{\sqrt{2}} \times e^{i\delta} \times \frac{1}{\sqrt{2}}\right)\right|^2 = I_0 \sin^2\left(\frac{\delta}{2}\right). \quad (34.2)$$

Note that $I_1 + I_2 = I_0$, and therefore energy is conserved. Note that practically the mirrors and the beam-splitters will absorb energy and transfer it into heat, and therefore strictly speaking $I_1 + I_2 < I_0$.

We shall note that the Mach-Zehnder is quite a toy-model. It illustrates the main method of analysis of interference by division of amplitude, but there is this mysterious phase shift addition, which must be wavelength dependent if the interferometer is to separate wavelengths. We are currently just sealing it as a black box just to get the idea of interferometry by division of amplitude, but in order to be practical we have to "look into the black box" and investigate the methods of designing this phase addition. We will talk about one of such a method in the next apparatus, the Michelson interferometer.

Summary

1. The Mach-Zehnder interferometer introduces a phase shift δ of the lower beam, which then gives an overall interference pattern at the two ports

$$I_1 = I_0 \cos^2\left(\frac{\delta}{2}\right), \qquad I_2 = I_0 \sin^2\left(\frac{\delta}{2}\right). \quad (34.3)$$

7 Michelson Interferometer

§35. Michelson Interferometer

Let us continue from the black box phase shift of δ in the Mach-Zehnder interferometer. In the Michelson interferometer, we add in an extra path difference to one of the two beams from the beam splitter that is later superposed, which is transferred into this phase difference δ by multiplying with the wavenumber. Light with a shorter wavelength gives a larger wavenumber; and hence a larger phase addition. Therefore, the intensity I with respect to the extra path difference Δ is different for light with different wavelengths, hence allowing us to do spectrometry.

A diagram of the Michelson interferometer is given in figure 35.1. The beam splitter splits the amplitude from a source such that the two beams are reflected at two mirrors M_1 and M_2 at different distances from the beam splitter, where the common distance gives an overall phase, but the difference in the two distances gives a relative phase shift of one beam with respect to the other. An extra complication is that there can be an extra phase difference caused by the beam reflected at mirror M_2 caused by the beam travelling *inside* the glass that the beam splitter is made off, which is compensated by placing a **compensation plate** between the beam splitter and the mirror M_1 giving the other beam exactly the same phase shift, so now all the phase difference is purely caused by the difference in the distances.

We shall note that the physics of the Michelson interferometer is exactly the same as the Mach-Zehnder interferometer, and as a result, if the differences between the distances from the beam splitter to the two mirrors M_1 and M_2 is d, then since the light travels through this extra distance twice before interference, the optical path length difference is given as $\Delta\text{OPL} = 2d$, and therefore the output intensity at port 1 is

$$I_{\text{out}} = \left|u_0 + u_0 \mathrm{e}^{\mathrm{i}\delta}\right|^2 = I_0 \cos^2\left(\frac{\delta}{2}\right), \qquad \delta = k \times \Delta\text{OPL} = 2kd = \frac{4\pi}{\lambda} d. \tag{35.1}$$

The Michelson interferometer is an extremely versatile interferometric apparatus. We may not only look at the ports as intensity detectors for light bouncing off the mirrors at normal incidence, but also consider light hitting the two mirrors at a small angle, which then gives us fringes. We can also use the interferometer with an extended source instead of a point source. In this course we go through three basic methods of operating the Michelson interferometer:

- move d, i. e. move one of the mirrors, which gives a "travelling Michelson", or a **Fourier-transform spectrometer**;
- hold d stationary and look at the fringes formed: these fringes are called **Haidinger fringes**, or **fringes of equal inclination**;
- set $d = 0$ and tilt the two mirrors such that they are at an angle to one another and look at the fringes: these fringes are called **Fizeau fringes**, or **fringes of equal thickness**.

We shall first look at the Fourier transform spectrometer and its applications.

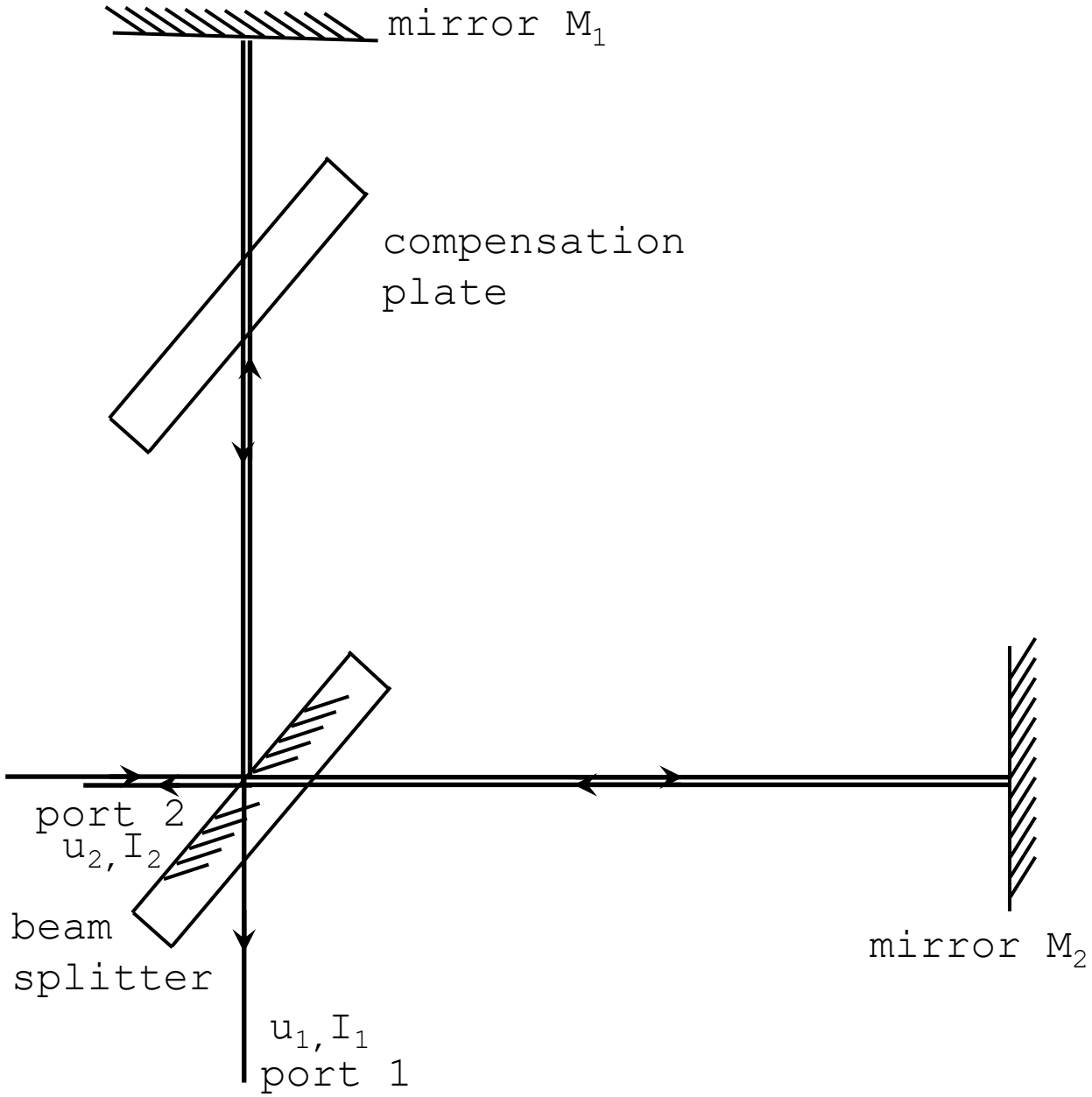

Figure 35.1: The setup of a Michelson interferometer.

Summary

1. The Michelson interferometer is formed such that the amplitude is divided and one of the two legs travels through an extra path length $2d$, which gives

$$I_{\text{out}} \;=\; I_0 \cos^2\left(\frac{\delta}{2}\right) \;=\; I_0 \cos^2\left(\frac{2\pi}{\lambda} d\right). \tag{35.2}$$

   The Michelson interferometer is very versatile and we shall explore its usage in the next few sections.

§36. Fourier-Transform Spectroscopy

Separation of Wavelengths

Let us now look at a Fourier-transform spectrometer, where d is moved and we record the received intensity. Recall that the intensity output from the Michelson interferometer reads

$$I_{\text{out}} \;=\; I_0 \cos^2\left(\frac{\delta}{2}\right), \tag{36.1}$$

which is algebraically identical to

$$I(\Delta) \;=\; \frac{1}{2} I_0 [1 + \cos(\delta)] \;=\; \frac{1}{2} I_0 + \frac{1}{2} I_0 \cos(k\Delta). \tag{36.2}$$

Note that we have the difference in optical path length $\Delta \;=\; \Delta\text{OPL} \;=\; 2d$, which is the object we vary in a Fourier-transform spectrometer. As a result, if we have a light with a wide range of N different frequencies ω_m where $m \;=\; 1, \cdots, N$, then they have different wavenumebers $k_m \;=\; \omega_m / c$, and each of them will contribute to a constant $\frac{1}{2} I_0$ term, but they will give different cosine terms as they have different wavenumbers. Thus, the intensity is

$$I(\Delta) \;=\; \langle I \rangle + \frac{1}{2} \sum_{m\,=\,1}^{N} I_m \cos(k_m \Delta), \qquad \langle I \rangle \;=\; \frac{1}{2} \sum_{m\,=\,1}^{N} I_m. \tag{36.3}$$

We shall note that each cosine term is a sum of two exponentials with opposite exponents, i. e.

$$\cos(k_m \Delta) \;\propto\; \mathrm{e}^{\mathrm{i} k_m \Delta} + \mathrm{e}^{-\mathrm{i} k_m \Delta}, \tag{36.4}$$

suggesting that if we obtain the intensity pattern with respect to the optical path length, and we do a Fourier transform of the cosine terms on a computer, then the spectrum

$$S(k) \;=\; \mathscr{T}_F[I(\Delta) - \langle I \rangle](k) \;=\; \int_{-\infty}^{\infty} \mathrm{d}\Delta\, \mathrm{e}^{\mathrm{i} k \Delta} \times [I(\Delta) - \langle I \rangle] \;\propto\; \sum_{m\,=\,1}^{N} I_m [\delta(k - k_m) + \delta(k + k_m)], \tag{36.5}$$

are pairs of Dirac-deltas centred at the input wavenumbers $\pm k_m$ — so we have successfully separated the wavelengths. Or, alternatively, to eliminate the Dirac-deltas with negative k_ms, we can consider doing a **Fourier cosine transform**

$$S_{\text{c}}(k) \;=\; \mathscr{T}_{\text{c}}[I(\Delta) - \langle I \rangle](k) \;=\; \int_{-\infty}^{\infty} \mathrm{d}\Delta\, \cos(k\Delta) \times [I(\Delta) - \langle I \rangle] \;\propto\; \sum_{m\,=\,1}^{N} I_m [\delta(k - k_m)], \tag{36.6}$$

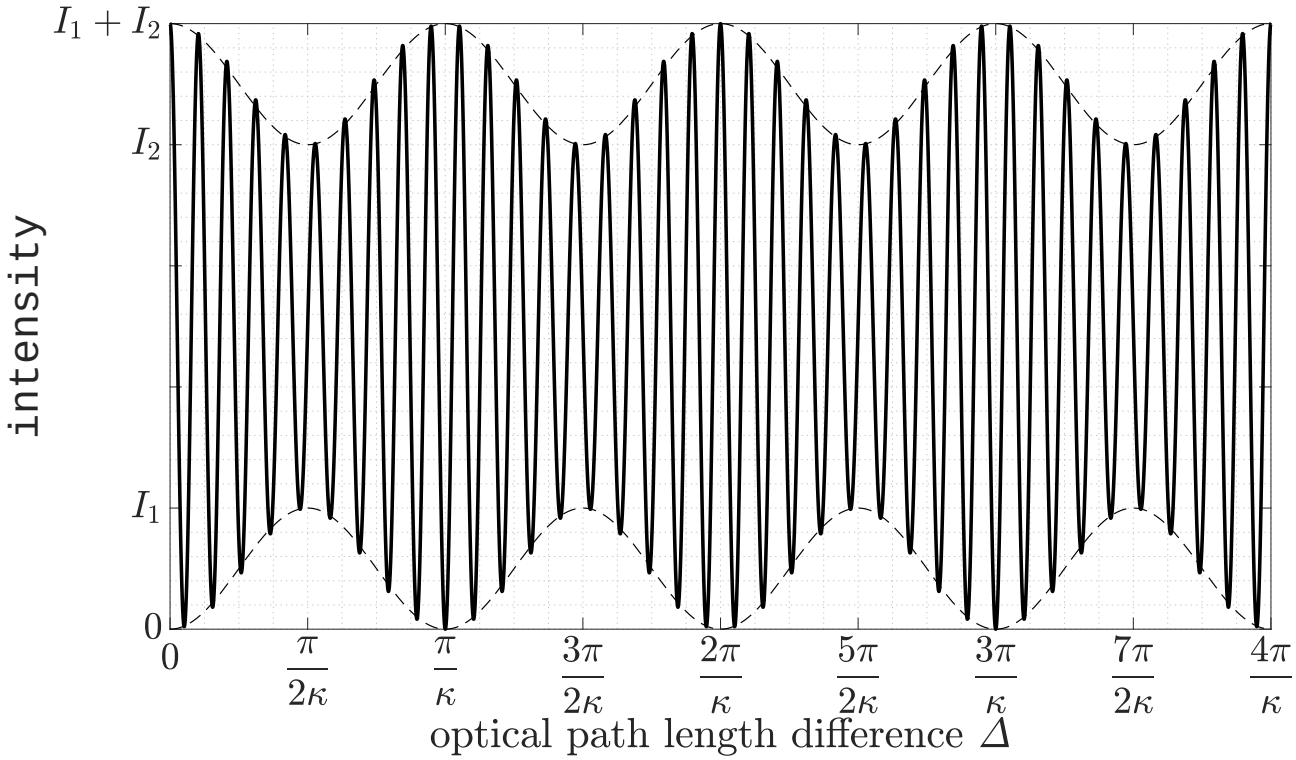

Figure 36.1: The intensity measurement of a Fourier-transform spectrometer of a source with two wavelengths.

a spectrum with *just* the Dirac-deltas centred on the positive k_ms.

Resolving two Wavelengths

Let us illustrate this method by sending in a light with two different frequencies k_1 and k_2 with intensities I_1 and I_2 respectively. Without loss of generality, let us set $I_1 > I_2$. Then, the intensity collected by varying the difference in distances from the beam splitter to the two mirrors d is

$$I(\Delta) = I_1 \cos^2\left(\frac{1}{2}k_1\Delta\right) + I_2\cos^2\left(\frac{1}{2}k_2\Delta\right) = \frac{I_1+I_2}{2} + \frac{I_1}{2}\cos(k_1\Delta) + \frac{I_2}{2}\cos(k_2\Delta), \quad (36.7)$$

where $\Delta = 2d$. Written in terms of the average wavenumber $K = \frac{1}{2}(k_1+k_2)$ and the half-difference in wavenumber $\kappa = \frac{1}{2}(k_2-k_1)$, this reads

$$I(\Delta) = \frac{I_1+I_2}{2}[1+\cos(K\Delta)\cos(\kappa\Delta)] + \frac{I_1-I_2}{2}[\sin(K\Delta)\sin(\kappa\Delta)]. \quad (36.8)$$

This is demonstrated in figure 36.1. Note that, the maxima and minima of the intensity pattern can be determined naively, just by comparing the phases travelled by the two beams of light: the beams of light are in phase if

$$k_2\Delta = k_1\Delta + 2\pi p, \quad \Rightarrow \quad \Delta = p\pi/\kappa \quad (36.9)$$

where p is an integer, and the beams of light are out of phase if

$$k_2\Delta = k_1\Delta + 2\pi p + \pi, \quad \Rightarrow \quad \Delta = p\pi/\kappa + \pi/(2\kappa). \quad (36.10)$$

We shall take special note that hitting a minima does not mean that there is no intensity, instead it means that the overall intensity is closer to the average of the two different intensities, also demonstrated in figure 36.1.

After this analysis, note that if the light we input is composed of two different wavenumbers, then by reading off the intensity variation against Δ and locating the maxima and minima, even without inputting it into a computer and doing a Fourier transform, we are able to work out the individual wavenumbers and intensities: we can work out κ by locating the position of maxima and minima of the large scale variations, and we can work out K by locating the position of maxima and minima of the small-scale variations. Then we can use the values of K and κ to work out the values of k_1 and k_2. Then, at the positions of large-scaled maxima and minima, we may find that the difference in intensities of consecutive small-scaled fringes are $I_1 + I_2$ and $I_1 - I_2$ respectively, and hence from there we may work out the differences in intensities of the two individual components.

Summary

1. If we input a range of wavenumbers k_m into the Fourier-transform spectrometer, and vary the difference between distances from the beam splitter to the two mirrors d to obtain the intensity of the light as a function of d, we can retrieve the wavenumbers we have input by doing a Fourier cosine transform of the intensity

$$S_{\mathrm{c}}(k) \;=\; \mathscr{T}_{\mathrm{c}}[I(\Delta) - \langle I\rangle](k) \;=\; \int_{-\infty}^{\infty} \mathrm{d}\Delta\, \cos(k\Delta) \times [I(\Delta) - \langle I\rangle] \;\propto\; \sum_{m\,=\,1}^{N} I_m[\delta(k - k_m)], \tag{36.11}$$

   where $\Delta \;=\; 2d$.

2. If we input light with just two different wavelengths into the Fourier-transform spectrometer, then we are able to find the wavenumbers k_1 and k_2 and the corresponding intensities I_1 and I_2 by just analysing the intensity distribution with respect to Δ, without actually doing the Fourier transform.

§37. Chromatic Resolving Power of a Fourier-Transform Spectrometer

Temporal Parts of the Intensity

We note that we have been hiding the temporal parts of the scalar amplitudes in the analysis previously without actually justifying whether we are allowed to do that. We shall justify this simplification here. Let us just consider the simple case where we input two different wavelengths, then, the intensities of the two components due to the temporal parts of the scalar amplitudes are given by

$$I_m \;\propto\; \left|u_m \mathrm{e}^{-\mathrm{i}\omega_m t}\right|^2 \tag{37.1}$$

where $m \;=\; 1, 2$. Then, the output will be

$$I(\Delta, t) \;=\; I_1(\Delta) + I_2(\Delta) + 2u_1u_2 \cos[(\omega_2 - \omega_1)t], \tag{37.2}$$

where the temporal part averages to 0 over time. Therefore if the detector is slow, then we can safely ignore the temporal part of the scalar amplitude, yet if the detector is fast, i. e. in the gigahertz range, then we will be able to measure the extra difference caused by this temporal variation, which has an effect on the intensity called **beat note**. For more than two sources,

the story is similar: we simply need to add on top of equation 36.3 a term that is dependent on time, which we average to 0 if we have a very slow detector.

Chromatic Resolving Power of the Fourier-Transform Spectrometer

We now think about resolving two different wavelengths using the Fourier-transform spectrometer, which follows the logic of §22: we need to find the minimum wavenumber difference resolvable by the Fourier-transform spectrometer. Recall that, to resolve two wavelengths, we look at the large scale variations of the intensity pattern, with one full period having a separation in Δ of π/κ, where

$$\kappa \;=\; (k_1 - k_2)/2 \;=\; \Delta k/2 \;=\; \pi \times \Delta\bar{\nu}. \tag{37.3}$$

In order to distinguish between the two wavelengths, let us say that we need to see one whole period of the large scale variation, which is limited by the greatest difference in the optical path lengths achievable by the spectrometer $\Delta_{\max}$. Let us say that the two wavelengths are just distinguishable, so the separation in Δ is $\Delta_{\max}$, then the difference in wavenumbers, i. e. the instrumental width, is

$$\Delta_{\max} \;=\; \frac{\pi}{\pi \times \mathrm{INST}_{\bar{\nu}}} \qquad \Rightarrow \qquad \mathrm{INST}_{\bar{\nu}} \;=\; \frac{1}{\Delta_{\max}}. \tag{37.4}$$

This means that the chromatic resolving power of the Fourier-transform spectrometer is given as

$$\mathscr{P} \;=\; \frac{\bar{\nu}_{\mathrm{mean}}}{\mathrm{INST}_{\bar{\nu}}} \;=\; \bar{\nu}_{\mathrm{mean}}\Delta_{\max} \;=\; \frac{\Delta_{\max}}{\lambda_{\mathrm{mean}}}, \tag{37.5}$$

which, in contrary to the grating spectrometer, is different for different wavenumbers.

Summary

1. We are safe to ignore the temporal part of the scalar amplitude when working out the intensity of the Fourier-transform spectrometer, unless our detector can detect signals with very high (gigahertz) frequencies, for which we shall see beat notes of the intensity appearing on our detector.

2. The instrumental width of and the chromatic resolving power of the Fourier-transform spectrometer are

$$\mathrm{INST}_{\bar{\nu}} \;=\; \frac{1}{\Delta_{\max}}, \qquad \mathscr{P} \;=\; \bar{\nu}_{\mathrm{mean}}\Delta_{\max} \;=\; \frac{\Delta_{\max}}{\lambda_{\mathrm{mean}}}, \tag{37.6}$$

   where $\Delta_{\max}$ is the largest difference in the optical path lengths from the beam splitter to the two mirrors achievable by the spectrometer, which limits the power of the Fourier-transform spectrometer.

§38. Finite Coherence Length and Line-Broadening

Intensity of a Source with a Finite Coherence Length

Previously we have suggested that a monochromatic light beam through a Fourier-transform spectrometer gives a cosine output, hence implicitly suggesting that we are able to see oscillations for any Δ we select. However in reality this is not the case: if Δ is too large, then the oscillations will disappear. In this course we focus on two reasons of why this might happen,

which are finite coherence length and line-broadening. Both of them have to do with the source which we now see as non-ideal.

We first introduce the concept of coherence length and study its implications. Previously we have stated that for an interferometer that uses the method of division by amplitude, the light source must have longitudinal or temporal coherence, or, the scalar amplitude $u(t)$ should be similar to the scalar amplitude $u(t+\tau)$. This is usually not the case, as most light sources have a finite **coherence time** t_c, after which the coherence is lost, i. e. $u(t)$ and $u(t+t_c)$ looks so different that no spatial oscillations are visible at all. The light travels a **coherence length** $\ell_c = c \times t_c$ in its coherence time, after which the light will be completely incoherent with itself. The simplest model of this is to assume that the source emits many "light pulses" that oscillates for a distance ℓ_c. Each of these light pulse can be thought as a wavetrain that is only ℓ_c long (and hence the E-fields are sinusoidal with a tophat envelope). The effect of passing light through an extra path difference Δ is to delay itself in time, and therefore when we see no spatial oscillations in the intensity pattern for a monochromatic light source, we immediately know that we have set Δ too large such that we have exceeded the coherence length of the light. A demonstration of this is shown in figure 38.1.

Applying this logic backwards, we can use the fact that we see no spatial oscillations being equivalent to the condition that the coherence length has been exceeded to measure the coherence length ℓ_c. To do this, we increase Δ from 0, and as the light suffers from decoherence, we see the maxima and minima closer and closer to the average intensity, as illustrated by figure 38.1. The moment when we see no oscillations corresponds to the optical path length difference Δ equal to the coherence length ℓ_c.

To measure how coherent the light is, we can use the function **visibility** defined as

$$\gamma = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \tag{38.1}$$

where the maximum and minimum refers to the small-scale oscillations. We can see from figure 38.1 that when the two waves that interferes have perfect coherence $\gamma = 1$, and when the two waves that interferes are completely incoherent $\gamma = 0$.

Effect of Line-Broadening on the Image of Fourier-Transform Spectroscopy

Recall that if the light source that we feed into the Fourier-transform spectrometer has discrete wavelengths, or, the intensity in terms of wavelengths has "lineshapes" (that is, the intensity of the light emitted from the source against frequency) as Dirac-deltas, then the intensity distribution function at the interferometer will be oscillatory for all Δ if the source has perfect temporal coherence. However all real light sources have line-broadening effects, and therefore no real sources has lineshapes as Dirac-deltas. This broadening effect will cause the spatial oscillations to decay as Δ increases, and the shape of the envelope that causes this decay is determined by the shape of the input light source. Some of the lineshapes and corresponding intensity distributions on the Fourier-transform spectrometer is shown in figure 38.2.

A Quantitative Discussion of Visibility and Coherence

A lot of our discussion in this chapter is based on light that is partially coherent, and the only quantitative measure for how coherent two light beams are that has been introduced is the visibility γ, where we have not provided any justification.

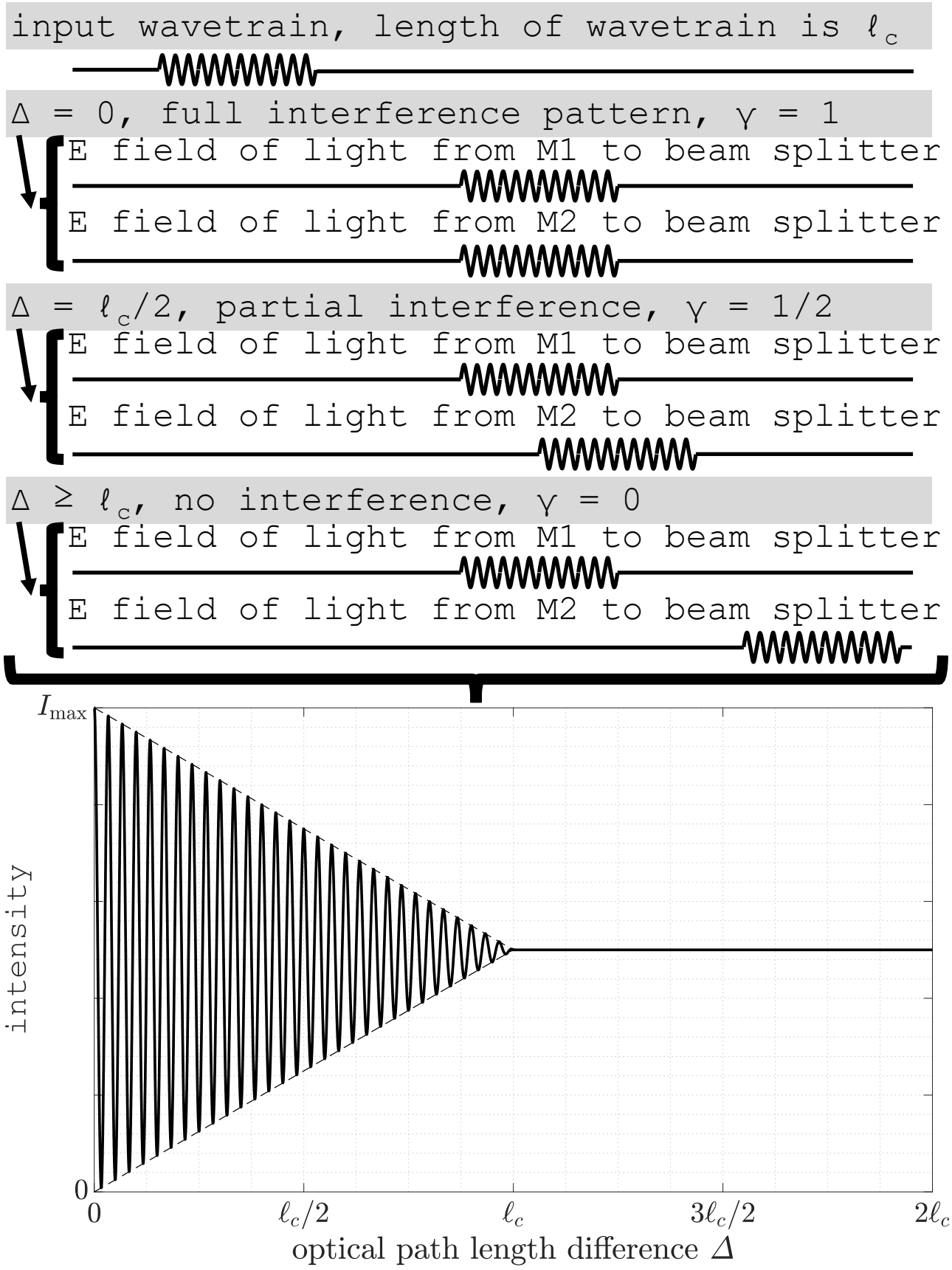

Figure 38.1: Passing two wavetrains with coherence length ℓ_c through a Michelson interferometer. The two wavetrains each have a tophat envelope, and the total envelope in the intensity can be seen as the convolution of these two tophats.

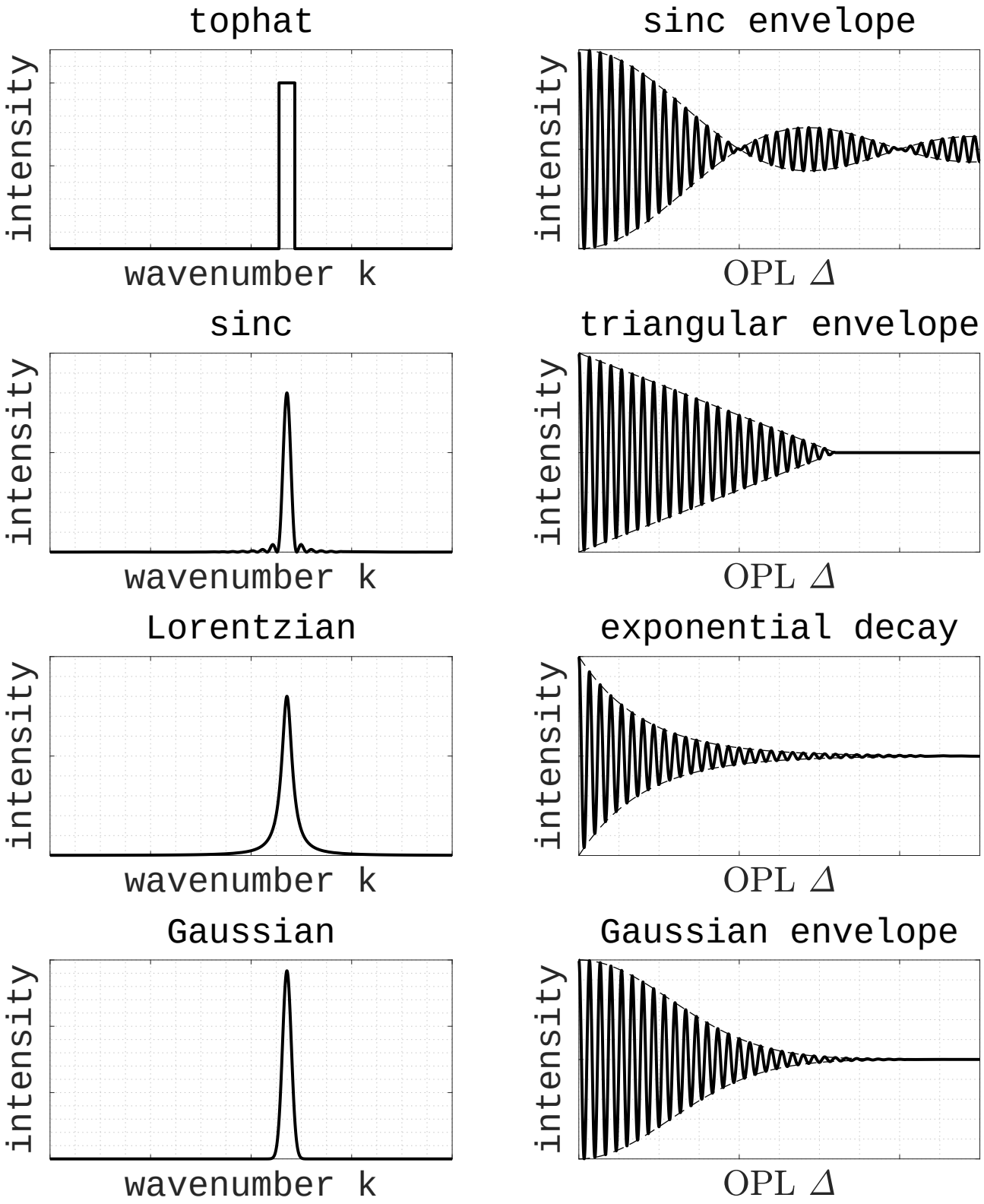

Figure 38.2: The effect on the intensity distribution of Fourier-transform spectroscopy by different lineshapes of the source. The left column is the different lineshapes, and the right column is the corresponding intensity distribution on the Fourier-transform spectrometer corresponding to the lineshape on the same row. Note that the wavenumber of the small-scale oscillatios on the right corresponds to exactly the centre of the lineshape on the left.

To justify that the visibility is a good measure for coherence, let us consider equation 35.1 again, where we have combined u_0 and $u_0 e^{i\delta}$, or, with their temporal parts added in, $u_0 e^{-i\omega t}$ and $u_0 e^{-i\omega t+i\delta}$. For a generalised discussion let us call these contributions u_A and u_B. Then, the average intensity over time is given as

$$\begin{aligned} I &= \left\langle |u_A + u_B|^2 \right\rangle_t \\ &= I_A + I_B + \langle u_A u_B^* + u_B u_A^* \rangle_t \\ &= I_A + I_B + 2\mathrm{Re}\left[\Gamma\right] \\ &= I_A + I_B + 2\sqrt{I_A I_B}\mathrm{Re}\left[\tilde{\gamma}\right] \end{aligned} \tag{38.2}$$

where $I_A = \left\langle |u_A|^2 \right\rangle_t$, $I_B = \left\langle |u_B|^2 \right\rangle_t$, $\Gamma = \langle u_A u_B^* \rangle_t$, and $\tilde{\gamma} = \langle u_A u_B^* \rangle_t / \sqrt{I_A I_B}$. For coherent light sources with equal intensities I_0, i. e. $u_A = u_0 e^{-i\omega t+i\delta}$ and $u_B = u_0 e^{-i\omega t}$, we have the frequency ω for the two contributions equal. In this case, the modulus of Γ equates to I_0 and the phase of Γ equates to δ. Then, $2\mathrm{Re}\left[\Gamma\right]$ is a sinusoidal interference term with respect to δ with amplitude $2I_0$, hence creating the interference pattern. On the other hand, if k and ω are unequal, then the average of the cross term in time vanishes and therefore we have $2\mathrm{Re}\left[\Gamma\right] = 0$. Therefore we note that $|\Gamma|$ is a good measure for the coherence of two light beams, equating to I_0 for coherent light with equal amplitudes and zero for incoherent light. Γ is given the name **mutual coherence function**. The division of $|\Gamma|$ by $\sqrt{I_A I_B}$ then gives $|\tilde{\gamma}|$, which equates to unity for coherent light and zero for incoherent light, getting the absolute intensity of the light beams out of the way and only focussing on how coherent they are. $\tilde{\gamma}$ is called the **degree of coherence**.

Now that we have convinced ourselves that $|\tilde{\gamma}|$ is a good measure of coherence, we only need to formulate a relation between γ and $|\tilde{\gamma}|$ to show that γ is indeed also a good measure for coherence. Noting that the extrema of $\mathrm{Re}\left[\tilde{\gamma}\right]$ is given as $\pm|\tilde{\gamma}|$, we have the extrema of I as

$$I_{\max} = I_A + I_B + 2\sqrt{I_A I_B}|\tilde{\gamma}|, \qquad I_{\min} = I_A + I_B - 2\sqrt{I_A I_B}|\tilde{\gamma}|, \tag{38.3}$$

and therefore using equation 42.3, we have

$$\gamma = \left[2\sqrt{I_A I_B}/(I_A + I_B)\right]|\tilde{\gamma}|, \tag{38.4}$$

where, in the case where the two beams of light are equal in amplitude, such as the beams in a Michelson interferometer, we set $I_A = I_B$, and then γ equates to $|\tilde{\gamma}|$ exactly.

Summary

1. Light sources have a finite coherence time t_c, which means that they have a finite coherence length $\ell_c = c \times t_c$. This can be measured by a Fourier-transform spectrometer: when the spatial variations vanish, then the optical path length difference Δ is equal to the coherence length ℓ_c.

2. Light sources will usually contain a line-broadening effect: the lineshapes are not exactly Dirac-deltas. This means that the intensity distribution on the Fourier-Transform spectrometer will have spatial oscillations decaying, with the envelope of this decay dependent on the lineshape of the light source.

3. The modulus of the degree of coherence $\tilde{\gamma} = \langle u_A u_B^* \rangle_t / \sqrt{I_A I_B}$ is a good measure of coherence between two light beams A and B. The degree of coherence and the visibility is linked by
$$\gamma = \left[2\sqrt{I_A I_B}/(I_A + I_B)\right]|\tilde{\gamma}|, \tag{38.5}$$
and therefore visibility is also a good measure of coherence.

§39. Doppler Broadening

Effect of a Moving Source on the Emitted Light

We have discussed some generalities in the effect of line-broadening in the previous section. We have stated that all sources have this effect without explaining why. This section presents one reason: that is, for a light source such as a discharge lamp, the atoms in the lamp that emits the light all move at individual velocities v, following Maxwell's distribution
$$f(v) \propto \mathrm{e}^{-v^2/{v_{\mathrm{th}}}^2}, \qquad v_{\mathrm{th}} = \sqrt{2k_{\mathrm{B}}T/m}. \tag{39.1}$$
This means that although the light emitted by each atom have the same frequency viewed in the rest from of these individual atoms, when they hit the detector their frequencies will be Doppler shifted. This effect is called **Doppler broadening** and the following discussion quantifies this.

We first investigate into the deviations in wavenumber caused by this effect. Recall that each photon has an energy
$$E = h\nu_0 = hc\bar{\nu}_0 \tag{39.2}$$
and a momentum
$$p = E/c = h\bar{\nu}_0, \tag{39.3}$$
where ν_0 and $\bar{\nu}_0$ are the frequency and wavenumber of the photon viewed in the rest frame of the atom. If the atom then moves away from the detector at a speed v, then it is clear that the relation between the energy and momentum of the photon seen by the detector and that of the atom's rest frame is given by
$$\begin{pmatrix} h\bar{\nu} \\ h\bar{\nu} \end{pmatrix} = \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} h\bar{\nu}_0 \\ h\bar{\nu}_0 \end{pmatrix}, \qquad \beta = \frac{v}{c}, \qquad \gamma = \frac{1}{\sqrt{1-\beta^2}}, \tag{39.4}$$
an inverse Lorentz transform. Explicitly performing the matrix multiplication, we have the relationship
$$\bar{\nu} = \beta\gamma\bar{\nu}_0 + \gamma\bar{\nu}_0 = (1+\beta)\left(1-\beta^2\right)^{-\frac{1}{2}}\bar{\nu}_0 \approx (1+\beta)\bar{\nu}_0 \tag{39.5}$$
for small β, or
$$\bar{\nu} - \bar{\nu}_0 = \frac{v}{c}\bar{\nu}_0. \tag{39.6}$$
Now we have a link between the velocity of the atom and the wavenumber of the atom, we are ready to see the effect of viewing such a source on a Michelson interferometer.

Fringe Pattern Seen by the Michelson Interferometer

Now that we have a link between v and $\bar{\nu}$, we are able to utilise Maxwell's distribution to write down the lineshape of the source by simply replacing v with $(\bar{\nu} - \bar{\nu}_0)c/\bar{\nu}_0$ using equation 39.6, giving a Gaussian
$$S \propto \mathrm{e}^{-(\bar{\nu}-\bar{\nu}_0)^2c^2\big/\left({v_{\mathrm{th}}}^2\bar{\nu}_0^2\right)} = \mathrm{e}^{-(k-k_0)^2/\Gamma^2}, \qquad \Gamma^2 = \frac{4\pi^2\bar{\nu}_0^2{v_{\mathrm{th}}}^2}{c^2} = \frac{8\pi^2k_{\mathrm{B}}T}{mc^2\lambda_0^2}. \tag{39.7}$$

Here we have used the relationship $\bar{\nu} = 1/\lambda = k/(2\pi)$. However what we would really want to know is the intensity of the fringes with respect to the optical path length difference Δ. To do this we consider an inverse Fourier transform of the Gaussian to give

$$I(\Delta) - \langle I \rangle = \mathscr{F}_F^{-1}[S] \propto Q(\Delta) \mathrm{e}^{-(\Gamma\Delta/2)^2}, \tag{39.8}$$

ditching the component caused by k_0 as that gives us the quick oscillating fringes into $Q(\Delta)$, which we are not particularly interested in. Therefore, the envelope of the fast oscillations of the fringes seen by the detector is given as

$$I(\Delta) = I_0 \pm I_0 \mathrm{e}^{-(\Gamma\Delta/2)^2}. \tag{39.9}$$

We inspect from figures 38.1 and 38.2 that a source that is line-broadened creates a similar pattern to the effect of finite coherence length, and therefore people usually define a coherence length for a line-broadened source analogously, that is, define the coherence length when the quick oscillations disappears as seen by the Michelson interferometer. However in the case of Doppler broadening, the quick oscillations do not vanish entirely, which creates a problem in defining the coherence length of the source. This is done differently in different literature. One such definition of the coherence length ℓ_c is the value of Δ such that the envelope is 1/e the width of the $\Delta = 0$ intensity, i. e.

$$\ell_c = 2\Gamma^{-1}. \tag{39.10}$$

Note that this logic can also be applied backwards, i. e. we can infer, for example, the temperature of the discharge lamp by looking at the fringes. To do so, we collect the intensity data against Δ and record ℓ_c, then use this to determine Γ. Then, the temperature of the lamp is given as

$$T = \frac{mc^2\Gamma^2\lambda^2}{8\pi^2 k_\mathrm{B}}. \tag{39.11}$$

Of course, the source may also be objects other than a discharge lamp, such as a distant star, but the same recipe applies. Note that the Gaussian envelope detected by the Fourier transform spectrometer can also serve as an experimental evidence of Maxwell's distribution of velocities in kinetic theory.

Summary

1. For a source emitting light with wavenumber $\bar{\nu}_0$ in its rest frame that is moving away from a detector at a velocity v, the wavenumber detected $\bar{\nu}$ and $\bar{\nu}_0$ follows the relation

   $$\bar{\nu} - \bar{\nu}_0 = v\bar{\nu}_0/c. \tag{39.12}$$

   For light from a discharge lamp each photon suffers from this, causing an envelope in the fringe pattern. This effect is called Doppler broadening.

2. We may define the coherence length $\ell_c = 2\Gamma^{-1}$ as the value of Δ such that the envelope is 1/e the width of the $\Delta = 0$ intensity. Then, the temperature of the source is given by

   $$T = \frac{mc^2\Gamma^2\lambda^2}{8\pi^2 k_\mathrm{B}}. \tag{39.13}$$

§40. Aether Drift (Michelson-Morley) Experiment

The aether drift experiment is probably the most famous experiment done with the Michelson interferometer, which demonstrated that aether, or the hypothetical unique medium of light which light travels at a speed c with respect to, does not exist, instead light travels at a speed c with respect to *all* reference frames.

To demonstrate how the experiment works, let us write down a few results predicted by a theory of space-time with aether. It is usually thought that the aether frame is stationary with respect to the Sun, and therefore if we build an Michelson interferometer on Earth, then the whole apparatus travels at a speed v, where v is the speed of the rotation of the point on Earth where the apparatus is located with respect to the Sun (hence taking both the orbital and rotational motion of the Earth into account). Let us point that velocity parallel to line that extends from the beam splitter to mirror M_2. Then, the apparatus travels in space as the light travels across the apparatus, demonstrated in figure 40.1.

Let us set the distances from the beam splitter to both mirrors M_1 and M_2 as L in the lab frame. By the geometry of the light travelling towards mirror M_1, we have the time t_1 of the light travelling towards M_1 as

$$L^2 + (vt_1)^2 \;=\; (ct_1)^2 \quad\Rightarrow\quad t_1 \;=\; \frac{L}{\sqrt{c^2 - v^2}}, \tag{40.1}$$

and therefore the optical path length of the light bouncing off at mirror M_1 is given as

$$\mathrm{OPL}_1 \;=\; 2ct_1 \;=\; 2c \times \frac{L}{\sqrt{c^2 - v^2}} \;=\; 2L\left(1 + \frac{v^2}{2c^2}\right) \tag{40.2}$$

in the limit $v \ll c$. For light travelling to the second mirror, light has to first "chase up" with the mirror M_2, then "meet up" with the beam splitter, where both the mirror and the beam splitter is travelling with a speed v; and therefore light travels through an optical path length

$$\mathrm{OPL}_2 \;=\; \frac{L}{c - v} \times c + \frac{L}{c + v} \times c \;=\; 2c \times \frac{L}{c^2 - v^2} \;=\; 2L\left(1 + \frac{v^2}{c^2}\right). \tag{40.3}$$

This gives the difference in optical path lengths as

$$\Delta \;=\; Lv^2/c^2, \tag{40.4}$$

and therefore the speed of the apparatus is

$$v \;=\; c\sqrt{\Delta/L}. \tag{40.5}$$

From this let us determine the lowest detectable speed of the laboratory frame with respect to the aether frame, if we compare the fringe pattern with the fringe pattern with $v = 0$, achieved by rotating the setup which floats on a giant pool of mercury. Say that we are able to distinguish the intensity change out of port 1 if it is deviated from the maximum by 0.01 times the separation between consecutive maxima, i. e. our criterion for distinguishability is $\Delta \geqslant 0.01\lambda$. If we then equip $L = 11\,\mathrm{m}$ which is close to the historical apparatus, and send light with wavelength $\lambda = 440\,\mathrm{nm}$ into the interferometer, then we have the range of detectable speeds as

$$v \;\geqslant\; 6\,000\,\mathrm{m\,s^{-1}}. \tag{40.6}$$

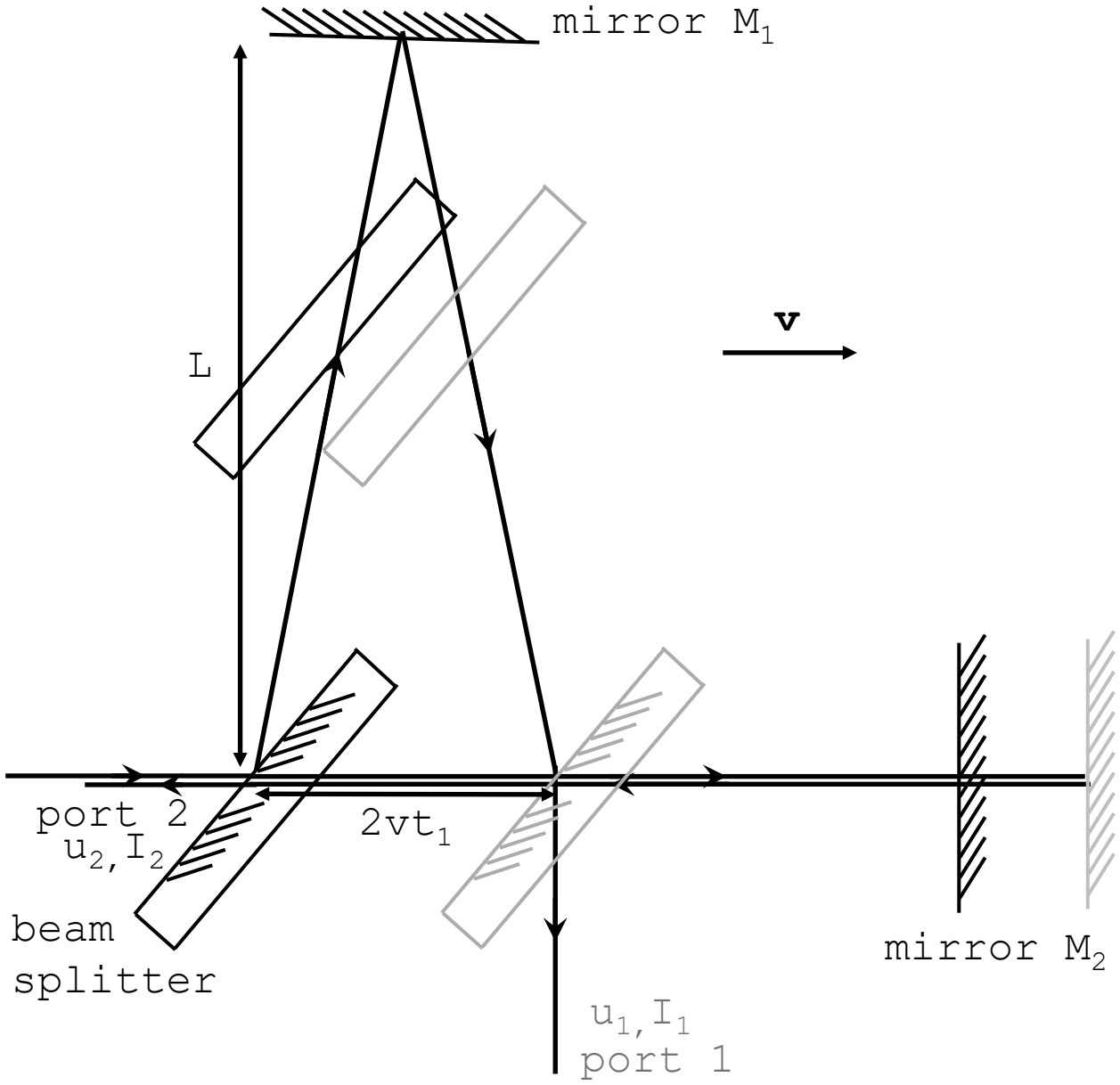

Figure 40.1: A demonstration of Michelson-Morley experiment. The grey apparatuses indicates the apparatuses travelled in time.

Note that the orbital speed of Earth about the Sun is $30\,000\,\mathrm{m\,s^{-1}}$, and therefore if there is an aether frame, the apparatus would detect a corresponding fringe shift. However nothing is found with the experiment repeated many times, which therefore motivated the theory of special relativity, a theory of space-time that predicts no aether frame. This is very inspiring, as it reveals that a null-experiment — an experiment with a negative outcome — motivates a new theory in physics; and therefore not only experiments with positive outcomes are important in physics.

Summary

1. The aether drift experiment is able to detect any measurable difference between the frame of the Earth and the frame of the aether with $v \geqslant 6\,000\,\mathrm{m\,s^{-1}}$, which is much smaller than the theoretical speed of the Earth with respect to the aether frame, if aether does exist. This is not experimentally detected, which suggests that there is no aether at all, and thus motivated the theory of special relativity.

§41. Haidinger (Equal Inclination) Fringes

Haidinger (Equal Inclination) Fringes

I have made a promise in §35 that we will have a look at a variety of uses of the Michelson interferometer, where the first variation of uses involves keeping d constant and just look at the output, but not only at exactly the output, but also at the small angles around it, and see whether we can see fringes.

To analyse this, we think about the mirror M_1 *as if it is* rotated about the beam splitter to the same axis as the mirror M_2, giving M_1', and port 1 rotated onto the same axis, giving port $1'$, demonstrated by figure 41.1. After this imaginary rotation, now the difference between distances from the beam splitter to M_1 and M_2, d, becomes the perpendicular distance from M_1' to M_2. We then view the fringes by focusing it with a lens with a focal length f, which means that the image will be formed at a distance $\rho = f\theta$ from the principal axis of the lens. Note that the setup has cylindrical symmetry: if we were to rotate the system about the normal axis, the physics would remain unchanged. Therefore, the lens actually focuses light onto **circles**, i. e. we shall see circular fringes, or **Haidinger Fringes**.

Note that by the geometry of the setup (which is sketched on figure 41.1), it is clear that the phase difference caused by this reflection for rays at a small inclination angle θ about the normal has an additional phase shift

$$\delta = 2kd\cos\theta, \tag{41.1}$$

which gives an intensity pattern

$$I(\theta) = I_0\left|1+\mathrm{e}^{\mathrm{i}\delta}\right|^2 = I_{\max}\cos^2(\delta/2) = I_{\max}\cos^2(kd\cos\theta). \tag{41.2}$$

Note that the intensity pattern is only dependent on the inclination angle θ, and therefore equal angles θ would correspond to equal intensity; therefore the fringes would trace out loci of equal inclination, which explains why sometimes Haidinger fringes are called "fringes of equal inclination".

By equating the optical path length difference with an integer number of wavelengths, we

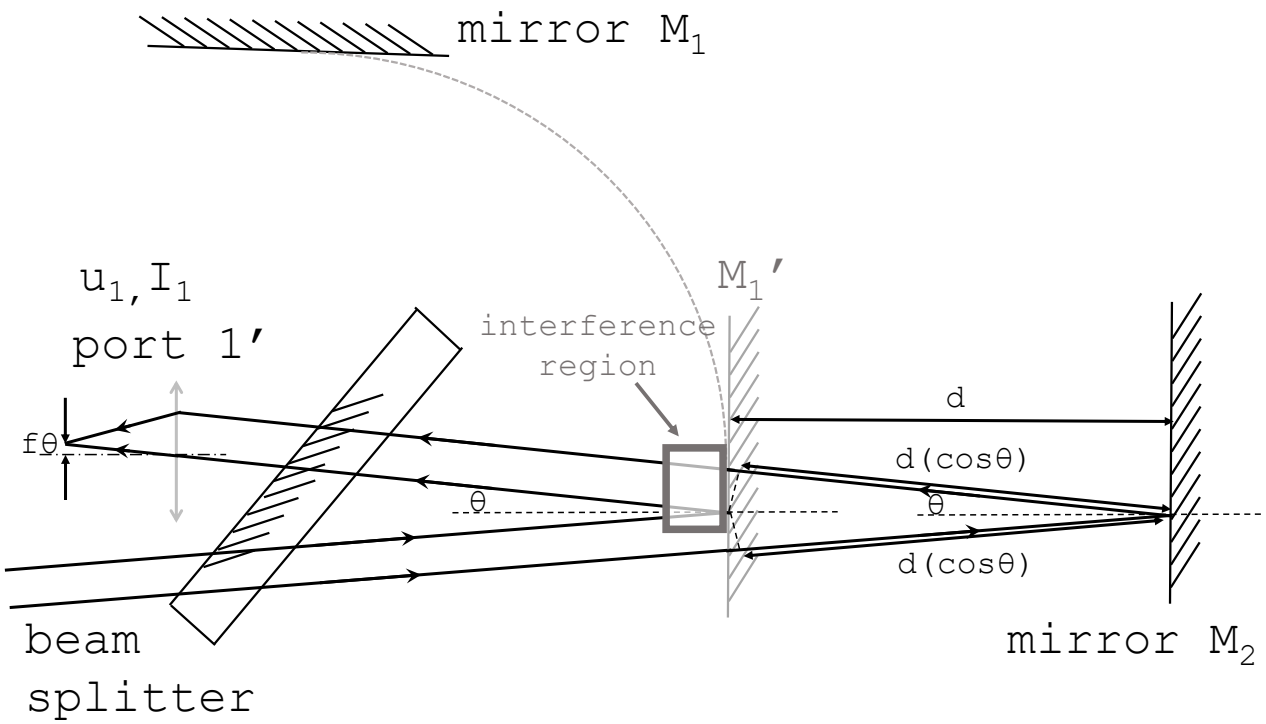

Figure 41.1: The setup that shows Haidinger fringes, or fringes of equal inclination.

have the location of maxima at

$$\Delta\text{OPL} \;=\; 2d\cos\theta_p \;\overset{!}{=}\; p\lambda \quad\Rightarrow\quad p\lambda \;=\; 2d\left(1-\frac{\theta_p^{\,2}}{2}\right) \;=\; 2d\left(1-\frac{\rho_p^{\,2}}{2f^2}\right), \tag{41.3}$$

where we assume that θ is small. This result can be easily verified by equation 41.2. Here ρ_p represents radii corresponding to bright fringes with order p. Note that in this setup, θ being close to 0 corresponds to the order p at a maximum instead of $p = 0$ in a transmission grating, as $\theta = 0$ means that the phase difference between the two interfering beams are at a maximum path difference $2d$, just by looking at the geometry of the setup.

Localisation of Haidinger Fringes

We have previously viewed Haidinger's fringes using a thin lens and we look at the image at the focal length of the thin lens f, which means that we assume that the image is formed at infinity. However we would like to find out whether we actually need to do that, which is illustrated by figure 41.2.

We note that, in order to see fringes of equal inclination, we would like to spatially separate light exiting towards different angles. For an extended light source, illustrated by the top figure of figure 41.2, this cannot be done unless we are at a very far distance away from the two mirrors compared to the size of the extended light source, otherwise the different angles are not spatially separated and are cluttered at the same spatial position. However for a single source, illustrated by the bottom figure of figure 41.2, different angles are separated spatially quite quickly after the light emerges from the two mirrors. As a result, we have to use a lens to focus at infinity for an extended light source, and this is not necessary for a single source: we can intercept the output light with a screen and see the fringes projected on the screen. The nomenclature that describes the type of fringes formed of an extended source is **fringes**

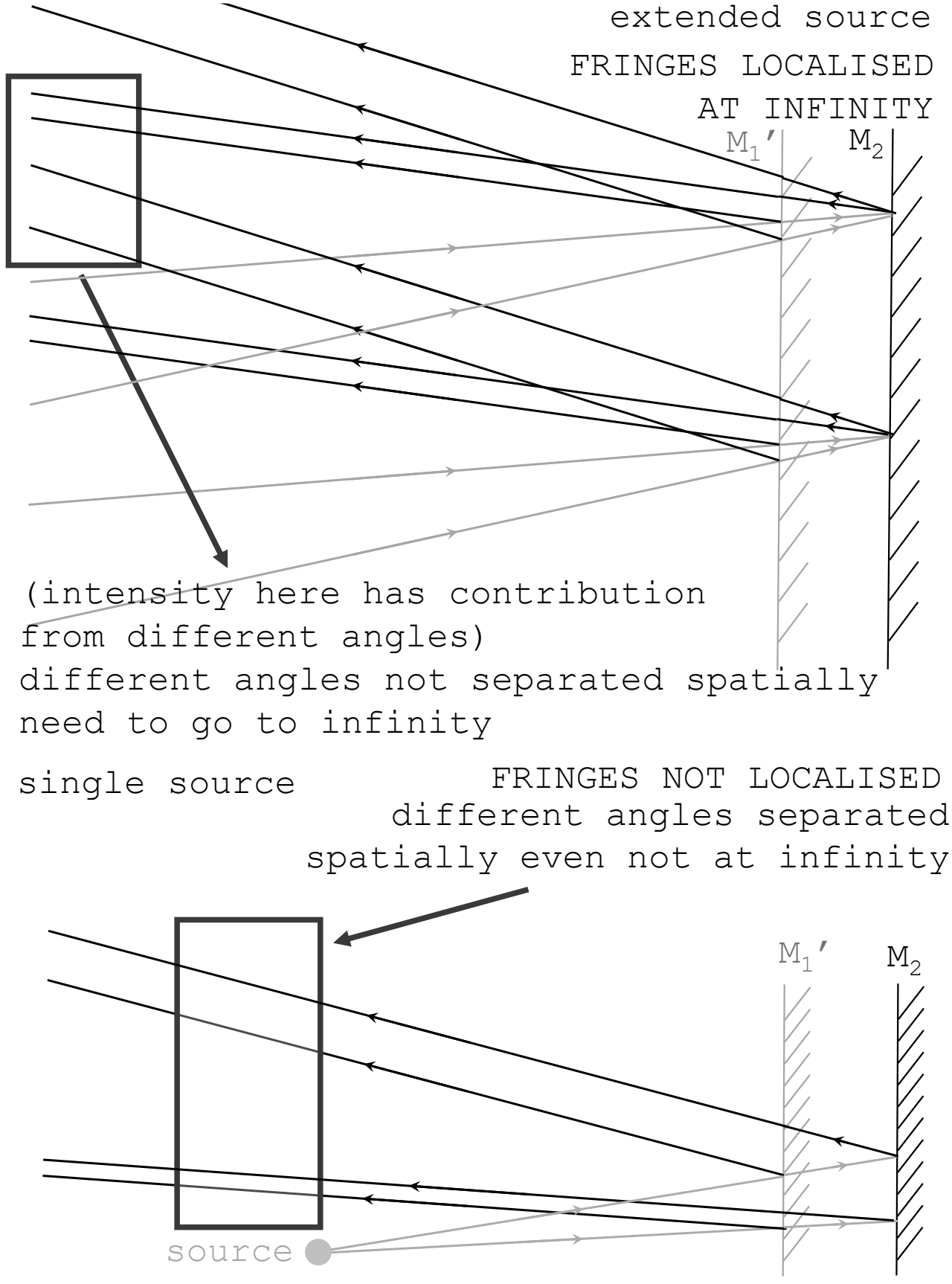

Figure 41.2: Localisation of Haidinger fringes for an extended source and a point source.

localised at infinity, and the nomenclature that describes the type of fringes formed of a point source is that the **fringes are not localised**.

Summary

1. We are able to see circular Haidinger fringes for a Michelson interferometer, where the intensity distribution is dependent on the angle of inclination only, and thus the name "fringes of equal inclination". The p^{th} order maxima has radii
$$p\lambda \;=\; 2d\left(1-\frac{\theta_p^{\,2}}{2}\right) \;=\; 2d\left(1-\frac{\rho_p^{\,2}}{2f^2}\right), \tag{41.4}$$
if the output is focussed by a lens with focal length f.
2. If the source that inputs into the Michelson interferometer is extended, then the fringes will be localised at infinity and has to be focussed by a lens with focal length f or be viewed at a very large distance compared to the size of the extended source. If the source is a point source, then the fringes will be not be localised, and can be viewed by intercepting the light with a screen.

§42. Fizeau (Equal Thickness) Fringes

Fizeau Fringes for Point Sources

We now set the Michelson interferometer to the third setup, where we set $d \;=\; 0$ and the mirrors wedged by a small angle Φ with respect to each other. Again the best way to think about this is to imagine that we rotate the mirror M_1 to M_1' such that we are viewing M_1 and M_2 together, and we also imagine that we rotate port 1 to port $1'$, so that we have everything on the same axis. This is demonstrated in figure 42.1.

Let us first investigate into the case where we have a point source P illuminating the two mirrors, at a distance D from the centre of the two wedged mirrors, where D takes into account of both the distance from the point source to the beam splitter and the distance from the beam splitter to the two mirrors. We shall now mirror image the point source by the wedged mirrors M_1' and M_2, and we find two imaginary sources P_1' and P_2. Just by tracing through the geometry (which takes a few steps), we find that the transverse distance between P_1' and P_2 is given by $2D\Phi$ — note that since Φ is very small, $2D\Phi \;\ll\; D$ in reality: figure 42.1 is a figure that exaggerates this distance greatly.

We now note that we have fringes from the interference from the two imaginary sources P_1' and P_2, which are fringes that emerges from a pair of slits. These fringes can be intercepted anywhere — they are not localised. For the ray emerging at a an angle θ, since the path difference between P_1' and P_2 is $2D\Phi\theta$ for small θs, we have the condition of maxima
$$p\lambda \;=\; 2D\Phi\theta, \tag{42.1}$$
where p is an integer. Note that since the mirrors are now wedged, the cylindrical symmetry in the Haidinger fringes setup is now lost, and therefore we see **straight** fringes. If we then look at the fringes at a distance $\mathscr{D}$ away, then we have the transverse distance x satisfying $\theta \;=\; x/\mathscr{D}$, and therefore the maxima are located at
$$x \;=\; \frac{\mathscr{D}}{2D\Phi}p\lambda. \tag{42.2}$$

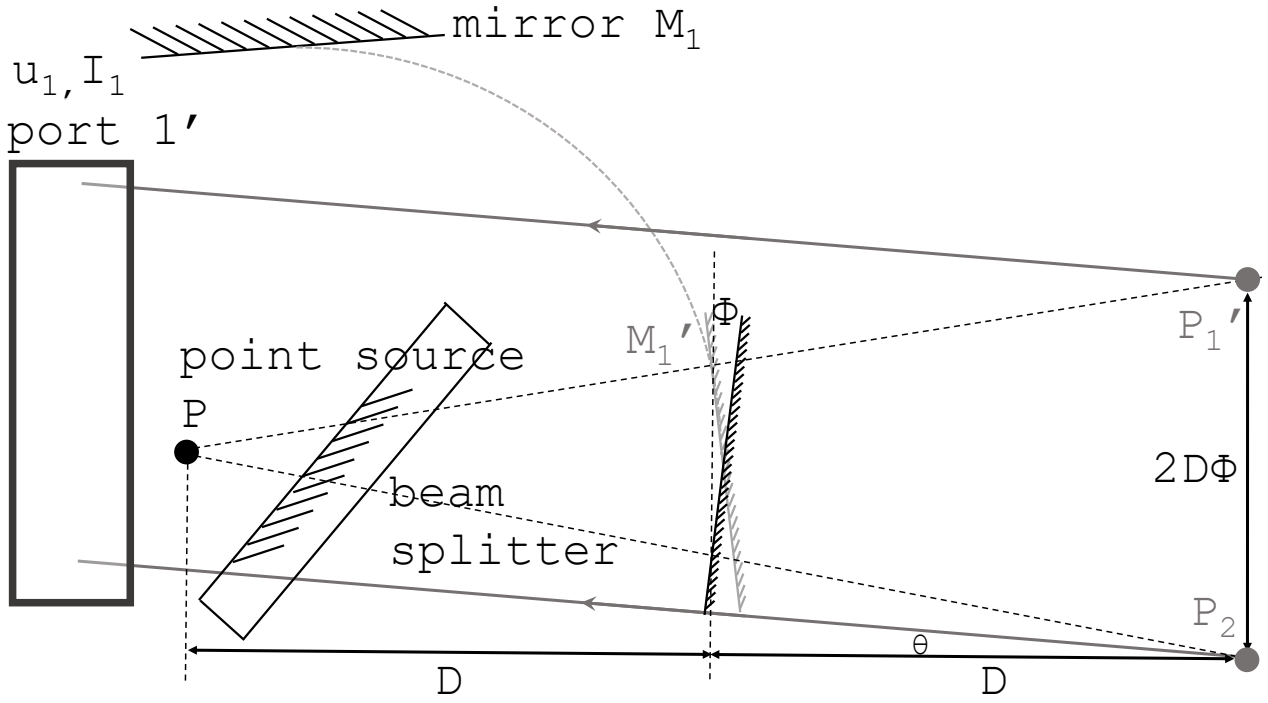

Figure 42.1: Generating Fizeau fringes for a point source.

Of course, if $\mathscr{D}$ is smaller than D, then we have apparatus in the way, and therefore it is physically difficult to view fringes there. The method to use is usually to select a lens that focuses on fringes at that distance, i. e. place a lens with a focal length f that is L away from the centre of mirrors M_1' and M_2, and look at the lens at a distance v away where

$$\frac{1}{L-\mathscr{D}}+\frac{1}{v} \;=\; \frac{1}{f}, \tag{42.3}$$

to see fringes at a distance $\mathscr{D}$ from the centre of M_1' and M_2.

Fizeau Fringes of Extended Sources, or Fringes of Equal Thickness

If we change the point sources into extended sources, like figure 42.2, then the imaginary sources will be large blocks that are not spatially separated. As a result, there is no way that we can see fringes that are not localised, or, in fact, whether we are able to see any fringes at all. However there is a solution to this problem, which is to consider the light from the extended source collimated and directed towards the wedged mirrors along the line that extends from the beam splitter to the middle of the two mirrors. We shall also assume that Φ is so small that the light will not be deflected from the normal upon reflection or refraction (which figure 42.3 has Φ too large, so this assumption certainly does not hold true and the light will certainly be deflected: the rays sketched in figure 42.3 is quite an exaggeration). Then, light at spatial distance x across the collimated beam from the centre of the wedged mirrors will have a difference in the optical path length $2\Phi x$ between the beam that is reflected upon M_1' and M_2 dependent on the transverse distance x. The two rays that are reflected off M_1' and M_2 then interferes, giving maxima at

$$p\lambda \;=\; 2\Phi x, \tag{42.4}$$

therefore we will still see straight fringes. Note that the condition is only dependent on $2\Phi x$, which are the distances between the wedged mirrors, or "thicknesses", and therefore each bright

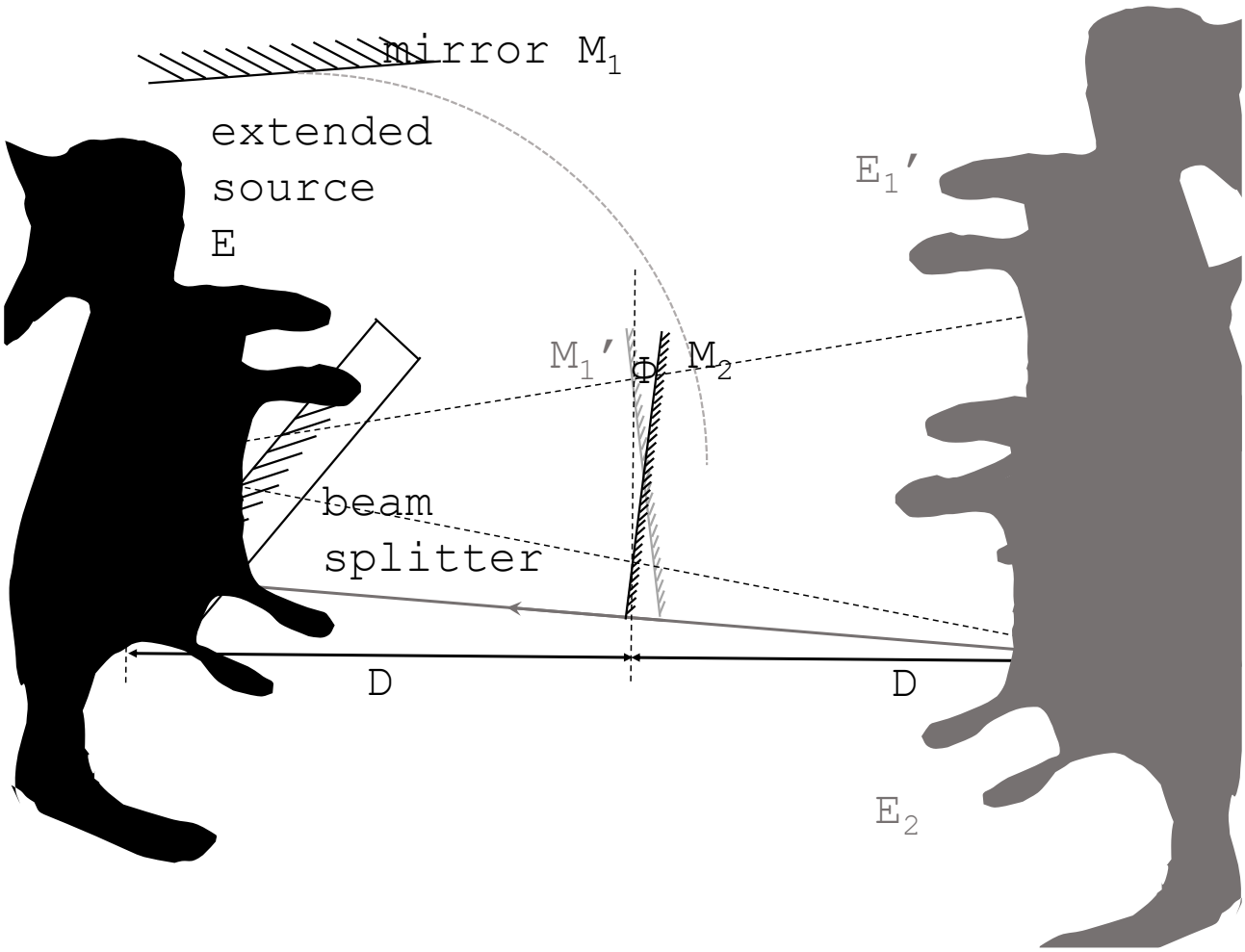

Figure 42.2: Inability of extended sources to create fringes that are not localised.

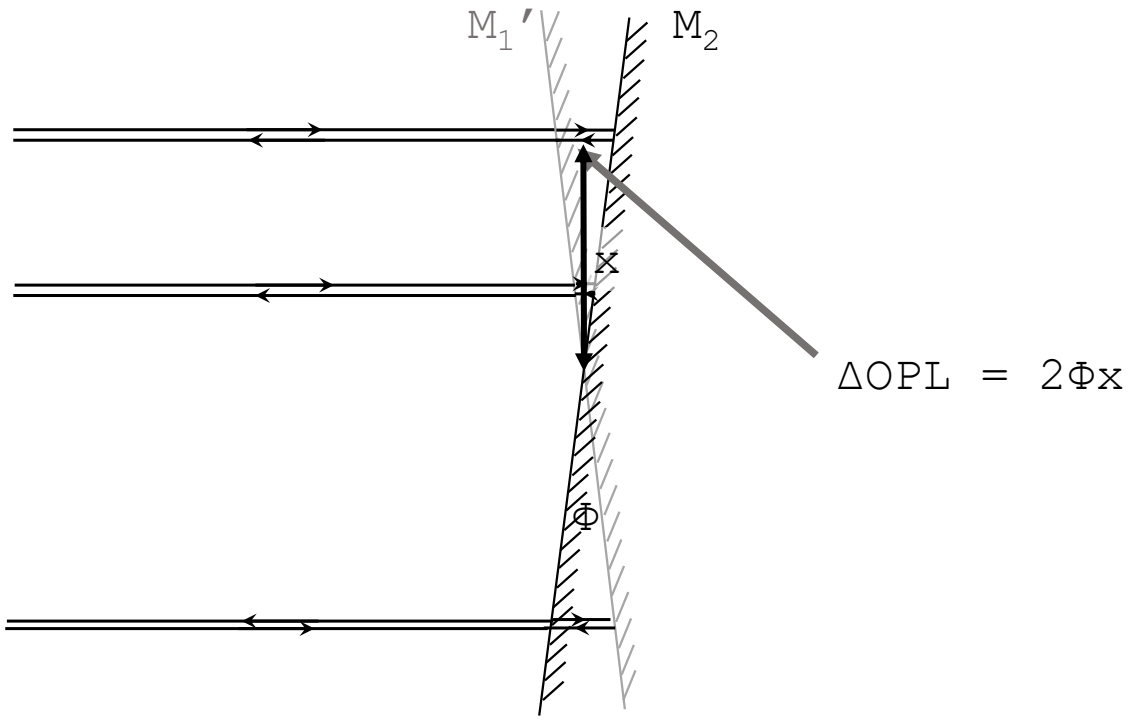

Figure 42.3: Fizeau fringes, or fringes of equal thickness, of an extended source.

fringe corresponds to sections on the wedged mirror with the same "thickness". That is why Fizeau fringes of this kind is also referred to as "fringes of equal thickness".

Finally we shall note that, in reality, no matter how small Φ is, there will always be a non-negligible deviation of the ray from the axis that links the beam splitter and the wedged mirrors. As a result, if we are intercepting the beam away from the mirrors, then we shall see no fringes at all, as fringes emerged at different thickness will mix together as they will travel out with different angles. This means we have to look at the beam *on the wedged mirrors*, which is achieved by placing a lens such that the lens images exactly the wedged mirrors. This can be verified as setting $D = \mathscr{D}$ in equation 42.2 gives exactly equation 42.4, so that the two-slit interference fringes and the fringes of equal thickness coincide — and they must be, as we have the interference between two beams originating from one point source, and hence it can only give one set of fringes. Thus, Fizeau fringes of an extended source is localised on the wedged mirrors.

Summary

We shall summarise the properties of the Fizeau fringes first, and then we shall pull the different types of fringes observable on the Michelson interferometer altogether.

1. Fizeau fringes for point sources are straight fringes that are not localised, same as two-slit interference fringes.

2. Fizeau fringes for extended sources are straight fringes that are localised on the wedged mirrors. They are also called fringes of equal thickness, which suggests that the fringes traces out equal distances between the wedged mirrors.

We now tabulate all the fringes observable on the Michelson interferometer.

	Haidinger fringes setup §41, figure 41.1	Fizeau fringes setup §42, figure 42.1
point source	not localised, circular	not localised, straight
extended source	localised at infinity, circular	localised at the wedged mirrors, straight

8 Fabry-Perot Etalon and Fabry-Perot Interferometer

§43. Fabry-Perot Etalon and Fabry-Perot Interferometer

The next device for the separation of wavelengths that we are going to look into is the Fabry-Perot etalon and the Fabry-Perot Interferometer, which is constructed with two very reflective mirrors with reflectivity (intensity of light reflected / intensity of light incident) $R = r^2$ and transmissivity (intensity of light transmitted / intensity of light incident) $T = t^2$ mounted parallel to each other at a distance d apart, with incident light shining on the interferometer at a small angle θ. If d is adjustable by, for example, a piezoelectric, then the apparatus is called a Fabry-Perot interferometer, otherwise if d is fixed then it is called a Fabry-Perot etalon. Here r and t are the change in scalar amplitudes when light is reflected or transmitted from the mirror, noting that some Fabry-Perot interferometers are air-spaced, so it can be possible that r carries a π phase change and therefore is negative.

The intensity detected by the Fabry-Perot arises from the interference between the infinite number of light rays transmitted through the etalon, as illustrated in figure 43.1. Again, we have cylindrical symmetry, and as a result, we should see circular fringes. If the input is an extended source, then the image is formed at infinity, exactly the same as Haidinger's fringes in a Michelson interferometer. We shall now work out the intensity distribution as a function of θ. Note that, as discussed in §41, the phase change per round-trip between the mirrors is given by $\delta = 2kd\cos\theta = 2nk_0 d\cos\theta$, where k_0 is the wavenumber of the light in vacuum and n is the refractive index of the material between the mirrors. As a result, the transmitted scalar amplitude is given as

$$u = u_0 t^2 \mathrm{e}^{\mathrm{i}\delta} \sum_{m=0}^{\infty} \left(r^2 \mathrm{e}^{\mathrm{i}\delta}\right)^m = u_0 t^2 \times \frac{1}{1 - r^2 \mathrm{e}^{\mathrm{i}\delta}}. \tag{43.1}$$

Using $I = u^* u$, this gives an intensity

$$I = I_{\max} \times \frac{1}{1 + \left[\dfrac{2\mathscr{F}}{\pi} \sin\left(\dfrac{\delta}{2}\right)\right]^2}, \qquad \mathscr{F} = \frac{\pi\sqrt{R}}{1-R}, \qquad I_{\max} = u_0{}^* u_0 \left(\frac{T}{1-R}\right)^2. \tag{43.2}$$

The graph of the intensity I against the phase change per round trip δ is shown in figure 43.2. Note that, if the mirrors are ideal and have no losses i. e. $R + T = 1$, then $I_{\max} = u_0{}^* u_0$. In reality this is impossible and a large fraction of light is absorbed by the mirrors, where the energy carried by the electromagnetic radiation is transferred into heat. $\mathscr{F}$ is called the **finesse** of the interferometer and is dependent on the quality of the mirrors used, and the larger the finesse, the sharper the fringes. This is demonstrated in figure 43.2.

Summary

1. The Fabry-Perot etalon or Fabry-Perot interferometer is made of two highly reflective mirrors with high reflectivity. If an extended source is used to generate the image, then we will see circular fringes localised at infinity. The finesse $\mathscr{F} = \pi\sqrt{R}/(1-R)$ represents the sharpness of the fringes, where the larger the finesse, the sharper the fringes.

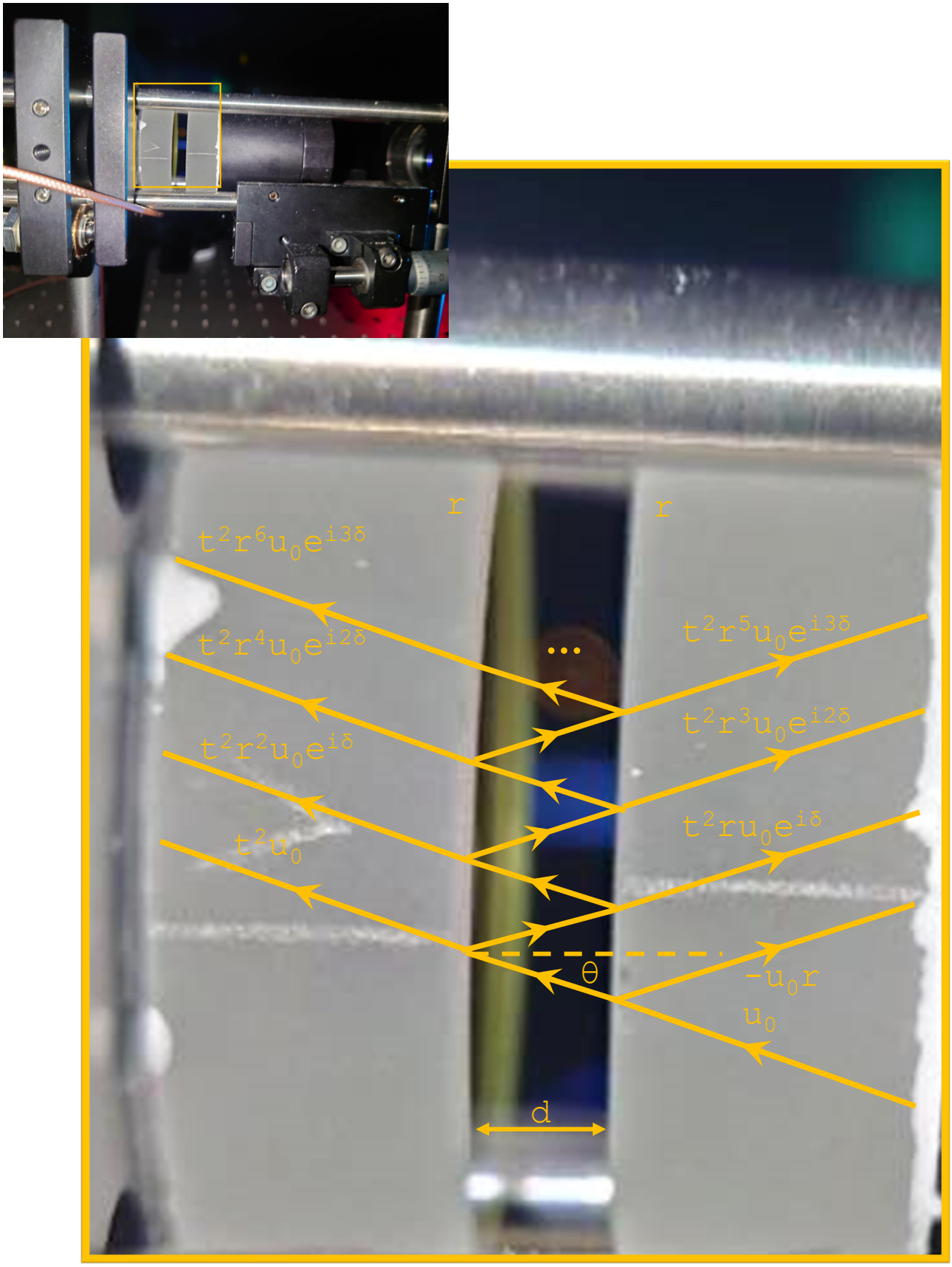

Figure 43.1: A Fabry-Perot interferometer.

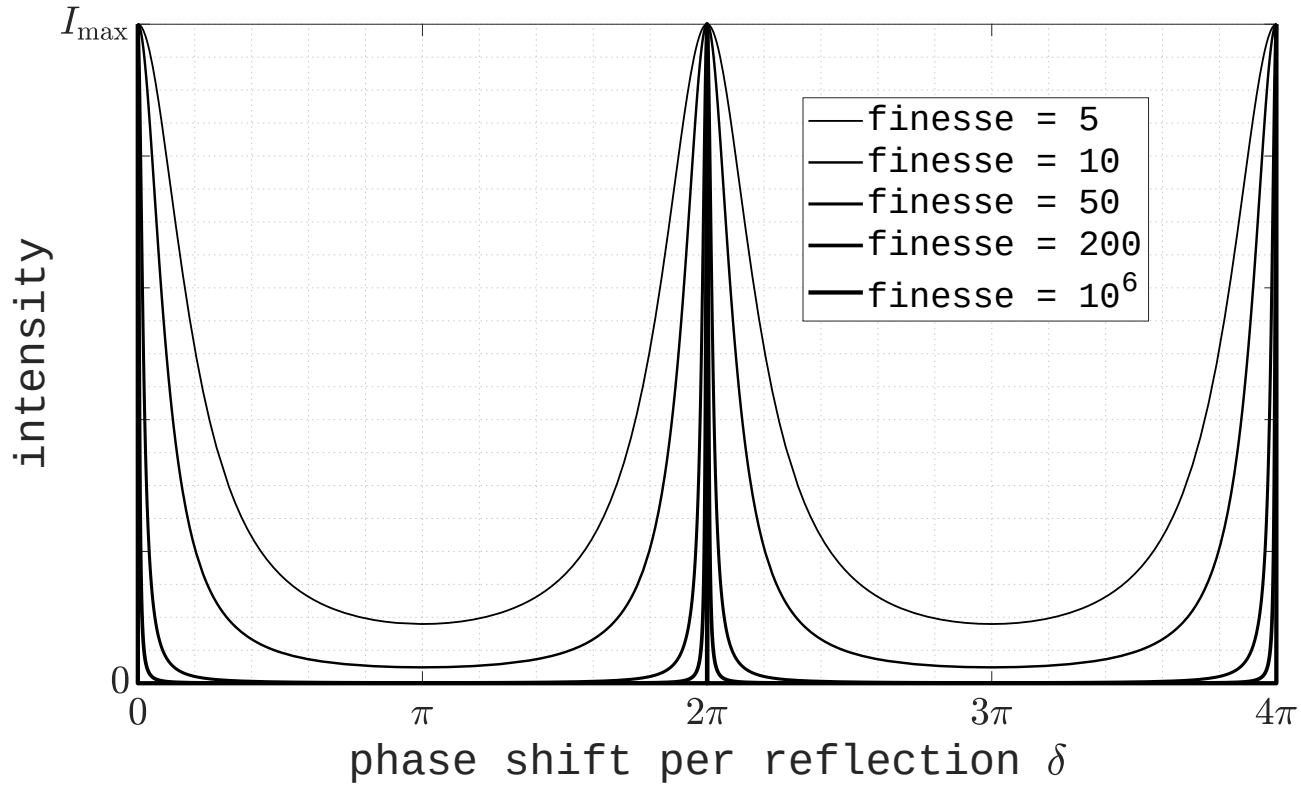

Figure 43.2: The intensity of a Fabry-Perot interferometer against the phase change per round-trip.

§44. Analysis of the Fringes of a Fabry-Perot Etalon

Radii of Fringes

We are interested in the radii of fringes of the Fabry-Perot etalon. Let us use the wavelength in the material between the mirrors here instead of the wavelength in a vacuum, then for the formation of a bright fringe, the geometrical path length must equate to an integer number p times the wavelength. Also let us focus the fringes onto a detector using a lens with a focal length f. This gives the condition on the angle θ_p and the radii ρ_p of the fringes on the detector as

$$2d\cos\theta_p \;=\; 2d\left(1-\frac{\theta_p^{\,2}}{2}\right) \;=\; 2d\left(1-\frac{\rho_p^{\,2}}{2f^2}\right) \;=\; p\lambda \tag{44.1}$$

for small θ_ps, i. e. located at exactly the same place at the Haidinger's fringes of a Michelson interferometer. This can be verified by minimising the denominator of equation 43.2, which means setting $\delta/2 \;=\; p\pi$, giving exactly the same condition. Note that again, same as Haidinger's fringes, p is maximised near $\theta \;=\; 0$, instead of minimised as in a diffraction grating. Also when θ_p is small, then $\cos\theta_p \;=\; 1$, and hence usually $p\lambda \;=\; 2d$ is a good approximation that we shall utilise in the calculations involving Fabry-Perot etalons. The plot of the intensity against the radius and the squared radius is shown in figure 44.1. The circular fringes seen in a Fabry-Perot etalon is shown in figure 44.2.

Separation of Wavelengths

Let us now attempt to resolve two very close wavelengths λ and Λ from an image , where light with wavelength λ has a maximum with p^{th} order at radius ρ_p and the light with wavelength Λ has a maximum with p^{th} order at radius r_p. This is illustrated in Figure 44.3, and a computer

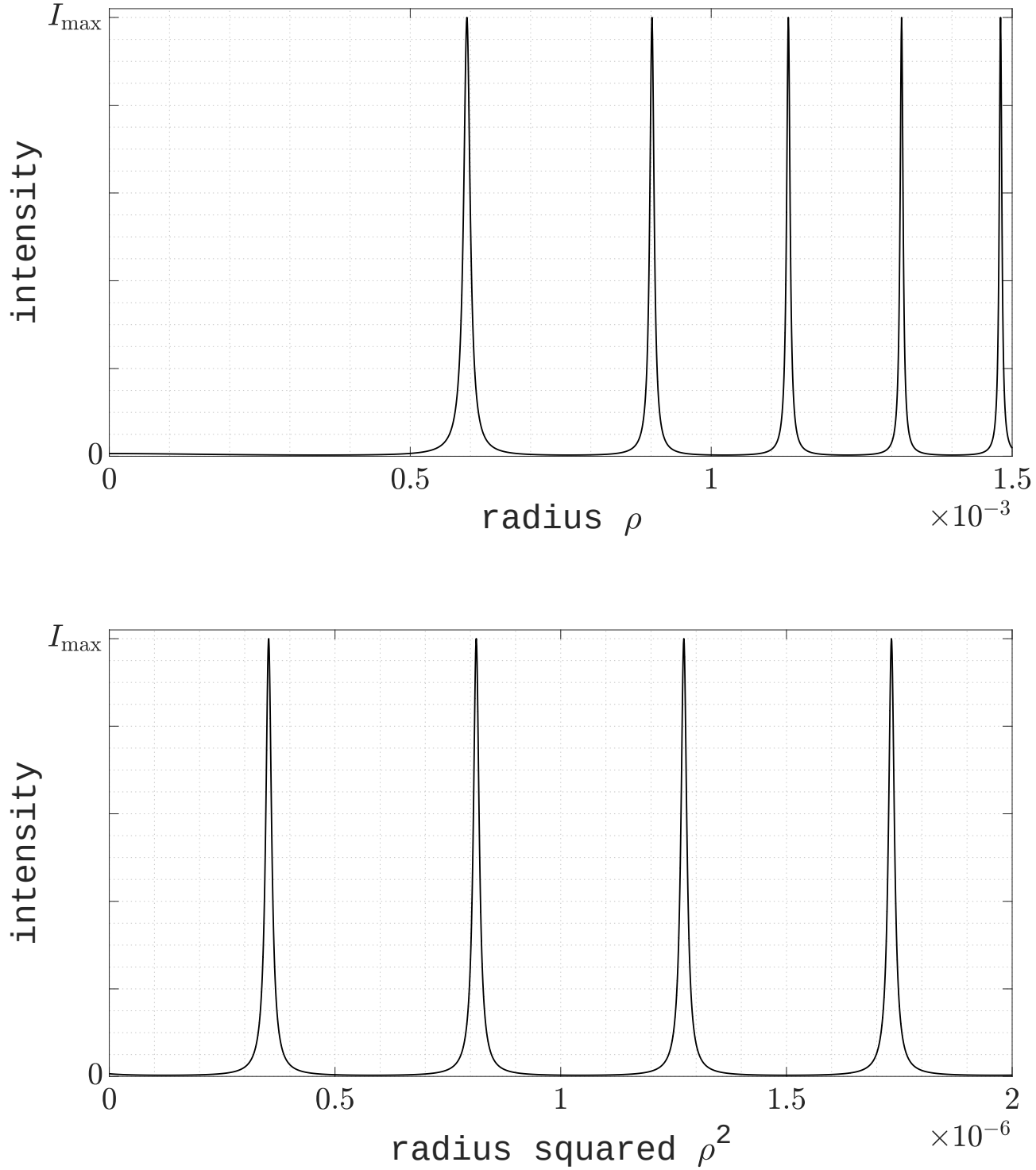

Figure 44.1: The intensity of the image on the detector against radius ρ and squared radius ρ^2. Here we select $\mathscr{F} = 30$, $d = 7\,\text{mm}$, $f = 0.1\,\text{m}$, and $\lambda = 6\,438\,\text{Å}$.

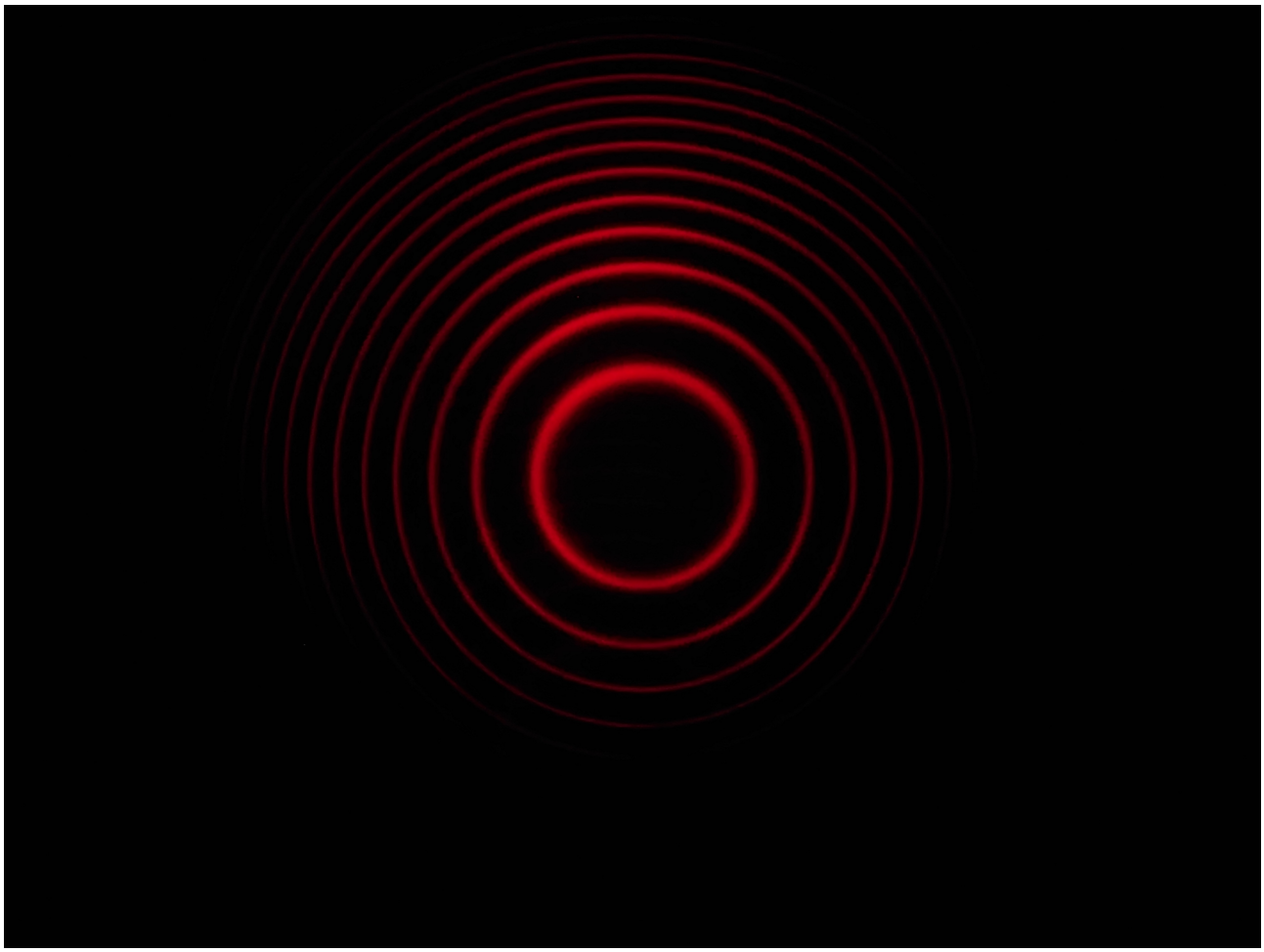

Figure 44.2: The circular fringes in a Fabry-Perot etalon.

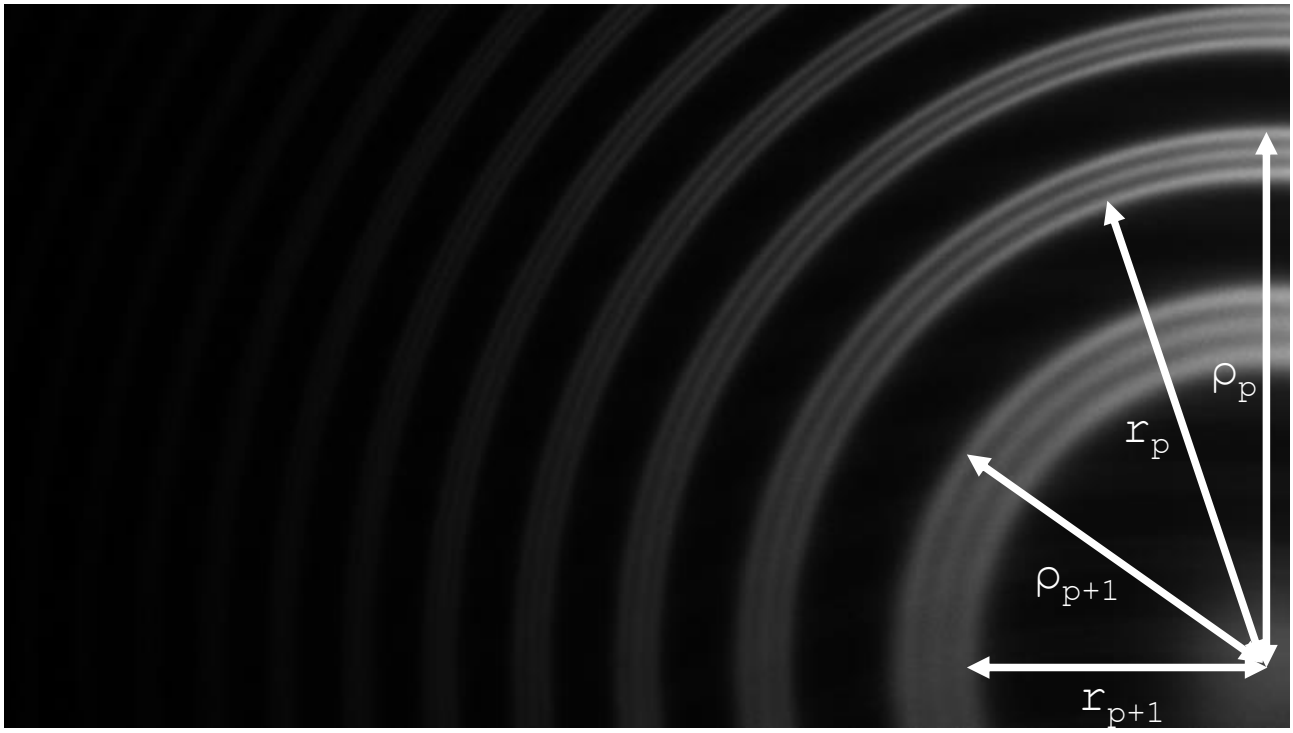

Figure 44.3: Reading off radii from the Fabry-Perot etalon. This image shows the wavelengths corresponding to the Zeeman splitting of the cadmium red ($5^1D_2 \to 5^1P_1$) transition under a magnetic field. The middle ring has a wavelength of 6 438 Å. As we are only discussing separating two wavelengths instead of three, we shall only use the rings corresponding to the lowest and highest wavelengths.

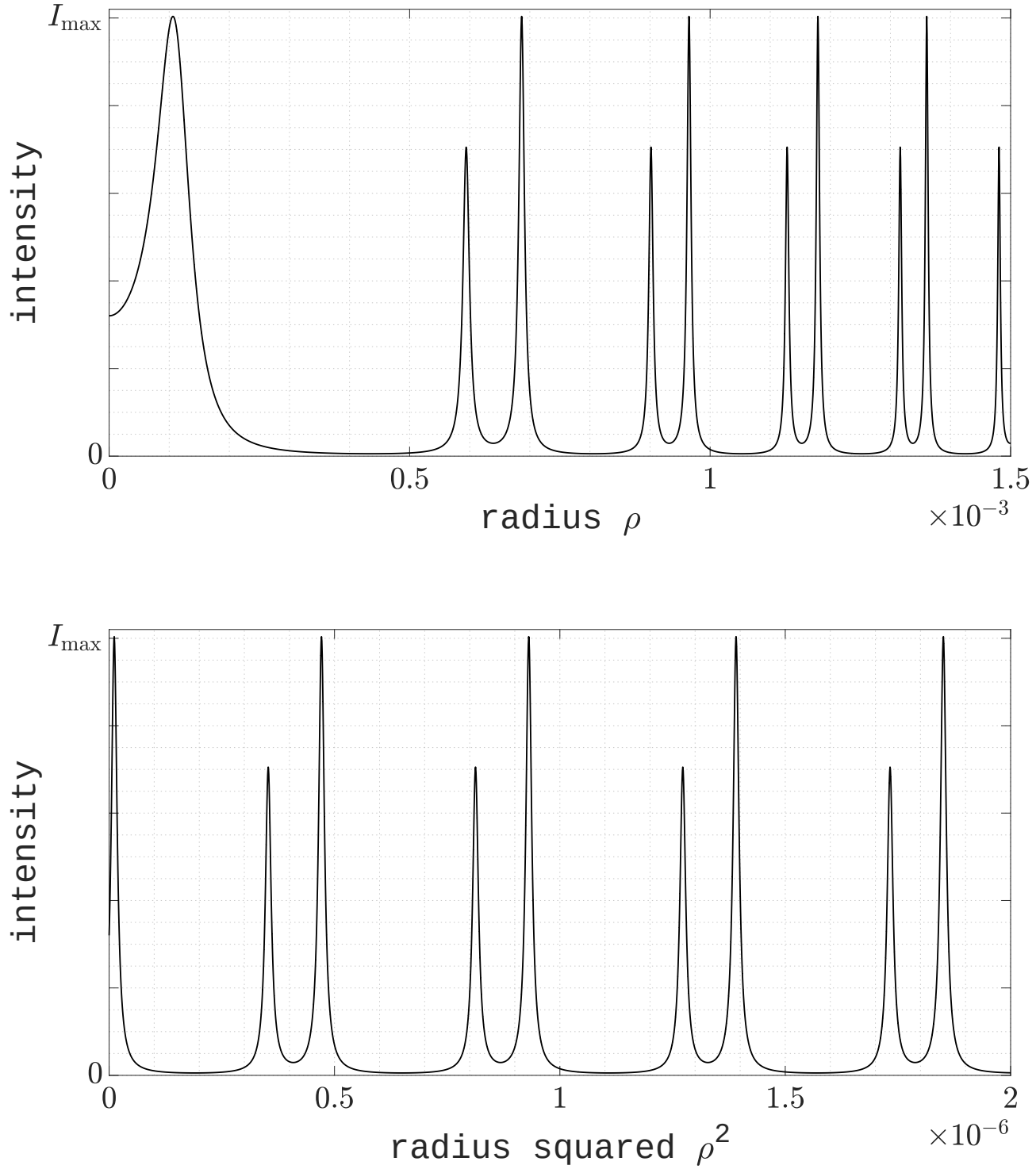

Figure 44.4: The intensity of the image consisted of a source with two wavelengths on the detector against radius ρ and squared radius ρ^2. Here we select $\mathscr{F} = 30$, $d = 7\,\text{mm}$, $f = 0.1\,\text{m}$, and $\lambda = 6\,438\,\text{Å}$ and $\lambda = 6\,438.11\,\text{Å}$.

simulation of the intensity distribution of a source with two wavelengths is illustrated in figure 44.4. From the image, we read off the radii of the p^{th} and $(p+1)^{\text{th}}$ order. Without loss of generality let us set $\Lambda > \lambda$. Our goal would be to find the difference in wavenumber

$$\Delta\bar{\nu} \;=\; 1/\lambda - 1/\Lambda \tag{44.2}$$

between λ and Λ. Using equation 44.1, we are able to write down the following equations for the p^{th} and $(p+1)^{\text{th}}$ order maxima

$$2d\left(1 - \frac{{\rho_p}^2}{2f^2}\right) \;=\; p\lambda, \tag{44.3}$$

$$2d\left(1 - \frac{{r_p}^2}{2f^2}\right) \;=\; p\Lambda, \tag{44.4}$$

$$2d\left(1 - \frac{{\rho_{p+1}}^2}{2f^2}\right) \;=\; (p+1)\lambda, \tag{44.5}$$

$$2d\left(1 - \frac{{r_{p+1}}^2}{2f^2}\right) \;=\; (p+1)\Lambda. \tag{44.6}$$

Rearranging the first three equations, we have

$$\frac{\text{eqn. 44.3} - \text{eqn. 44.4}}{\text{eqn. 44.3} - \text{eqn. 44.5}} \;=\; \frac{{\rho_p}^2 - {r_p}^2}{{\rho_p}^2 - {\rho_{p+1}}^2} \;\overset{!}{=}\; \frac{p(\Lambda - \lambda)}{\lambda}. \tag{44.7}$$

Multiplying both sides with $1/\Lambda$ and utilising equation 44.2 and the approximation $p\Lambda \;=\; 2d$ gives

$$\Delta\bar{\nu} \;=\; \frac{1}{2d} \times \frac{{\rho_p}^2 - {r_p}^2}{{\rho_p}^2 - {\rho_{p+1}}^2}. \tag{44.8}$$

Therefore we have successfully found the difference in wavenumbers of the two colours that we are looking for. However there are further complications, which will be discussed in the next section.

Summary

1. The location of p^{th} maxima is at radii ρ_p

$$2d\left(1 - \frac{{\rho_p}^2}{2f^2}\right) \;=\; p\lambda, \tag{44.9}$$

   where the order p is maximised when ρ_p is closest to 0. For small ρ_ps, we can approximate $p\lambda \;=\; 2d$.

2. To find the difference in wavenumbers of the two wavelengths sent into the Fabry-Perot interferometer, we first locate the p^{th} and $(p+1)^{\text{th}}$ maxima and find their radii. Then we can find the difference using the relation

$$\Delta\bar{\nu} \;=\; \frac{1}{2d} \times \frac{{\rho_p}^2 - {r_p}^2}{{\rho_p}^2 - {\rho_{p+1}}^2}. \tag{44.10}$$

§45. Instrumental Range and Width of a Fabry-Perot Etalon

We are now able to find the difference in wavenumbers of the two wavelengths that we send into the Fabry-Perot etalon. However note that, similar to a grating, there are upper and lower limits of $\Delta\bar{\nu}$ that we can input for determination of $\Delta\bar{\nu}$, which we shall discuss now.

Instrumental Range of a Fabry-Perot Etalon

Our master equation 44.8 can only work if we are certain that the rings that are taken from figure 44.3 are with orders p and $p+1$. We are only sure about this if $\Delta\bar{\nu}$ is not too large, so that we keep the radius r_p larger than ρ_{p+1}. This means that the largest resolvable wavenumber, or the instrumental range (free spectral range), is given by setting $r_p = \rho_{p+1}$ in equation 44.8, i. e.

$$\mathrm{FSR}_{\bar{\nu}} = 1/(2d). \tag{45.1}$$

Note that even if we have $\Delta\bar{\nu} < \mathrm{FSR}_{\bar{\nu}}$, there is still an ambiguity. We have previously stated that we assume $\Lambda > \lambda$ without loss of generality, yet when we are looking at the fringes, we are unable to locate whether the rings that we identify as λ and Λ has this condition $\Lambda > \lambda$. If it is indeed the other way round, then r_p labelled in figure 44.3 is in fact r_{p+1}, as $r_{p+1} > \rho_{p+1}$ if $\Lambda < \lambda$. In this case, we have

$$\frac{\text{eqn. 44.6} - \text{eqn. 44.5}}{\text{eqn. 44.3} - \text{eqn. 44.5}} = \frac{{\rho_{p+1}}^2 - {r_{p+1}}^2}{{\rho_{p+1}}^2 - \rho_p^2} \overset{!}{=} \frac{(p+1)(\lambda-\Lambda)}{\lambda}. \tag{45.2}$$

Then, noting that since p is large, $p \approx p+1$, and therefore we may use the approximation $(p+1)\Lambda = 2d$, and hence

$$\Delta\bar{\nu}' = \frac{1}{2d} \times \frac{{r_{p+1}}^2 - {\rho_{p+1}}^2}{\rho_p^2 - {\rho_{p+1}}^2}. \tag{45.3}$$

To check whether the difference in wavenumbers is $\Delta\bar{\nu}$ or $\Delta\bar{\nu}'$, we will need to change the distance between the mirrors d, and see which one of the wavenumber difference can be retrieved.

Instrumental Width and Chromatic Resolving Power of a Fabry-Perot Etalon

We shall now look into the lowest difference in $\bar{\nu}$ we can distinguish, i. e. the instrumental width $\mathrm{INST}_{\bar{\nu}}$. To find that we recall the Rayleigh criterion: the two wavelengths are indistinguishable if their separation is smaller than the width of each individual peak. To find the width of each peak, we need to find the full-width at half-maxima of the intensity distribution, equation 43.2

$$I = I_{\max} \times \frac{1}{1 + \left[\dfrac{2\mathscr{F}}{\pi}\sin\left(\dfrac{\delta}{2}\right)\right]^2}. \tag{45.4}$$

For $I = I_{\max}/2$, it is clear that we require

$$\frac{2\mathscr{F}}{\pi}\sin\left(\frac{\delta}{2}\right) = 1. \tag{45.5}$$

We then set δ close to a peak, i. e. $\delta = 2p\pi + \mathrm{INST}_\delta/2$, or $\delta/2 = p\pi + \mathrm{INST}_\delta/4$, where p is an integer, then we expand equation 45.5 about $\delta/2 = p\pi$ to first order, giving

$$\frac{2\mathscr{F}}{\pi}\sin(p\pi) + \left\{\frac{\mathrm{d}}{\mathrm{d}(\delta/2)}\left[\frac{2\mathscr{F}}{\pi}\sin\left(\frac{\delta}{2}\right)\right]\right\}_{\delta/2\,=\,p\pi} \times \frac{\mathrm{INST}_\delta}{4} = \frac{\mathscr{F}}{2\pi} \times \mathrm{INST}_\delta \overset{!}{=} 1, \tag{45.6}$$

giving $\text{INST}_\delta = 2\pi/\mathscr{F}$. Then, since $\delta = 2kd\cos\theta = 4\pi\bar{\nu}d\cos\theta$ is linear on $\bar{\nu}$, we have

$$\frac{\text{FSR}_{\bar{\nu}}}{\text{INST}_{\bar{\nu}}} = \frac{\text{FSR}_\delta}{\text{INST}_\delta} = \mathscr{F} \quad \Rightarrow \quad \text{INST}_{\bar{\nu}} = \frac{1}{2\mathscr{F}d}. \tag{45.7}$$

Here we have used the fact that the when the phase difference between the two wavelengths is FSR_δ, the difference in δ is too large such that the two rings are crossing each other, and therefore we have a phase difference of 2π, i. e. $\text{FSR}_\delta = 2\pi$. With this, the chromatic resolving power is given as

$$\mathscr{P} = \frac{\bar{\nu}}{\text{INST}_{\bar{\nu}}} = 2\mathscr{F}d\bar{\nu} = \mathscr{F}p, \tag{45.8}$$

using the approximation $2d = p\lambda = p/\bar{\nu}$ again. To illustrate the power of a Fabry-Perot etalon, we compare the etalon with a grating, which has a chromatic resolving power of Np, with N in the range of 10^3. Note that very good etalons also have $\mathscr{F}$ in the range of 10^3 so they draw on this battle. However note that the grating has a maximum diffraction order of order unity, but an etalon can have a diffraction order $p = 2d/\lambda$ on the magnitude of 10^6. Therefore an etalon has a much higher refraction power.

Finally, from the relation

$$\text{FSR}_\delta/\text{INST}_\delta = \mathscr{F}, \tag{45.9}$$

we have an interesting and sometimes handy interpretation of the finesse: it is the ratio of the separation between consecutive maxima and the width of a single maxima in figure 43.2.

Summary

1. A Fabry-Perot etalon has an instrumental range $\text{FSR}_{\bar{\nu}} = 1/(2d)$. Even if we have $\Delta\bar{\nu} < \text{FSR}_{\bar{\nu}}$ we could still have an ambiguity as we are unsure of the wavelengths corresponding to the two set of rings. The other wavenumber separation is given as

$$\Delta\bar{\nu}' = \frac{1}{2d} \times \frac{{r_{p+1}}^2 - {\rho_{p+1}}^2}{\rho_p^2 - {\rho_{p+1}}^2}. \tag{45.10}$$

   To resolve this ambiguity we need to change the distance between mirrors d.

2. A Fabry-Perot etalon has an instrumental range $\text{INST}_{\bar{\nu}} = 1/(2\mathscr{F}d)$, and a resolving power $\mathscr{F}p$, which is much larger than a grating.

§46. Linearisation of the Intensity Pattern of a Fabry-Perot Etalon

Linearisation of the Intensity Pattern of a Fabry-Perot Etalon

We shall note that, from figure 44.1 and 44.4, the intensity of a Fabry-Perot is not linear on the radius ρ. However we shall note that it is linear on ρ^2. Therefore it makes a lot of sense to analyse the pattern with respect to ρ^2 instead of ρ; or, since $\rho = \theta/f$, we may analyse the pattern in θ^2. Since we have the condition in maximum

$$p\lambda \Rightarrow 2d\cos\theta_p \quad \Rightarrow \quad \bar{\nu} = \frac{p}{2d}\sec\theta, \tag{46.1}$$

upon expansion of $\sec\theta$, we yield

$$\bar{\nu} = \frac{p}{2d}\left(1 + \frac{1}{2}{\theta_p}^2\right). \tag{46.2}$$

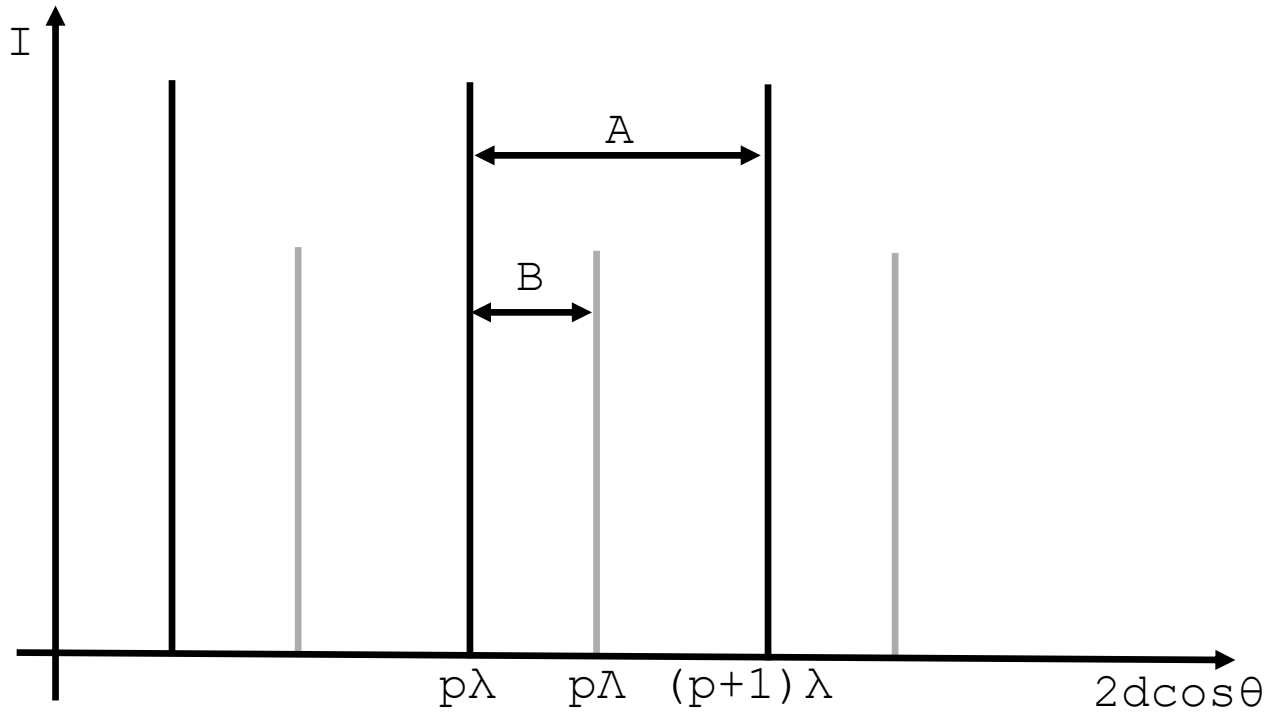

Figure 46.1: The intensity plot of the Fabry-Perot etalon of two distinct wavelengths λ and Λ against $\delta \;=\; 2d\cos\theta$.

We therefore note that

$$\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}(\theta_p^2)} \;=\; \frac{p}{2d}, \tag{46.3}$$

i. e. the wavenumber $\bar{\nu}$ is linear on θ_p^2. This is called the **dispersion relation**.

Using the dispersion relation, we are able to calculate the free spectral range in another way. Since we can calculate $\bar{\nu}$ from both the p^{th} and $(p+1)^{\text{th}}$ order fringe, we have, from equation 46.2,

$$4d\bar{\nu} \;=\; p\big[2+(\theta_p)^2\big] \;=\; (p+1)\big[2+(\theta_{p+1})^2\big], \tag{46.4}$$

giving

$$\mathrm{FSR}_{\theta^2} \;=\; (\theta_p)^2-(\theta_{p+1})^2 \;=\; 2/p \tag{46.5}$$

in the limit where $p \;\gg\; 1$. Therefore, just by seeing the FSR_{θ^2} as an error on θ_p, and using the error correction formula to propagate this error to $\bar{\nu}$, we have

$$\mathrm{FSR}_{\bar{\nu}} \;=\; \frac{\mathrm{d}\bar{\nu}}{\mathrm{d}(\theta_p^2)} \times \mathrm{FSR}_{\theta^2} \;=\; \frac{p}{4d} \times \frac{2}{p} \;=\; \frac{1}{2d}. \tag{46.6}$$

This equation also gives us an alternative way of analysing the pattern originated from two wavelengths from a Fabry-Perot etalon. If we have an intensity plot with respect to the phase $2d\cos\theta$, then we are able to measure the difference in phase between two orders and between two wavelengths within the same order. Let us label them A and B respectively. Then, we have

$$\frac{B}{A} \;=\; \frac{p(\Lambda-\lambda)}{(p+1)\lambda-p\lambda} \;=\; \frac{p(\Lambda-\lambda)}{\lambda}. \tag{46.7}$$

Then, using the relationship $\Delta\bar{\nu} = 1/\lambda - 1/\Lambda$ and $p\Lambda = 1/(2d) = \text{FSR}_{\bar{\nu}}$, by dividing through Λ, we yield

$$\Delta\bar{\nu} = (B/A) \times \text{FSR}_{\bar{\nu}}, \tag{46.8}$$

which is yet another method that allows us to find the difference in wavenumbers of the two wavelengths.

Instrumental Range as the Smallest Wavenumber Allowed for a Standing Wave

We shall note that the instrumental range

$$\text{FSR}_{\bar{\nu}} = 1/(2d) \tag{46.9}$$

is the reciprocal of twice the distance between mirrors. Note that, if we are to fix a standing wave between the two mirrors, then the allowed wavelengths is

$$\lambda_p = 1/(2dp), \tag{46.10}$$

and therefore the allowed wavenumber is

$$\bar{\nu}_p = 1/\lambda_p = p \times 2d = p \times \text{FSR}_{\bar{\nu}}, \tag{46.11}$$

i. e. the instrumental range is the smallest wavenumber allowed for a standing wave between the two mirrors, and any higher order standing wave must have a wavenumber as an integer number of the instrumental range.

Summary

1. The intensity pattern of the Fabry-Perot etalon can be linearised if we plot the intensity with respect to θ_p^2. This allows us to develop an alternative derivation for the expression of the instrumental range of the Fabry-Perot etalon.

2. The instrumental range is the smallest wavenumber allowed for a standing wave between the two mirrors, and any higher order standing wave must have a wavenumber as an integer number of the instrumental range.

§47. Cavity Round-Trip in a Fabry-Perot Etalon

Light as a Self-Consistent Steady State

An alternative method of using the Fabry-Perot is a "trap" for a light ray, containing a circulating light inside the two mirrors. We then equip the scalar amplitude of the light at different parts of the cavity as u_1, u_2, u_3, and u_4, as denoted in figure 47.1. The four scalar amplitudes are then

$$u_1 = tu_0 + ru_4, \quad u_2 = u_1 e^{i\delta/2}, \quad u_3 = ru_2, \quad u_4 = u_3 e^{i\delta/2}. \tag{47.1}$$

Substituting everything into the equation for u_1, we have

$$u_1 = tu_0 + r^2 e^{i\delta} u_1 \qquad \Rightarrow \qquad u_1 = \frac{tu_0}{1 - r^2 e^{i\delta}} \qquad \Rightarrow \qquad u = tu_2 = \frac{u_0 t^2 e^{i\delta}}{1 - r^2 e^{i\delta}}, \tag{47.2}$$

agreeing with equation 43.1 up to an overall phase.

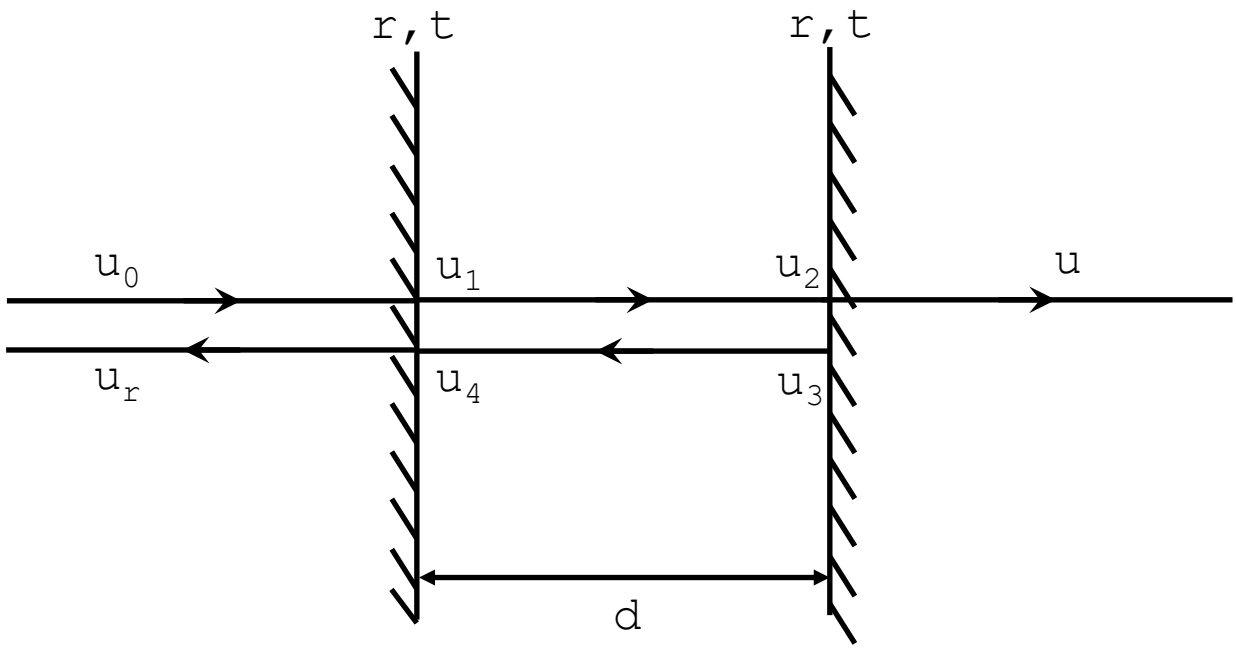

Figure 47.1: Light in a self-consistent steady state.

Reflected Intensity and Energy Conservation

We are also able to find the reflected scalar amplitude u_r, shown in figure 47.1. This is given as

$$u_r \;=\; -ru_0 + tu_4 \;=\; -ru_0 + \frac{u_0 r t^2 \mathrm{e}^{\mathrm{i}\delta}}{1 - r^2 \mathrm{e}^{\mathrm{i}\delta}}. \tag{47.3}$$

Multiplying it by its complex conjugate, we have the reflected intensity as

$$I_r \;=\; I_0 \times \frac{2R(1-\cos\delta)}{(1-R)^2} \times \frac{1}{1 + \left[\dfrac{2\mathscr{F}}{\pi}\sin\left(\dfrac{\delta}{2}\right)\right]^2}, \tag{47.4}$$

which after addition with equation 43.2, under the condition where $R + T \;=\; 1$, gives I_0, and therefore the energy is conserved.

Finesse and the Number of Reflections

We shall now consider the ability of a Fabry-Perot cavity to trap light. Since each time the light reflects, the intensity decreases by a factor of R, after m bounces the intensity remaining in the cavity is given as

$$I_m \;=\; I_0 R^m \;\approx\; I_0 \mathrm{e}^{-(1-R)m} \tag{47.5}$$

for R close to unity i. e. $1 - R$ close to 0. As a result, the intensity decreases to I_0/e after $(1-R)^{-1} \;=\; \mathscr{F}/\pi$ reflections, and therefore the finesse $\mathscr{F}$ is also characteristic of the number of reflections light stays in the cavity.

Summary

1. Light in a Fabry-Perot cavity can be thought as a self-consistent steady state. By considering it as such we obtain the scalar amplitude agreeing with equation 43.1 up to an overall phase.

2. We can also obtain the reflected scalar amplitude, which we can find the corresponding intensity, which if the mirrors are lossless, then the reflected and transmitted intensity adds up to the incident intensity, hence showing the conservation of energy.

3. The finesse characterises the ability of a Fabry-Perot cavity in trapping light: the intensity will decay by a factor of 1/e after $(1-R)^{-1} = \mathscr{F}/\pi$ reflections from the mirror.

9 Multi-Layer Coatings

§48. Light Propagation in a Multi-Layer Coating

When light enters a region with a different refractive index, it is general that some light is reflected and some light is transmitted. However, for example when one is wearing glasses, most of the light turns out to be transmitted through the glass and only a very small fraction of the light is reflected. This is achieved by placing multi-layer coatings between the glass and air, and in this part we investigate into the use of this technology.

Boundary Conditions

We consider a light ray from a medium with refractive index $n_0 = 1$ and impedance Z_0 incident on a dielectric medium with refractive index n_1 and impedance Z_1. Let the light ray be equipped with electric and magnetic fields with parallel components E_0 and H_0 before it hits the boundary, and E_1 and H_1 after it hits the boundary. Also equip the reflected light with fields E_0', H_0', E_1', and H_1'. This is shown in figure 48.1. Note that, since the parallel components of $\mathbf{E}$ and $\mathbf{H}$ must be continuous at the boundary due to Maxwell's equations, we must have

$$E_0 + E_0' = E_1 + E_1'; \tag{48.1}$$
$$H_0 - H_0' = H_1 - H_1'. \tag{48.2}$$

Using the relation

$$\frac{Z_1}{Z_0} = \frac{\sqrt{\mu_0/(\varepsilon_0\varepsilon_r)}}{\sqrt{\mu_0/\varepsilon_0}} = \frac{1}{\sqrt{\varepsilon_r}} = \frac{n_0}{n_1}, \tag{48.3}$$

for non-magnetic media and the fact that the impedance is the ratio of electric and magnetic fields, we may reform equation 48.2 into

$$E_0 - E_0' = \frac{n_1}{n_0}(E_1 - E_1'). \tag{48.4}$$

By adding and subtracting equations 48.1 and 48.4, we are able to construct a matrix equation

$$\begin{pmatrix} E_0 \\ E_0' \end{pmatrix} = \mathscr{M}_{01} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix}, \qquad \mathscr{M}_{01} = \frac{1}{2}\begin{pmatrix} 1 + n_1/n_0 & 1 - n_1/n_0 \\ 1 - n_1/n_0 & 1 + n_1/n_0 \end{pmatrix} \tag{48.5}$$

which relates the electric fields of light at the boundary between the two media.

Pile of Dielectric Layers

We shall now consider light propagating through a pile of dielectric layers with refractive indices n_m and thickness d_m instead of a single layer, coming from air and ending up in glass with refractive index n_{T}. This configuration is shown in figure 48.2. We shall assume no reflected light inside the glass.

When light travels in the m^{th} layer, the electric field of the light leaving the layer picks up a phase shift term $\mathrm{e}^{\mathrm{i}k_m d_m}$ compared to the the electric field of the light entering that layer, i. e.

$$E(z + d_m) = \mathrm{e}^{\mathrm{i}k_m d_m} E(z); \tag{48.6}$$
$$E(z)' = \mathrm{e}^{\mathrm{i}k_m d_m} E(z + d_m)' \quad \Rightarrow \quad E(z + d_m)' = \mathrm{e}^{-\mathrm{i}k_m d_m} E(z)'. \tag{48.7}$$

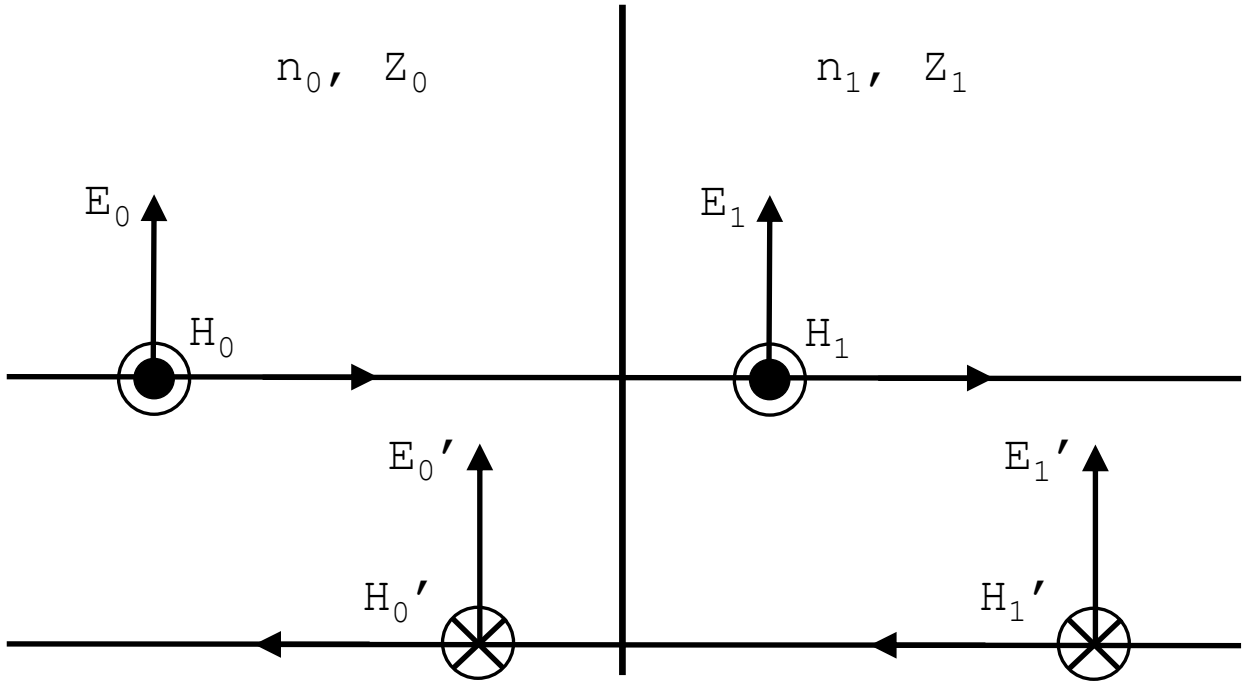

Figure 48.1: Light incident on a boundary between two media.

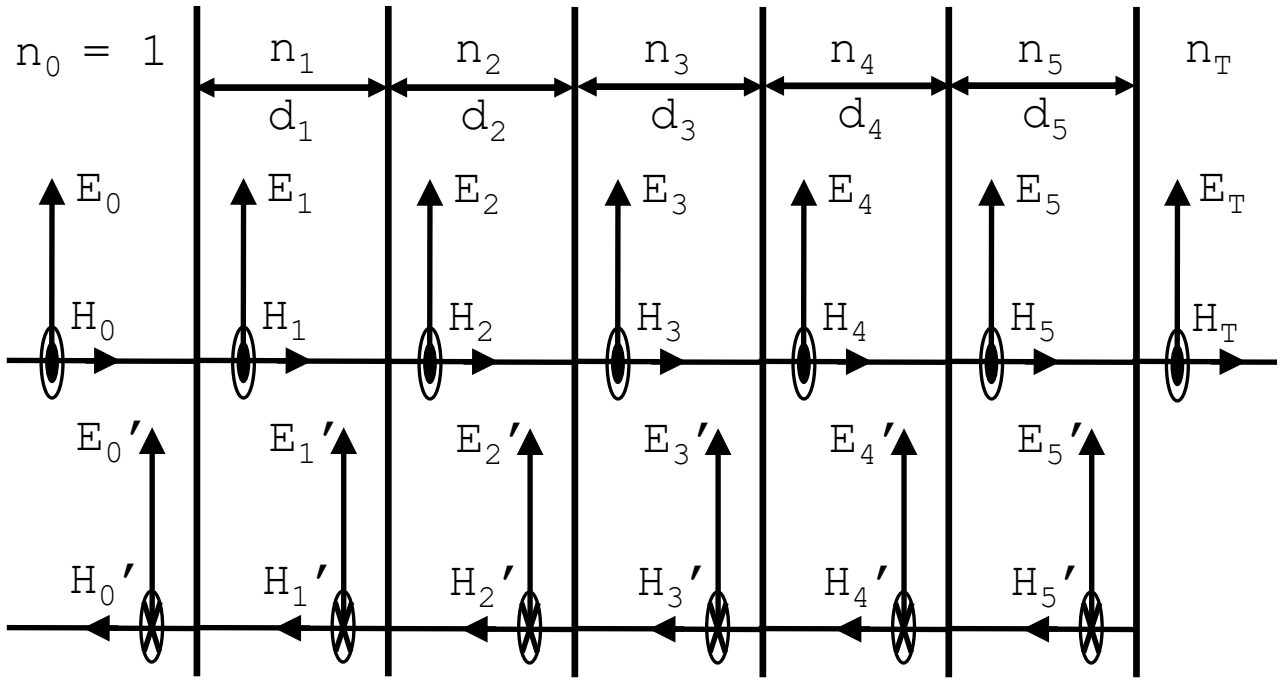

Figure 48.2: Light propagating through a pile of dielectric layers. Here the labels E, H, E', and H' denotes the E and H fields on the left end of the layer — after transmission through the layer these fields will then be phase shifted.

Note that here k_m is the wavenumber of the light inside the medium i. e. n_m times the wavenumber in a vacuum. We may also cast this into matrix form, giving

$$\begin{pmatrix} E(z) \\ E(z)' \end{pmatrix} = \mathscr{M}_{d_m} \begin{pmatrix} E(z+d_m) \\ E(z+d_m)' \end{pmatrix}, \qquad \mathscr{M}_{d_m} = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}k_m d_m} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}k_m d_m} \end{pmatrix}. \tag{48.8}$$

Then, the transmitted electric field E_T and the incident and reflected electric fields E_0 and E_0 are thus related by

$$\begin{pmatrix} E_0 \\ E_0' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E_\mathrm{T} \\ 0 \end{pmatrix} = \begin{pmatrix} AE_\mathrm{T} \\ CE_\mathrm{T} \end{pmatrix}, \qquad \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \left(\prod_{m=1}^{k} \mathscr{M}_{m-1,m} \mathscr{M}_{d_m} \right) \mathscr{M}_{k\mathrm{T}}. \tag{48.9}$$

Practically we would like to design anti-reflection (AR) and high-reflection (HR) dielectric coatings, that is, to minimise or maximise the reflection coefficient

$$R = |C/A|^2. \tag{48.10}$$

We shall take steps to see how this can be achieved.

Summary

1. When light from a medium with refractive index n_0 hits a dielectric with refractive index n_1, the boundary conditions enforce the outgoing and reflected rays to satisfy the equation

$$\begin{pmatrix} E_0 \\ E_0' \end{pmatrix} = \mathscr{M}_{01} \begin{pmatrix} E_1 \\ E_1' \end{pmatrix}, \qquad \mathscr{M}_{01} = \frac{1}{2} \begin{pmatrix} 1+n_1/n_0 & 1-n_1/n_0 \\ 1-n_1/n_0 & 1+n_1/n_0 \end{pmatrix}. \tag{48.11}$$

2. For light travelling through a pile of dielectric layers with refractive indices n_m and thickness d_m, we have the matrix equation

$$\begin{pmatrix} E_0 \\ E_0' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E_\mathrm{T} \\ 0 \end{pmatrix} = \begin{pmatrix} AE_\mathrm{T} \\ CE_\mathrm{T} \end{pmatrix}, \qquad \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \left(\prod_{m=1}^{k} \mathscr{M}_{m-1,m} \mathscr{M}_{d_m} \right) \mathscr{M}_{k\mathrm{T}} \tag{48.12}$$

where

$$\mathscr{M}_{d_m} = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}k_m d_m} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}k_m d_m} \end{pmatrix}. \tag{48.13}$$

§49. Single Layer Coatings

A Single $\lambda/4$ Layer

Now we are acquainted with the basic tools to analyse this problem, we may consider the simplest type of an anti-reflection (AR) coating. This is constructed by having a single layer of dielectric with refractive index n between air with refractive index 1 and glass with refractive index n_T. We equip the dielectric with thickness $\lambda/4$ where λ is the wavelength of the light inside the dielectric and demonstrate that perfect anti-reflection can be achieved with an appropriately chosen n. Then, using the fact that now

$$\mathrm{e}^{\mp \mathrm{i}k\frac{\lambda}{4}} = \mathrm{e}^{\mp \mathrm{i}\frac{2\pi}{\lambda}\times\frac{\lambda}{4}} = \mathrm{e}^{\mp \mathrm{i}\frac{\pi}{2}} = \mp\mathrm{i}, \tag{49.1}$$

we have the matrix defined in equation 48.9 as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1+n & 1-n \\ 1-n & 1+n \end{pmatrix} \begin{pmatrix} -\mathrm{i} & 0 \\ 0 & \mathrm{i} \end{pmatrix} \frac{1}{2} \begin{pmatrix} 1+n_\mathrm{T}/n & 1-n_\mathrm{T}/n \\ 1-n_\mathrm{T}/n & 1+n_\mathrm{T}/n. \end{pmatrix} \tag{49.2}$$

A short calculation then gives

$$A \;=\; -\frac{\mathrm{i}}{2}\Big(\frac{n_{\mathrm{T}}}{n}+n\Big), \qquad C \;=\; -\frac{\mathrm{i}}{2}\Big(\frac{n_{\mathrm{T}}}{n}-n\Big), \qquad R \;=\; \left|\frac{n_{\mathrm{T}}-n^2}{n_{\mathrm{T}}+n^2}\right|^2, \tag{49.3}$$

and therefore for perfect anti-reflection ($R \;=\; 0$) we require

$$n \;=\; \sqrt{n_{\mathrm{T}}}. \tag{49.4}$$

Glass has refractive index $n_{\mathrm{T}} \;=\; 1.52$ and therefore an optimum anti-reflective material should have a refractive index $n \;=\; 1.23$. One $\lambda/4$ layer of commonly used anti-reflective material, magnesium fluoride, MgF_2, has a refractive index $n \;=\; 1.37$ which gives $R \;=\; 0.011$, i. e. the reflection coefficient can be reduced to about only 1% with a single layer.

A Fabry-Perot Interferometer

We shall then attempt to do a sanity check of our formalism for the reflection coefficient using the Fabry-Perot interferometer. Let us consider light bouncing between two dielectric with refractive index n in air with refractive index 1. The configuration for a transmission maximum at normal incidence requires the distance between the dielectric to be an integer multiple of half-wavelengths, giving $\mathrm{e}^{\mathrm{i}kd} \;=\; \pm 1$. We shall attempt to show that this retrieves the reflection coefficient calculated using equation 47.4:

$$I_r \;=\; I_0 \times \frac{2R_n(1-\cos\delta)}{(1-R_n)^2} \times \frac{1}{1+\left[\dfrac{2\mathscr{F}}{\pi}\sin\left(\dfrac{\delta}{2}\right)\right]^2}, \tag{49.5}$$

(R in equation 47.4 is displayed by R_n here to resolve ambiguity of R) which gives $I_r \;=\; 0$ when δ is an integer multiple of 2π, i. e. the reflected rays destructively interferes giving no reflected intensity. Let us demonstrate that using our new formalism.

Here the matrix defined in in equation 48.9 is

$$\begin{aligned}\begin{pmatrix} A & B \\ C & D \end{pmatrix} \;&=\; \frac{1}{2}\begin{pmatrix} 1+1/n & 1-1/n \\ 1-1/n & 1+1/n \end{pmatrix}\begin{pmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{pmatrix}\frac{1}{2}\begin{pmatrix} 1+n & 1-n \\ 1-n & 1+n \end{pmatrix} \\ &=\; \frac{1}{4}\begin{pmatrix} 1+1/n & 1-1/n \\ 1-1/n & 1+1/n \end{pmatrix}\begin{pmatrix} \pm(1+n) & \pm(1-n) \\ \pm(1-n) & \pm(1+n) \end{pmatrix},\end{aligned} \tag{49.6}$$

and hence

$$C \;=\; \pm\frac{1}{4}\left[\left(1-\frac{1}{n}\right)(1+n)+\left(1+\frac{1}{n}\right)(1-n)\right] \;=\; 0, \tag{49.7}$$

therefore $R \;=\; I_r \;=\; 0$ as expected.

Summary

1. For a perfect anti-reflection coating between air with refractive 1 and glass with refractive index n_{T}, we require a $\lambda/4$ layer of $n \;=\; \sqrt{n_{\mathrm{T}}}$.
2. Our new formalism is able to verify that the reflected intensity of a Fabry-Perot interferometer is 0 given that light is at normal incidence and the distance between dielectrics is an integer multiple of half-wavelengths.

§50. Impedance Matching

Note that the formalism that we have developed is suitable for describing travelling waves in a vacuum or in a dielectric, however the mathematical starting points we have selected are exactly the same as confined electromagnetic waves in a transmission line, with voltage and current corresponding to the electric and magnetic fields. This thought is inspiring: can we apply the methods that we employed in the theory of transmission lines to simplify calculations in an optics context?

To think about impedance matching we must define an appropriate impedance for each layer. Let us think about a simple case illustrated in figure 48.2 where we have five layers between air and glass. Let us further simplify the case to assume these layers are all $\lambda/4$ layers, that is, their thickness corresponds to one quarter of the wavelength inside these media. It is instructive to define the impedance in the glass as

$$Z_{\mathrm{T}} \;=\; Z_0/n_{\mathrm{T}} \tag{50.1}$$

as this is the usual definition of impedance. To define suitable impedances for the rest of the system we recall two results from the theory of $\lambda/4$-lines:

- for complete transmission of energy we require

$$\mathrm{Z_{in}} \;=\; \mathrm{Z}^2/\mathrm{Z_L} \tag{50.2}$$

  where Z is the impedance of the line, $\mathrm{Z_L}$ is the impedance of the load, and $\mathrm{Z_{in}}$ is the input impedance;

- the reflection coefficient of a line connected to a load is

$$R \;=\; |(\mathrm{Z_L} - \mathrm{Z}_c)/(\mathrm{Z_L} + \mathrm{Z}_c)|^2 \tag{50.3}$$

  where $\mathrm{Z_L}$ is the impedance of the load and Z_c is the combined impedance of the input impedance and line impedance. For complete transmission we need $R \;=\; 0$, or $\mathrm{Z_L} \;=\; \mathrm{Z}_c$ i. e. impedance matching.

Although we have claimed that these results are "recalled", they can be proved based on principles of optics using the methods in previous sections.

To use these results in the current setup, let us first think of the layer with refractive index n_5 as the $\lambda/4$-line and everything else contributes to the input impedance $\mathrm{Z_{in,5}}$. We then define the layer to have an impedance $Z_5 \;=\; Z_0/n_5$. Then by equation 50.2 we have

$$\mathrm{Z_{in,5}} \;=\; Z_5^2/Z_{\mathrm{T}}, \tag{50.4}$$

where we consider Z_5 as the "line impedance" and Z_{T} as the "load impedance". We then take one step back and think of the fifth layer and glass as the load and air and the first three layers as the input impedance. In this case the "combined impedance" in equation 50.3 is $\mathrm{Z_{in,5}}$ and this therefore must equate to the new load impedance for the suppression of the reflection coefficient. Then applying equation 50.2 again, we have

$$\mathrm{Z_{in,4}} \;=\; Z_4^2/\mathrm{Z_{in,5}} \;=\; (Z_4/Z_5)^2 Z_{\mathrm{T}}. \tag{50.5}$$

Applying this method repeatedly then yields

$$\mathrm{Z}_{\mathrm{in},1} \;=\; (Z_1 Z_3 Z_5)^2 / \big[(Z_2 Z_4)^2 Z_{\mathrm{T}}\big]. \tag{50.6}$$

Generalising this to $2p$ and $2p+1$ layers we have

$$\mathrm{Z}_{\mathrm{in},1} \;=\; (Z_1 Z_3 \cdots Z_{2p-1})^2 Z_{\mathrm{T}} / (Z_2 Z_4 \cdots Z_{2p})^2 \tag{50.7}$$

and

$$\mathrm{Z}_{\mathrm{in},1} \;=\; (Z_2 Z_4 \cdots Z_{2p})^2 / \big[(Z_1 Z_3 \cdots Z_{2p+1})^2 Z_{\mathrm{T}}\big] \tag{50.8}$$

respectively. Now, since the boundary 01 is exactly the air-coating boundary, we have no "line". Therefore, the input impedance $\mathrm{Z}_{\mathrm{in},1}$ is therefore the load impedance for the entire system and equation 50.3 becomes

$$R \;=\; |(\mathrm{Z}_{\mathrm{in},1} - Z_0)/(\mathrm{Z}_{\mathrm{in},1} + Z_0)|^2 \tag{50.9}$$

Hence, we have the reflection coefficient at the air-coating boundary as

$$\begin{aligned} R \;&=\; \left| \frac{(Z_1 Z_3 \cdots Z_{2p-1})^2 Z_{\mathrm{T}} / (Z_2 Z_4 \cdots Z_{2p})^2 - Z_0}{(Z_1 Z_3 \cdots Z_{2p-1})^2 Z_{\mathrm{T}} / (Z_2 Z_4 \cdots Z_{2p})^2 + Z_0} \right|^2 \\ &=\; \left| \frac{n_0 (n_2 n_4 \cdots n_{2p})^2 - (n_1 n_3 \cdots n_{2p-1})^2 n_{\mathrm{T}}}{n_0 (n_2 n_4 \cdots n_{2p})^2 + (n_1 n_3 \cdots n_{2p-1})^2 n_{\mathrm{T}}} \right|^2 \end{aligned} \tag{50.10}$$

for $2p$ layers and

$$\begin{aligned} R \;&=\; \left| \frac{(Z_2 Z_4 \cdots Z_{2p})^2 / \big[(Z_1 Z_3 \cdots Z_{2p+1})^2 Z_{\mathrm{T}}\big] - Z_0}{(Z_2 Z_4 \cdots Z_{2p})^2 / \big[(Z_1 Z_3 \cdots Z_{2p+1})^2 Z_{\mathrm{T}}\big] + Z_0} \right|^2 \\ &=\; \left| \frac{n_0 (n_1 n_3 \cdots n_{2p+1})^2 n_{\mathrm{T}} - (n_2 n_4 \cdots n_{2p})^2}{n_0 (n_1 n_3 \cdots n_{2p+1})^2 n_{\mathrm{T}} + (n_2 n_4 \cdots n_{2p})^2} \right|^2 \end{aligned} \tag{50.11}$$

for $2p+1$ layers. These results can also be obtained directly using the matrix method in the previous section.

We therefore infer that, to make an anti-reflective coating we require $R = 0$ and therefore we require $Z_0 = \mathrm{Z}_{\mathrm{in},1}$, i. e. we need impedance matching at the 01 boundary. Instead if we would like to make a mirror, or a high-reflective coating, then we instead need $Z_0 \ll \mathrm{Z}_{\mathrm{in},1}$, i. e. an impedance mismatch between the 01 boundary.

Summary

1. Using the key thought that, for complete transmission we require the load impedance equal to the combined contribution of the line and input impedances, we are able to calculate the reflection coefficient for a multi-layer stack, that are

$$R \;=\; \left| \frac{n_0 (n_2 n_4 \cdots n_{2p})^2 - (n_1 n_3 \cdots n_{2p-1})^2 n_{\mathrm{T}}}{n_0 (n_2 n_4 \cdots n_{2p})^2 + (n_1 n_3 \cdots n_{2p-1})^2 n_{\mathrm{T}}} \right|^2 \tag{50.12}$$

for $2p$ layers and

$$R \;=\; \left| \frac{n_0 (n_1 n_3 \cdots n_{2p+1})^2 n_{\mathrm{T}} - (n_2 n_4 \cdots n_{2p})^2}{n_0 (n_1 n_3 \cdots n_{2p+1})^2 n_{\mathrm{T}} + (n_2 n_4 \cdots n_{2p})^2} \right|^2 \tag{50.13}$$

for $2p+1$ layers.

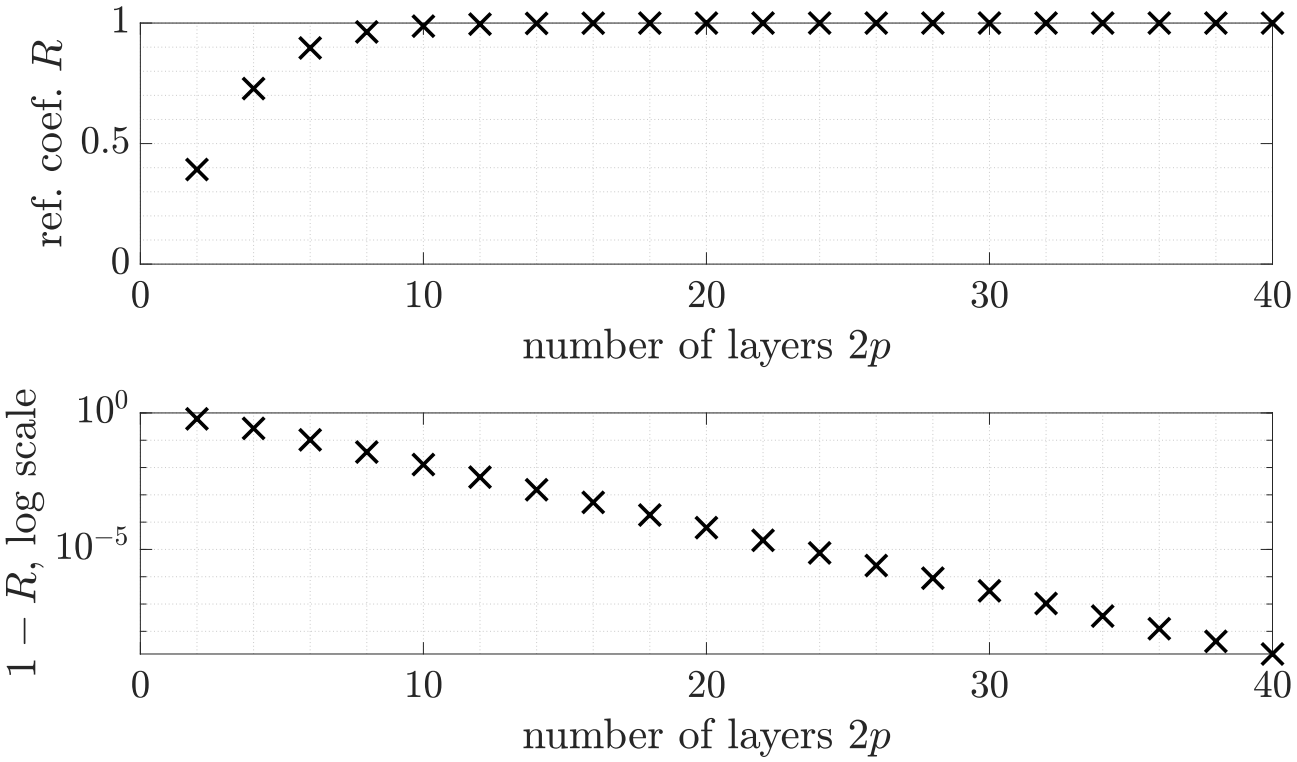

Figure 51.1: The reflection coefficient R against the number of layers $2p$ for a multi-layer high-reflection coating with $n_0 = 1.00$, $n_L = 1.35$, $n_H = 2.30$, and $n_\mathrm{T} = 1.50$.

§51. Multi-Layer Stacks

A Multi-Layer High-Reflection Coating

Previously we have looked into a single layer anti-reflection coating, let us see how to extend this theory to make a multi-layer high-reflection coating.

The simplest way to do this is to consider alternating $\lambda/4$-layers with refractive indices n_H and n_L with the layer at the air-coating boundary equipped with refractive index n_H. We shall set $n_H > n_L$ to create the impedance mismatch, and also set the total number of layers to be $2p$, that is, p dielectric layers of each refractive index. Then, using equation 50.10, the reflection coefficient is given as

$$R = \left| \frac{{n_0}^2 {n_L}^{2p} - {n_H}^{2p} {n_\mathrm{T}}^2}{{n_0}^2 {n_L}^{2p} + {n_H}^{2p} {n_\mathrm{T}}^2} \right|^2 . \tag{51.1}$$

To observe an increase in performance for an increase of layers let us adopt the toy parameters $n_0 = 1.00$, $n_L = 1.35$, $n_H = 2.30$, and $n_\mathrm{T} = 1.50$. This relationship in shown in figure 51.1. We observe that the reflection coefficient quickly goes towards 1, with only $R = 0.39$ for two layers and $R = 0.999\,937$ for twenty layers, making it a very good high-reflection coating. As a comparison, aluminium has $R \approx 0.88$ which is a lot less reflective than the coating that we have just created.

Interference Filters

The high-reflection coating we have just created is built based on the fact that the layers in between are $\lambda/4$-layers. This means that we have to built the coating for one specific wavelength, and light with other wavelengths behave differently. For example, figure 51.2 is the behaviour of the coating that we have just created for light with different wavelengths.

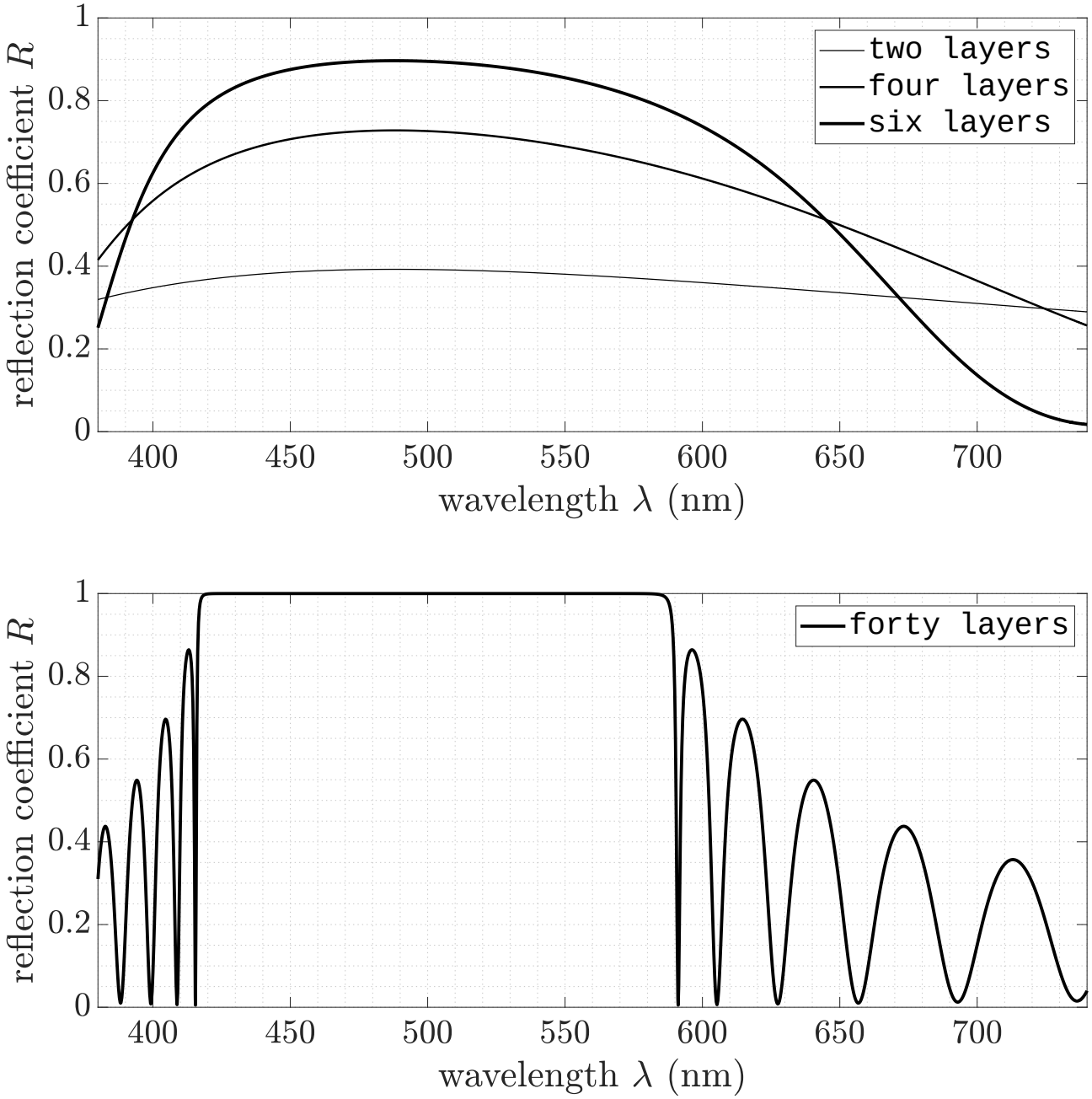

Figure 51.2: The reflection coefficient R against the wavelength λ for a multi-layer high-reflection coating with $n_0 = 1.00$, $n_L = 1.35$, $n_H = 2.30$, and $n_\mathrm{T} = 1.50$, built with $\lambda = 488\,\mathrm{nm}$.

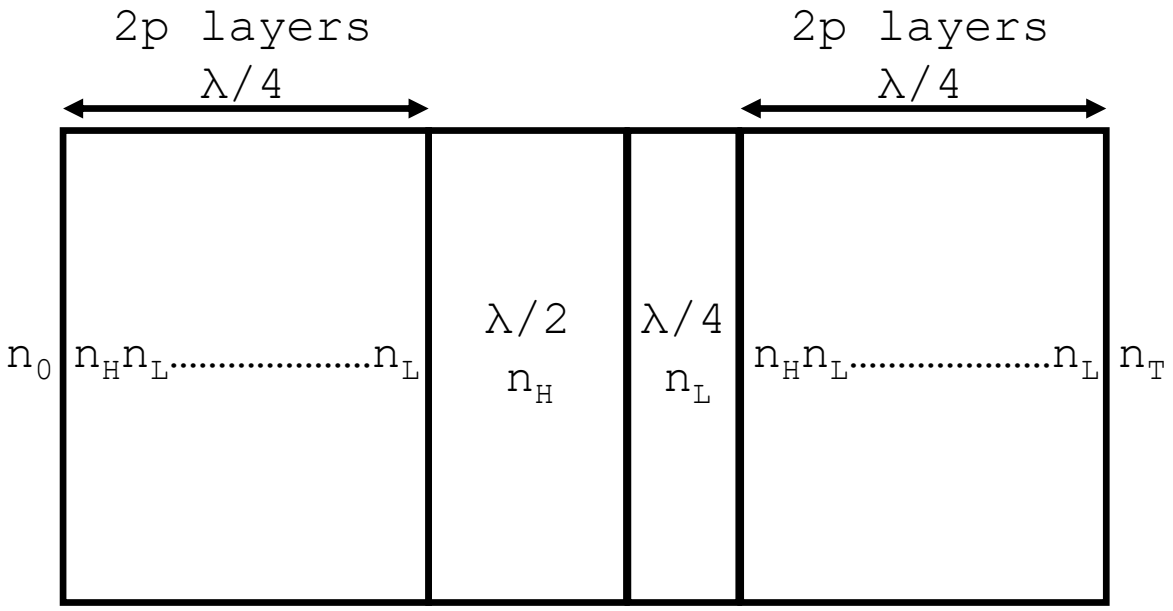

Figure 51.3: A simple scheme for an interference filter that selects a single wavelength.

We observe that figure 51.2 have very sharp peaks and troughs, therefore we may pose the question of whether we can build "interference filters" which blocks some wavelengths (or, is a high-reflection coating of these wavelengths) and lets other wavelengths through (acts as an anti-reflection coating).

The answer is yes, and a simple example is a coating that selects one specific wavelength. Such a filter can be built with two high-reflection coatings on each side sandwiching a $\lambda/2$-layer in the middle illustrated in figure 51.3. Then, the $\lambda/2$-layer acts as a solid Fabry-Perot etalon, selecting the required wavelength to transmit and reflecting all other wavelengths. Figure 51.4 demonstrates the performance of such a filter. Other filters can also be built, such as coatings that acts as a high-pass filter, band-pass filter, or a "notch" (blocking a specific range of wavelengths and letting through the rest).

Summary

1. By depositing a multi-layer coating with alternating $\lambda/4$-layers with refractive indices n_H and n_L, we have impedance mismatch between air and glass, giving a high-reflective coating with
$$R = \left| \frac{n_0^2 {n_L}^{2p} - {n_H}^{2p} {n_\mathrm{T}}^2}{n_0^2 {n_L}^{2p} + {n_H}^{2p} {n_\mathrm{T}}^2} \right|^2 . \tag{51.2}$$
2. We are able to create multi-layer interference filters that acts as a high-reflection coating for some wavelengths and a low-reflection coating for others. Examples of these are filters that selects a single wavelength, a high-pass, a band-pass, and a "notch" filter.

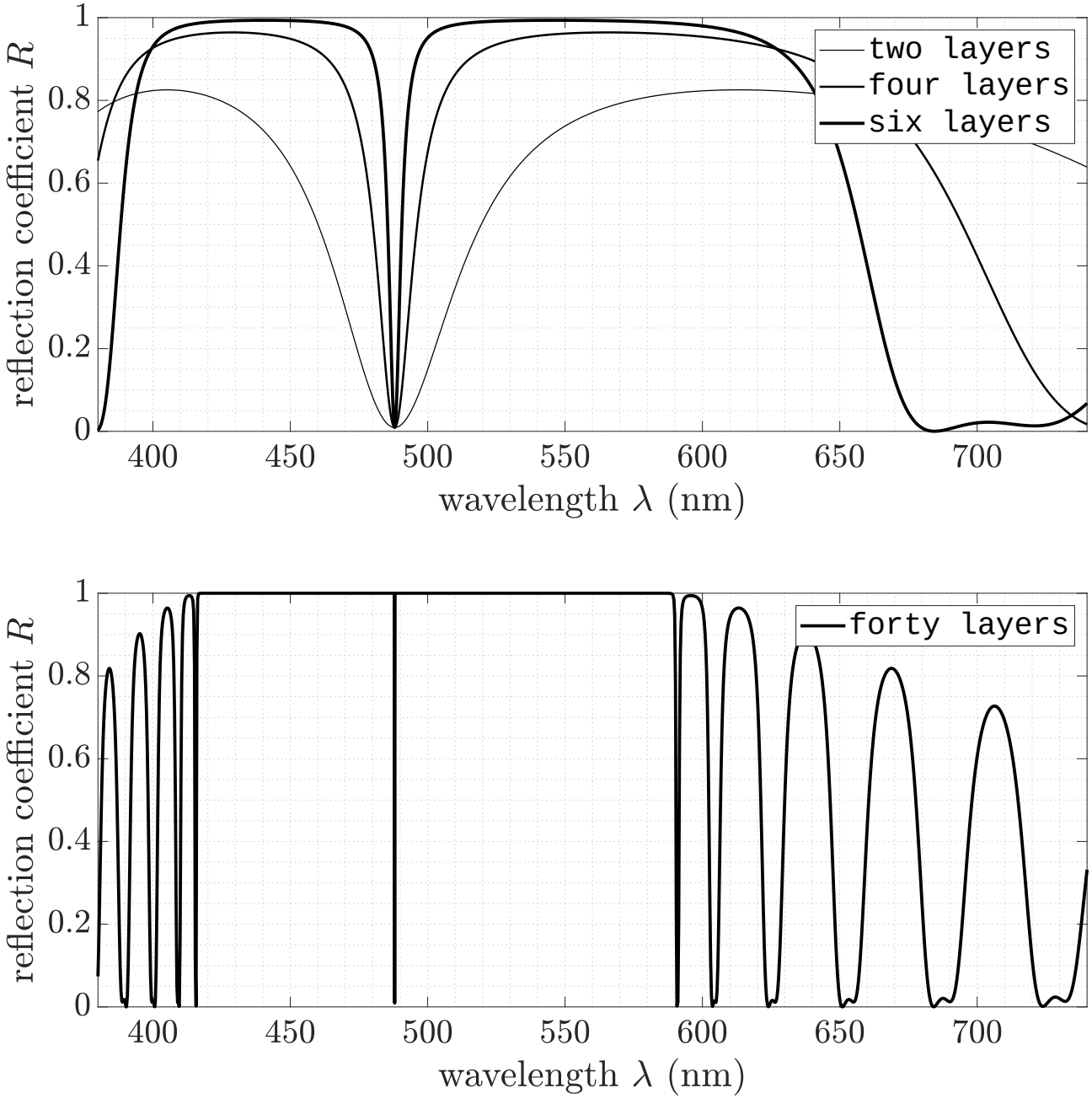

Figure 51.4: The reflection coefficient R against wavelength λ for an interference filter built according to the scheme of figure 51.3 with $n_0 = 1.00$, $n_L = 1.35$, $n_H = 2.30$, $n_\mathrm{T} = 1.50$, and $\lambda = 488\,\mathrm{nm}$. Note that the number of layers in the caption is the number of layers *per stack*, i. e. $2p$ in figure 51.3.

10 Lasers

§52. Interaction between Atoms and Radiation

What are Lasers?

To drive a large interferometer, for example a Michelson interferometer with long arm lengths, we need a light source that provides highly coherent and highly focussed sources of light. This is usually done by a device called a laser.

We shall now have a look at the principle of **lasers**. Lasers is a shorthand for **light amplification by stimulated emission of radiation**, where the idea is to trap some atoms or crystals (or "artificial atoms" in general, as long as they have got some discrete energy levels) in a cavity, and shoot light into them, making the atoms amplify the incoming radiation such that the output intensity is larger than the input intensity. To consider how this works, we need to first consider the interaction between the atoms and light. Then we note that energy does not come from nowhere — we need to think hard about how to "pump" the atoms i. e. to put the energy in. Then we need to see how to maintain the light output continuously. Finally we shall look at why the construction of lasers gives rise to the highly coherent radiation that we are looking for.

Interactions in a Two-Level System

The core actions of a laser happens between two states $|1\rangle$ and $|2\rangle$ of an atom, and what we hope is that after we feed in some radiation and some additional energy to the atom, the atom will de-excite back from $|2\rangle$ to $|1\rangle$, emitting some radiation with this de-excitation, and hence amplifying the incoming radiation; and the de-excitation will be equipped with exactly the frequency

$$\omega_{21} \;=\; (E_2 - E_1)/\hbar, \tag{52.1}$$

where E_1 and E_2 are the energies corresponding to states $|1\rangle$ and $|2\rangle$ respectively, and hence the radiation is coherent. The patch of space that contains such atoms is the source of light amplification, and is called the **laser medium**. In the two-level system, there are three possible interactions between the atoms and light. The following scheme sets up a simple model of the three interactions.

- **Absorption**: the atoms in state $|1\rangle$ can absorb energy and excite to $|2\rangle$, which has a rate proportional to the number of atoms in $|1\rangle$, N_1, and the energy density at ω_{21}, which we denote by ρ_{em}. The constant of proportionality of this process is denoted by the **Einstein's B-coefficient** B_{12}. Putting these together, the rate of absorption R_A is

$$R_A \;=\; B_{12} N_1 \rho_{em}. \tag{52.2}$$

- **Stimulated emission**: if the system is incident upon an incoming photon with frequency ω_{12}, then an atom in state $|2\rangle$ may de-excite to $|1\rangle$, emitting another photon to exit the system coherent with the incoming photon in the same direction. The rate of this process is proportional to the number of atoms in $|2\rangle$, N_2, and the energy density at ω_{21}, ρ_{em}. The constant of proportionality is given by another Einstein's B-coefficient B_{21}. Putting these together, the rate of absorption R_{ST} is

$$R_{ST} \;=\; B_{21} N_2 \rho_{em}. \tag{52.3}$$

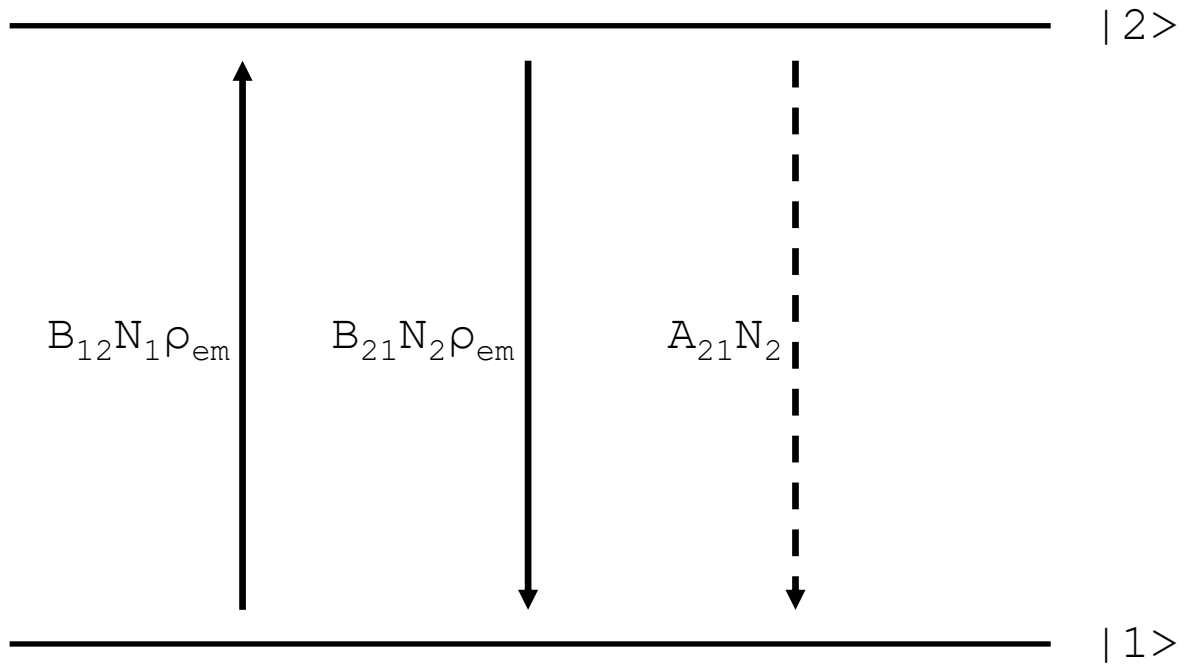

Figure 52.1: The three interactions that can occur between a medium of atoms with two states $|1\rangle$ and $|2\rangle$ and light.

- **Spontaneous emission**: a more complicated theory of quantum optics suggests that it is possible to quantise the background electromagnetic field, which would form energy levels similar to a quantum harmonic oscillator. This means that, even we are dealing with a complete vacuum, there is a zero point energy, usually called **vacuum fluctuations**. This is able to provide some energy to the atoms at the upper level, allowing it to de-excite into the ground state, where the light exits the system in completely random directions at random phases. This process is called spontaneous emission. The rate of this process is proportional to the number of atoms in state $|2\rangle$, N_2, and the constant of proportionality of this process is denoted by the **Einstein's A-coefficient** A_{21}. Putting these together, the rate of absorption R_{SP} is
$$R_{SP} \;=\; A_{21}N_2. \tag{52.4}$$

The interactions is demonstrated in figure 52.1.

Summary

1. Laser, or light amplification of stimulated emission by radiation, is a process that amplifies the incoming radiation. Laser light is coherent and focussed.
2. There are three processes of interaction between a medium of two-state atoms and light, they are
   - absorption, with rate $R_A \;=\; B_{12}N_1\rho_{em}$;
   - stimulated emission, with rate $R_{ST} \;=\; B_{21}N_2\rho_{em}$;
   - and spontaneous emission, with rate $R_{SP} \;=\; A_{21}N_2$.

§53. Einstein's A and B coefficients

Einstein's Rate Equations

We can gather the three processes in the previous section, and write down the rates of change of the number of atoms in states $|2\rangle$ and $|1\rangle$, which we denote by $\dot{N}_2$ and $\dot{N}_1$ respectively, as follows,

$$\dot{N}_2 \;=\; -A_{21}N_2 - B_{21}N_2\rho_{em} + B_{12}N_1\rho_{em}; \tag{53.1}$$

$$\dot{N}_1 \;=\; A_{21}N_2 + B_{21}N_2\rho_{em} - B_{12}N_1\rho_{em}. \tag{53.2}$$

For a laser to be ran sustainably at the steady state we require the two rates to be 0. Using this, we may rearrange our equations into the form

$$\frac{N_2}{N_1} \;=\; \frac{B_{12}\rho_{em}}{A_{21} + B_{21}\rho_{em}}. \tag{53.3}$$

Next we shall attempt to find Einstein's A and B coefficients.

Balancing of Emission and Absorption processes

We note that, although previously we have talked about $|1\rangle$ and $|2\rangle$ as two states, in principle they are two energy levels, which means that they could be many degenerate states with the same energy. For example, the hydrogen atom with $n = 2$ and $\ell = 1$ is three-fold degenerate with $m_\ell = -1, 0$, or 1. If we then consider the transition between the three degenerate states $|2, 1, m_\ell\rangle$ and the non-degenerate state $|1, 0, 0\rangle$, then since there are three different states an atom might de-excite from the upper energy level, and only one state for the atom on the ground state to be excited, for this system spontaneous emission is three times more likely than absorption, i. e. $B_{12} = 3B_{21}$. In general with the two energy levels g_1- and g_2-fold degenerate, we have the relation between Einstein's B coefficients

$$g_2B_{21} \;=\; g_1B_{12}. \tag{53.4}$$

In this set of notes, we ignore the degeneracies, i. e. we set $g_1 = g_2 = 1$, and therefore $B_{21} = B_{12} = B$.

Thermal Equilibrium

To determine the relationship between Einstein's A and B coefficients we then consider thermal equilibrium of the system. Note that the populations between the two levels satisfy Boltzmann statistics

$$\frac{N_2}{N_1} \;=\; \mathrm{e}^{-\beta(E_2-E_1)} \;=\; \mathrm{e}^{-\beta\hbar\omega}, \tag{53.5}$$

where $\omega = \omega_{21}$ and $\beta = 1/(k_\mathrm{B}T)$. Additionally, the energy density in the medium satisfies the Planck distribution

$$\rho_{em} \;=\; \frac{\hbar\omega^3}{\pi^2c^3} \times \frac{1}{\mathrm{e}^{\beta\hbar\omega} - 1}. \tag{53.6}$$

Substituting these two results into Einstein's rate equation in the steady state

$$N_1B\rho_{em} \;=\; N_2B\rho_{em} + N_2A_{21}, \tag{53.7}$$

we have

$$B \times \frac{\hbar\omega^3}{\pi^2c^3} \;=\; B \times \frac{\hbar\omega^3}{\pi^2c^3}\mathrm{e}^{-\beta\hbar\omega} + A_{21}\big(1 - \mathrm{e}^{-\beta\hbar\omega}\big). \tag{53.8}$$

For this to work at any temperatures, we therefore must have the relation between Einstein's A and B coefficients as

$$A_{21} \;=\; B \times \frac{\hbar\omega^3}{\pi^2 c^3}. \tag{53.9}$$

An alternative method to derive the relations between the Einstein's A and B coefficients is to make ρ_{em} the subject of Einstein's rate equation in the steady state, taking into account of Boltzmann statistics, and equating it to the Planck distribution,

$$\rho_{em} \;=\; \frac{A_{21}/B_{21}}{\dfrac{N_1}{N_2}\times\dfrac{B_{12}}{B_{21}} - 1} \;=\; \frac{A_{21}/B_{21}}{\dfrac{B_{12}}{B_{21}}\times \mathrm{e}^{\beta\hbar\omega} - 1} \;\stackrel{!}{=}\; \frac{\hbar\omega^3/(\pi^2c^3)}{\mathrm{e}^{\beta\hbar\omega}-1}, \tag{53.10}$$

which gives $B_{12} \;=\; B_{21} \;=\; B$ and $A_{21} \;=\; B \times \hbar\omega^3/(\pi^2c^3)$ by inspection.

Now we are able to see whether we may amplify an incoming light ray i. e. get the laser running, and we shall do that in the next section.

Summary

1. Einstein's rate equations relate the three allowed processes of interaction between light and two-state atoms, which gives, in the steady state
$$\frac{N_2}{N_1} \;=\; \frac{B_{12}\rho_{em}}{A_{21} + B_{21}\rho_{em}}. \tag{53.11}$$
2. Balancing the emission and absorption processes gives the relation between the two Einstein B coefficients
$$B_{12} \;=\; B_{21} \;=\; B. \tag{53.12}$$
3. Taking into account Boltzmann statistics of the relative population of atoms in the two energy levels and Planck's distribution for the background radiation, we have the relation between Einstein's A and B coefficients
$$A_{21} \;=\; B \times \frac{\hbar\omega^3}{\pi^2 c^3}. \tag{53.13}$$

§54. Condition for Amplification

Radiative Pumping

Previously I have stated that what we want is to pump energy in for the system to amplify the input light. However one might wonder whether we are able to amplify using light — that is, incoherent radiation in, coherent radiation out. This short calculation shows that for a two-level system it is simply impossible.

To reach amplification, we require the rate of stimulated emission to be larger than the rate of absorption, that is

$$R_{ST} \;>\; R_A \quad\Rightarrow\quad BN_2\rho_{em} \;>\; BN_1\rho_{em} \quad\Rightarrow\quad \Delta N \;=\; N_2 - N_1 \;>\; 0, \tag{54.1}$$

i. e. we require a **population inversion** between the two states. However if we look back at the equation for the relative populations, equation 53.3, this gives

$$\frac{N_2}{N_1} \;=\; \frac{B\rho_{em}}{A_{21} + B\rho_{em}} \;=\; \frac{1}{A_{21}/(B\rho_{em}) + 1}, \tag{54.2}$$

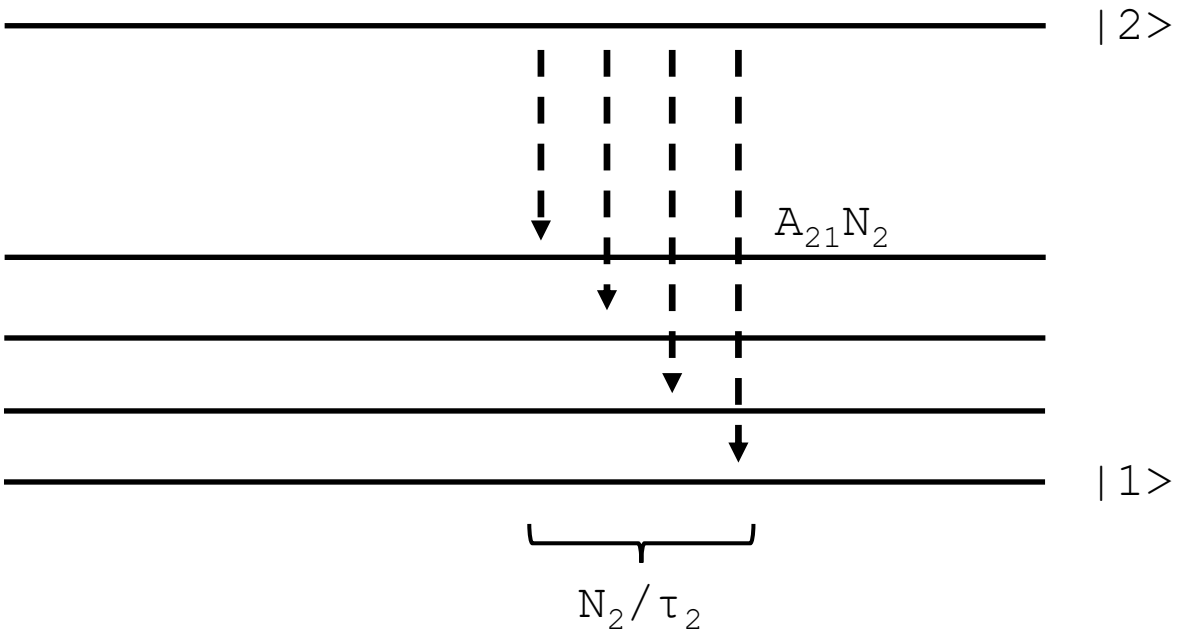

Figure 54.1: The comparison between spontaneous emission and decay.

which is strictly less than 1, i. e. we simply cannot amplify the light by purely pumping the system with radiation. This motivates the use for another pump mechanism.

Alternative Pump Mechanism

Now that pumping by radiation in the two-level laser is impossible (although they will come back to this idea on the discussion on three- and four-level lasers), we simply assume that the interaction between incident light and the medium of atoms, i. e. processes with rates proportional to Einstein's B coefficients, is much less significant than the parameters in our pump mechanism and the process of spontaneous emission. This means that we strip off all the terms involving Einstein's B coefficients from our calculation. We now consider inputting energy in a form of pumping particles into the states $|1\rangle$ and $|2\rangle$ separately, with pump rates P_1 and P_2 respectively. However we note that the when we actually reach population inversion, the population of these two states does not obey Boltzmann statistics, and therefore the population of the two levels $|1\rangle$ and $|2\rangle$ will simultaneously decrease by de-excitation to many other lower levels. The rate of this is proportional to the population of the level, and the constant of proportionality is given by $1/\tau$, where τ is called the **fluorescence lifetime** of the level. Note that the process of spontaneous emission between the two levels in question is also due to this de-excitation process, illustrated by figure 54.1, and therefore we must have

$$1/\tau_2 \geqslant A_{21}. \tag{54.3}$$

With this in mind, we have the interaction between this alternative pump mechanism and the atoms is illustrated in figure 54.2. We can therefore construct rate equations

$$\dot{N}_2 = P_2 - N_2/\tau_2; \tag{54.4}$$

$$\dot{N}_1 = P_1 + A_{21}N_2 - N_1/\tau_1. \tag{54.5}$$

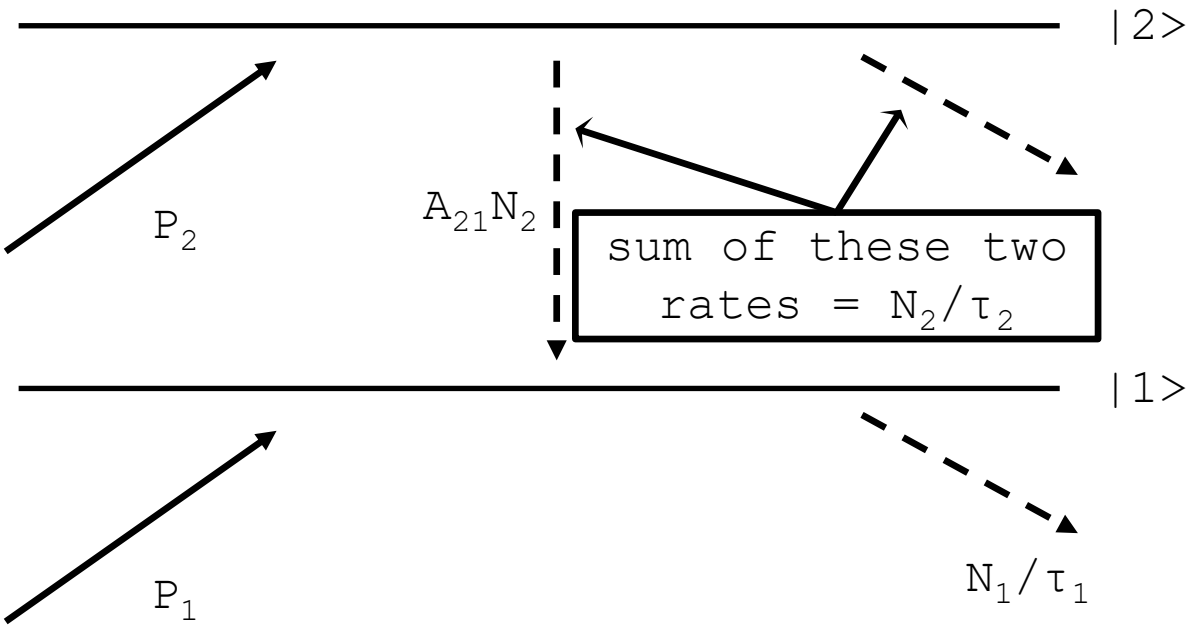

Figure 54.2: Interaction between the medium of atoms and the pump mechanism.

In the steady state case, we again have $\dot{N}_1 = \dot{N}_2 = 0$, and therefore, by rearranging the above rate equations, we have the difference in populations

$$\Delta N = N_2 - N_1 = P_2\tau_2(1 - A_{21}\tau_1) - P_1\tau_1. \tag{54.6}$$

Therefore, to reach population inversion, we must have

$$\frac{P_2}{P_1} \times \frac{\tau_2}{\tau_1} \times (1 - A_{21}\tau_1) > 1. \tag{54.7}$$

Hence, for population inversion to be achieved in a steady state, we need the following necessary but not sufficient condition

$$A_{21}\tau_1 < 1 \tag{54.8}$$

to keep the value of the bracket positive. In addition to this, it is apparent that we need at least one of the following:

- selective pumping: we pump state $|2\rangle$ at a higher rate compared to state $|1\rangle$, i. e. $P_2 > P_1$;
- favourable lifetime ratio: we need the lifetime of the upper level τ_2 to be longer than the lifetime of the lower level τ_1, i. e. $\tau_2 > \tau_1$;
- if we add in a discussion of how the degeneracy of states contributes to the equation, then it turns out that $g_1 > g_2$ may also do the job, i. e. making sure that the population per state of the lower energy level is small. The corresponding necessary but not sufficient condition with degeneracies factored in is

$$\frac{g_2}{g_1}A_{21}\tau_1 < 1. \tag{54.9}$$

Now we know how to reach population inversion, let us think about how practically we can build a laser. In practice we have amplified the light by just pumping light into the medium with the atoms once, but humans have a natural inclination to greediness, we would like to re-direct the light back into the same medium and amplify it again, and re-direct the amplified light back to the same medium to amplify it once more, and repeat. The way to make this happen is discussed in the next few sections.

Summary

1. To reach amplification, we require $\Delta N > 0$, i. e. population inversion. Pumping with purely radiation in a medium of atoms with two levels is not possible, we need an alternative pumping mechanism.

2. For the alternative pumping mechanism to reach population inversion, we need both

   - the condition
   $$A_{21}\tau_1 < 1, \tag{54.10}$$
   - and at least one of the three:
   $$P_2 > P_1, \quad \tau_2 > \tau_1, \quad \text{or} \quad g_1 > g_2. \tag{54.11}$$

§55. Sustainable Lasing

Optical Gain

Now that population inversion has been achieved, let us amplify some light. This means that, we simply have the population inverted N_1 and N_2 as our given condition (so we disregard the mathematics by the pump mechanism discussed previously), and now we only look at terms relating to Einstein's B coefficients, focussing on how they affects the density of photons emitted. The absorption and stimulated emission processes decreases and increases the photon density of the laser medium by $BN_1\rho_{em}$ and $BN_2\rho_{em}$ respectively, giving the overall increase in photon density as

$$\dot{\mathscr{N}} = B(\Delta N)\rho_{em}, \tag{55.1}$$

where $\mathscr{N}$ is the number of photons per unit volume in the laser medium. However, what we are actually interested is how the intensity I, or energy per unit area per unit time, changes with distance x. To make this link, we shall first note that the intensity I and the energy density ρ_{em} is linked through

$$I = \frac{\text{energy}}{\text{area} \times \text{time}} = \frac{\text{energy}}{\text{volume}} \times \frac{\text{length}}{\text{time}} = \rho_{em}c, \tag{55.2}$$

then noticing that each particle has energy $\hbar\omega$. Therefore, the change in intensity $\mathrm{d}I$ against length $\mathrm{d}x$ is

$$\begin{aligned} \mathrm{d}I &= \frac{\text{energy}}{\text{area} \times \text{time}} = \text{length} \times \frac{\text{number of particles}}{\text{volume} \times \text{time}} \times \frac{\text{energy}}{\text{particle}} \\ &= \mathrm{d}x\,\dot{\mathscr{N}} \times \hbar\omega = \mathrm{d}x\,\hbar\omega B(\Delta N)I/c = \mathrm{d}x\,(\Delta N)\sigma I, \end{aligned} \tag{55.3}$$

where the **scattering cross-section** σ is defined by

$$\sigma = \hbar\omega B/c, \tag{55.4}$$

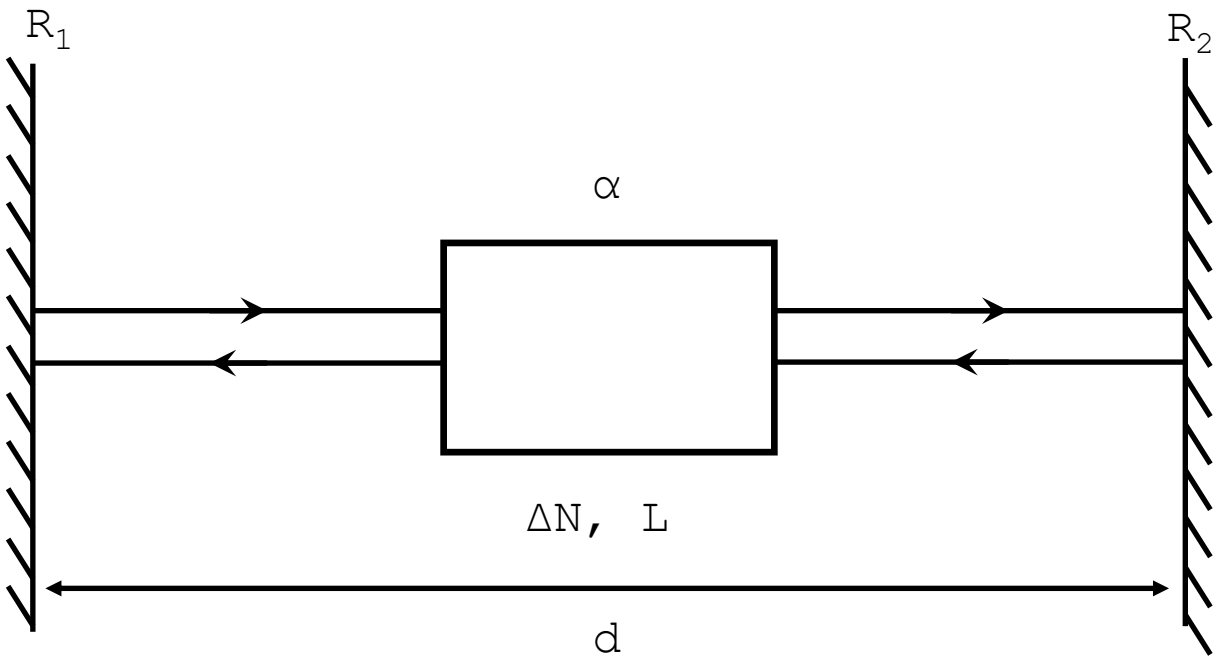

Figure 55.1: A laser in a Fabry-Perot cavity.

and we define the **gain coefficient**

$$\alpha \;=\; (\Delta N)\sigma. \tag{55.5}$$

Now that we have the laser medium set up with a gain, where each time some light goes through it gets amplified, let us maximise our greediness and amplify as much as possible.

Amplification by Oscillation

One of the methods (where we discuss other methods in §57) of amplifying the light repeatedly is to seal the laser medium in a Fabry-Perot cavity, where the light bounces between the two mirrors, and gets amplified by the laser medium every bounce. Note the reflectivities of the two mirrors R_1 and R_2 have values close to unity but not exactly. This means that there must be laser light escaping from both mirrors. Therefore, for the laser to lase, for every time the laser makes a round-trips between the mirrors, it must gain more intensity than it loses.

The intensity lost per round-trip can be expressed by the reflectances: it is simply $(1 - R_1R_2)I_{\text{circ}}$ as the remaining light between the mirrors is given by $R_1R_2I_{\text{circ}}$, where I_{circ} is the intensity circulating between the mirrors. The intensity gained per circulation is, rather apparently from 55.3, $\left[\mathrm{e}^{2(\Delta N)\sigma L} - 1\right]I_{\text{circ}}$, where L is the length of the laser medium, noting that light travels through the laser medium twice per circulation. Setting the intensity gain per round-trip to be larger than the intensity lost, we have

$$\left[\mathrm{e}^{2(\Delta N)\sigma L} - 1\right]I_{\text{circ}} \;\geqslant\; (1 - R_1R_2)I_{\text{circ}}, \tag{55.6}$$

where we note that the light travels through the laser medium twice per round-trip, and thus the distance in the medium per round-trip is given by $2L$. For R_1 and R_2 close to unity, the term in the brackets is very close to zero. Therefore, the above condition is algebraically equivalent to

$$\Delta N \;\geqslant\; N_{\text{threshold}} \;=\; \frac{1}{2L\sigma}(1 - R_1R_2). \tag{55.7}$$

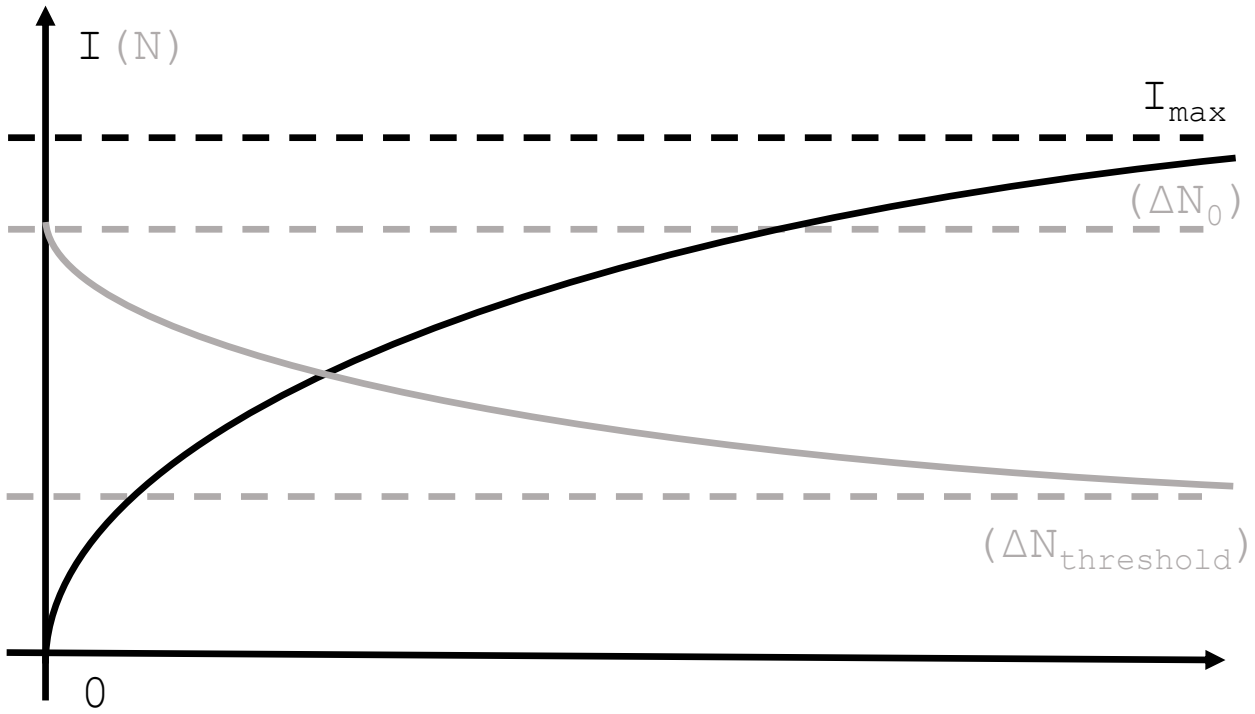

Figure 55.2: The number of particles (in grey) and the exit intensity (in black) against time for a laser after it has been turned on.

This clearly shows that population inversion alone is insufficient for sustainable lasing, we also need this stronger condition on top of population inversion.

Previously we have justified ourselves of ignoring the Einstein's B coefficients by suggesting this invention of the alternative pumping scheme, now we shall review our justification. When we are actually starting to think about what happens during lasing, there is an intensity circulating between the mirrors and passing through the laser media. This gives an energy density $\rho_{em} = I/c$ inside the system, which is how the terms involving Einstein's B coefficients emerge. Thus, we note that our treatment really only works in the case where there is no light intensity in the system; otherwise we immediately have the terms involving Einstein B coefficients kicking in, and these terms will start to equilibrate our population immediately after the laser starts running. Eventually the laser will reach a steady state with the population difference ΔN to be exactly $N_{\text{threshold}}$, and the laser will be running at maximum intensity. This process is demonstrated by figure 55.2.

Summary

1. When light passes through the laser medium, the change in intensity per unit distance is given by the equation
$$\mathrm{d}I = \mathrm{d}x\,(\Delta N)\sigma I = \mathrm{d}x\,\alpha I, \tag{55.8}$$
where σ is the scattering cross-section and α is the gain coefficient.

2. To amplify light sustainably we trap it under a Fabry-Perot cavity where the reflectivities of the two mirrors, R_1 and R_2, are very close to unity. When the laser is switched on, the existing intensities will equilibrate the populations to a steady state with threshold

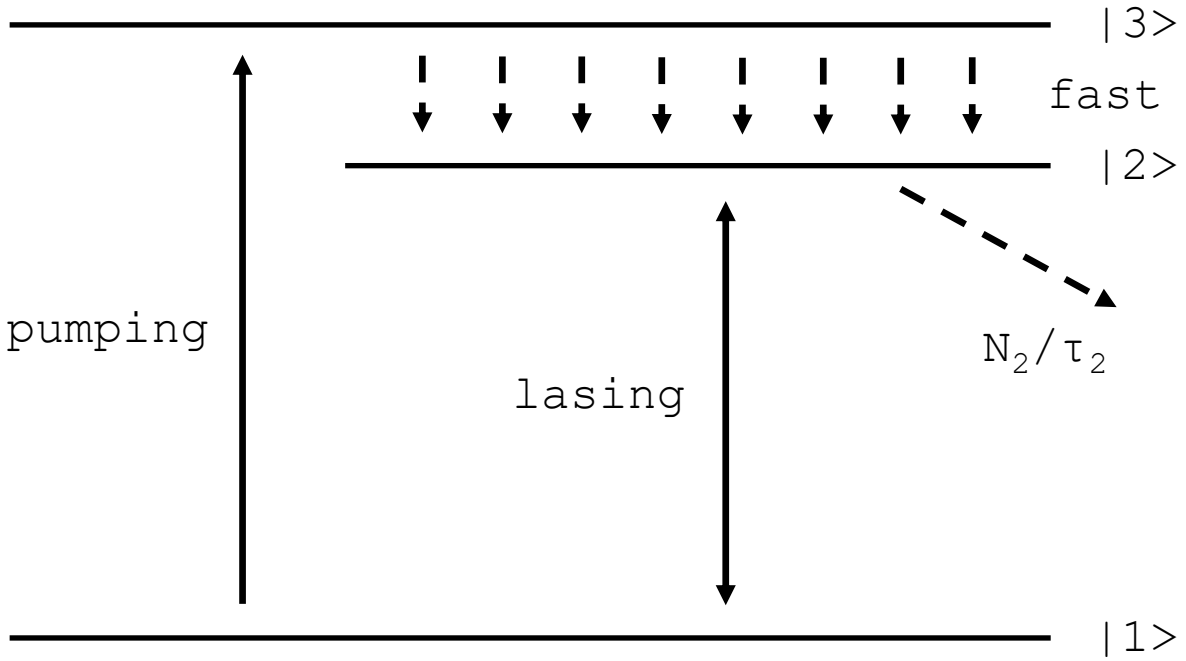

Figure 56.1: A three-level laser.

population difference

$$N_{\text{threshold}} \;=\; \frac{1}{2L\sigma}(1 - R_1R_2), \tag{55.9}$$

which, at that threshold population difference, the laser runs at maximum intensity.

§56. Laser Media

Three-Level Lasers

Previously we have illustrated that we need two main objects in the construction of a laser: the laser medium and the laser oscillator. We will talk about these two objects in a bit more detail in this section and the next section. In this section we look at different types of laser media. We have previously stated that radiative pumping cannot drive a two-level laser to population inversion, and as a result the only method to drive the laser using radiation is to change the two-level system completely, e. g. to use a three-level laser instead.

A schematic diagram for a three-level laser is given by figure 56.1. The pumping scheme pumps particles from states $|1\rangle$ to $|3\rangle$ at a rate P, where the particles experiences a fast decay process from state $|3\rangle$ to $|2\rangle$, and the lasing process happens between states $|2\rangle$ and $|1\rangle$. In this case, the population in state $|3\rangle$ is negligible, and the particles are all in states $|1\rangle$ and $|2\rangle$. This means that, for population inversion, we need at least half the atoms in the excited state, i. e. if we assume that the pump rate is the same as the decay rate, then

$$P \;=\; N_2/\tau_2 \qquad \Rightarrow \qquad P\tau_2 \;=\; N_2 \;>\; (N_1 + N_2)/2. \tag{56.1}$$

whichis difficult to achieve. Nevertheless, it could still be done, with a powerful pump mechanism. In the early ages of lasers, the lasers are simply pumped by flashlights. Alternatively, research groups at the Weizmann Institute of Science Solar Tower, Israel, have powered their

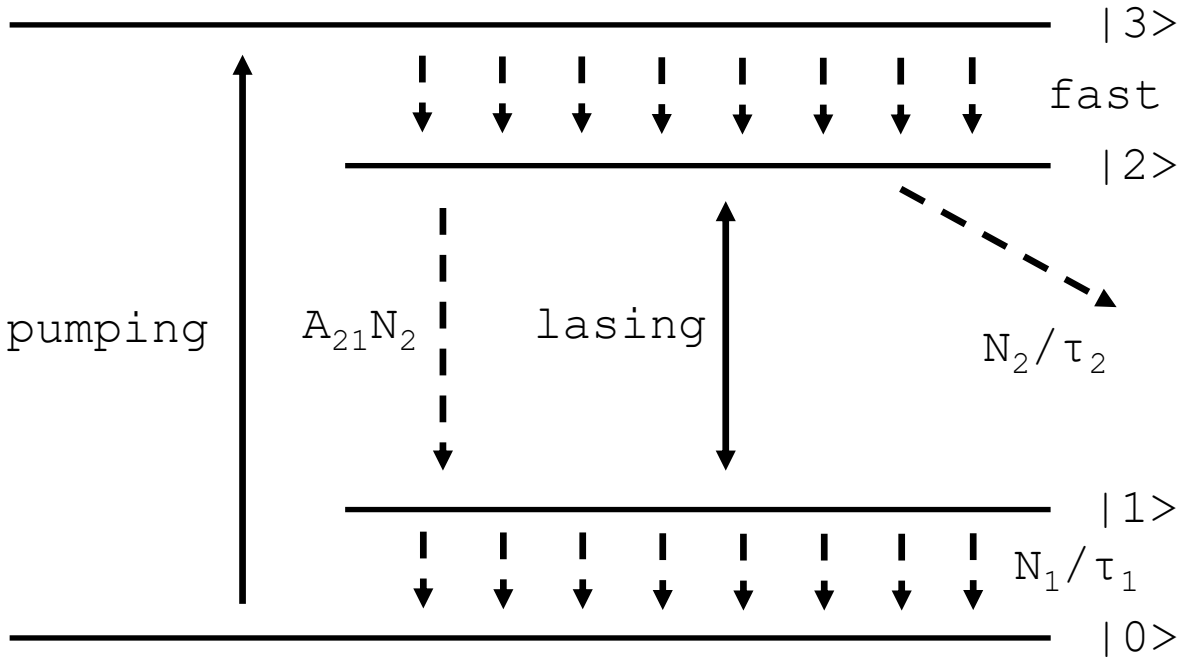

Figure 56.2: A four-level laser.

lasers with the Sun. Others powers their lasers by LEDs. We can also use one laser to pump another laser, for example, to lase at a different wavelength.

Four-Level Lasers

To get around the issue such that half the total population must be in the upper state, a more usual design for the laser medium is the four-level laser. A diagram of such a device works is figure 56.2. In a four level system, the transition from state $|1\rangle$ to state $|0\rangle$ is so fast that $|1\rangle$ is nearly empty. As long as the state is emptied out at a faster rate than it is filled in, we can lase sustainably, which means that we only need a condition

$$1/\tau_1 \; > \; A_{21}, \tag{56.2}$$

which is a lot easier to satisfy.

An example of the four-level laser is the Nd:YAG laser, which is based on an insulating solid, in which Nd^{3+} ions are dropped into an Yttrium Aluminium Garnet (YAG) crystalline host. In this case, levels $|1\rangle$ and $|2\rangle$ would be the two levels $^4I_{11/2}$ and $^4F_{3/2}$ levels for the Nd^{3+} ion, which has an energy gap allowing it to emit a wavelength of 1 064 nm. However, it is very common practice to use non-linear crystals to frequency double, triple, or quadruple the frequency of the laser, giving light with wavelengths of 532 nm, 354 nm, and 266 nm respectively. To pump the Nd:YAG laser, we can have either have it run on a continuous wave setup by pumping it with an LED, or run it on a pulsed setup using flashlight.

Summary

1. To design the laser medium for more convenient population inversion such that we can use radiation as a pumping scheme, we can use a three-level laser. For that to work, we

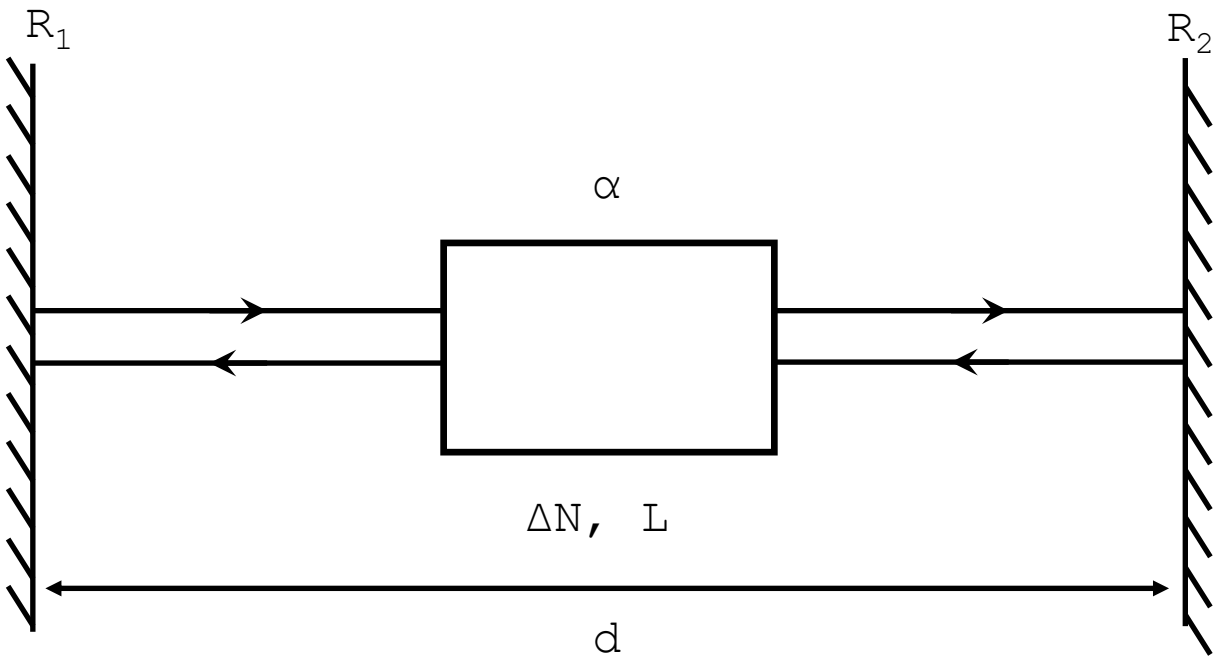

Figure 57.1: A Fabry-Perot cavity as a laser oscillator.

need at least half the atoms to be in the upper state, which is difficult to achieve, but can be achieved by powerful pumping method.

2. A four-level laser have very few atoms in state $|1\rangle$, which allows population inversion to be achieved more easily.

§57. Laser Oscillators

We shall then have a look at some designs of laser oscillators. The simplest laser oscillator is to have no laser oscillator. This gives the light simply amplified once with

$$I_{\text{out}} \;=\; I_{\text{in}}\mathrm{e}^{\alpha L}, \tag{57.1}$$

which simply amplifies the spontaneous emission, creating an **amplified spontaneous emission (ASE)** laser. Of course, there are more sophisticated laser oscillators, which we shall survey through the most popular two: the Fabry-Perot cavity and the ring resonator.

Fabry-Perot Cavity as a Laser Oscillator

Let us have a Fabry-Perot cavity as the laser oscillator, demonstrated in figure 57.1. We note that since there are mirrors on both ends, the boundary condition implies that the wavenumber inside must be an integer multiple of the fundamental, with a wavenumber $1/\lambda \;=\; 1/(2nd) \;=\; \mathrm{FSR}_{\bar{\nu}}$. Therefore the allowed wavenumbers are

$$\bar{\nu} \;=\; p \times \mathrm{FSR}_{\bar{\nu}} \;=\; p \times \frac{1}{2nd}, \tag{57.2}$$

where p is an integer. However, the lasers that we will be using are monochromatic i. e. have one wavenumber only. To do this, we simply match the frequency of the atomic transition

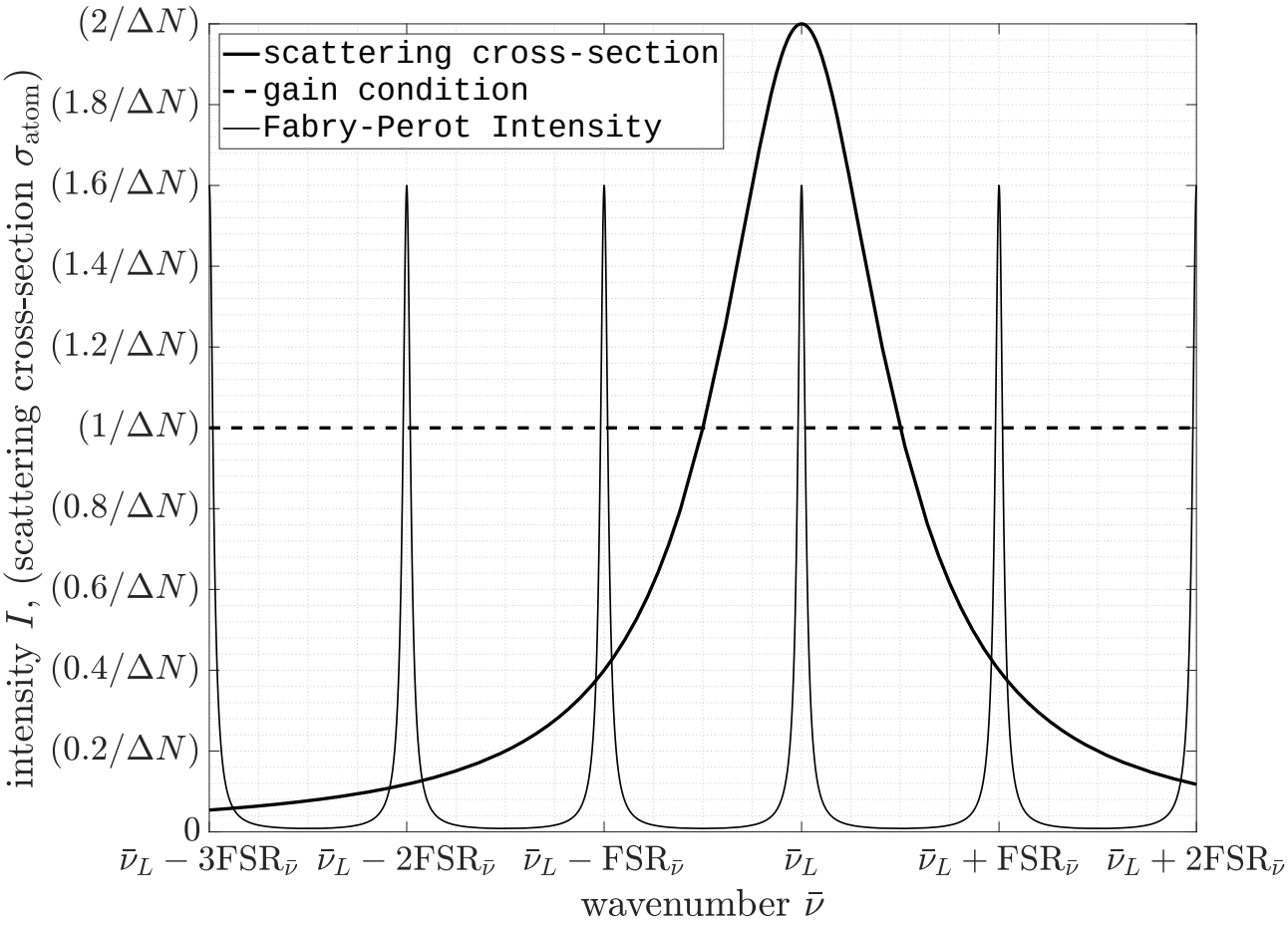

Figure 57.2: The intensity and the scattering cross-section of the atom against wavenumber. Note that the vertical tick marks is for the scattering cross-section; the Fabry-Perot intensity has relative scale, which is not displayed.

to a specific mode of the standing wave trapped in the Fabry-Perot cavity. However there is an additional complication, which is the fact that although we can assume the frequency of the atomic transition to be happening at one frequency only, i. e. we assume the scattering cross-section is only applicable for one single frequency, it is in fact a Lorentzian function in wavenumber. In order to lase only one mode, we need to make sure that the scattering cross-section function drops off below the gain condition quickly enough such that no other mode is lased. This is illustrated by figure 57.2, where single-mode lasing has been successfully achieved, noting that the scattering cross-section function is only larger than $1/\Delta N$ for a single wavenumber $\bar{\nu}$, which is the lasing wavenumber $\bar{\nu}$, the wavenumber that is going to lase. We note that in this specific case illustrated by figure 57.2, we have the width of the scattering cross-section larder than the instrumental width of the Fabry-Perot cavity, which is generally the case. As a rule of thumb, the width of the actual laser beam will be narrower than both of them, i. e.

$$\Delta\bar{\nu}_{\text{atom}} \;>\; \text{INST}_{\bar{\nu}} \;>\; \Delta\bar{\nu}_{\text{laser}}. \tag{57.3}$$

In a Fabry-Perot driven oscillator, there will be a standing wave in the medium, which means that the intensity will be varying in space. Near the nodes, the intensity is small, means that the electric field will also be small at all times, meaning that the population inversion will

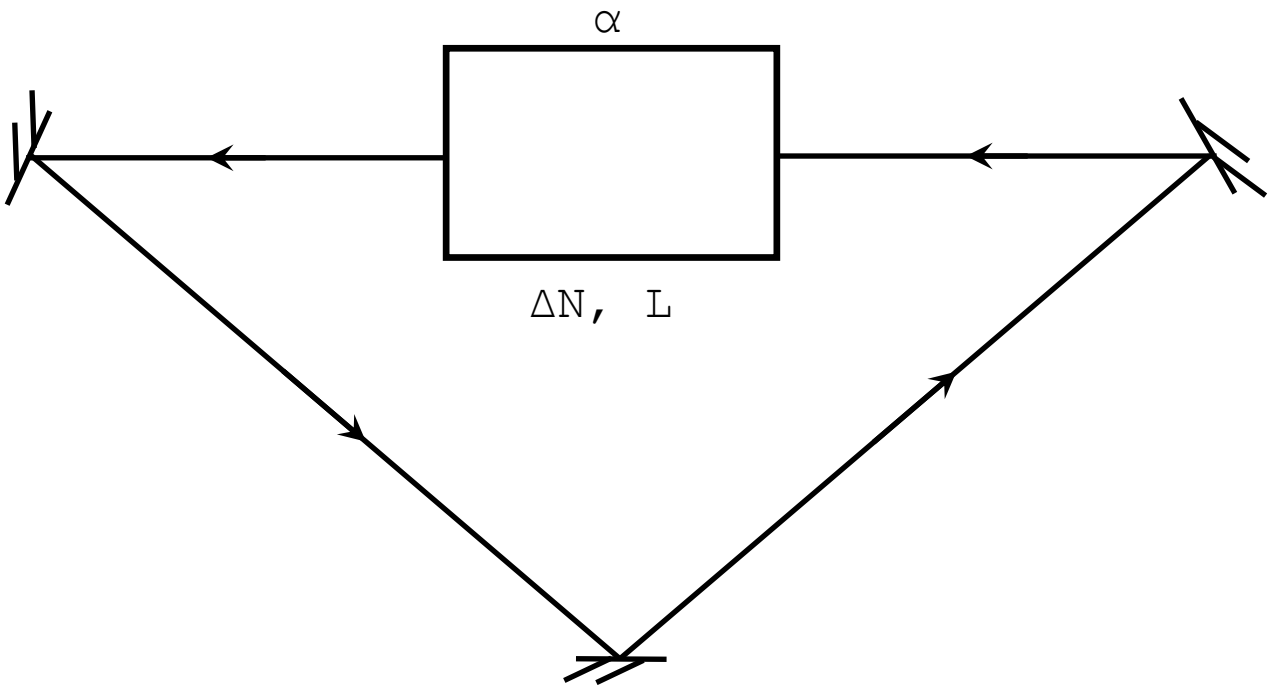

Figure 57.3: A ring resonator.

struggle to achieve, which is an effect called **spatial hole burning**. This motivates another method, called the ring resonator.

Ring Resonator

A ring resonator, with an example illustrated in figure 57.3, directs light into a ring with circumference L. This now have the allowed wavenumbers as

$$\bar{\nu} = p/(nL), \tag{57.4}$$

where p is an integer. This allows light to be trapped in a ring and, by a similar effect to the Fabry-Perot resonator, amplifies light by directing it many times through the laser medium. Note that the arbitrary integer p still enters the condition for maximum, and as a result we still need to put into effort to making the laser to run at a single wavelength to make sure that the laser is running at a single frequency.

Summary

1. A Fabry-Perot cavity can be an oscillator of light. It amplifies light every round-trip, however it can allow many different wavenumbers to be amplified simultaneously. To create a laser running at a single wavelength, we must make sure that the scattering cross-section function is larger than $1/\Delta N$ for one frequency only, which is the wavenumber that the laser lases. The cavity will have nodes which has a lower electric field, causing spatial hole burning, limiting the amplified intensity.

2. A ring resonator directs light in a closed ring, and hence passing the light through the laser medium many times, reaching amplification.

11 Polarisation

§58. Polarised v Unpolarised Light

Now we shall switch gears and look at light specifically as a transverse wave, and examine the additional behaviour light has due to its transverse nature, which is polarisation. We shall first look at Maxwell's equation in vacuum and then look at Maxwell's equation in an anisotropic medium, such as a crystal, where light behaves very differently compared to an isotropic medium, which we shall look into and consider how this can be exploited in experiments.

Electric Field Vector

Previously we have looked at Maxwell's equations in a vacuum, where Gauß's law

$$\operatorname{div}\mathbf{E} \;=\; 0 \quad\Rightarrow\quad \mathbf{ik}\cdot\mathbf{E} \;=\; 0 \tag{58.1}$$

gives the fact that $\mathbf{E}$ must be perpendicular to $\mathbf{k}$, and Faraday's law

$$\operatorname{curl}\mathbf{E} \;=\; -\partial_t\mathbf{B} \quad\Rightarrow\quad \mathrm{i}\mathbf{k}\wedge\mathbf{E} \;=\; \mathrm{i}\omega\mathbf{B} \tag{58.2}$$

gives the fact that $\mathbf{E}$ must be perpendicular to $\mathbf{B}$. We have suggested that these two facts motivates the notation

$$u \;=\; u_0\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\,\mathrm{e}^{-\mathrm{i}\delta} \quad\Rightarrow\quad \operatorname{Re}[u] \;=\; \operatorname{Re}\left[u_0\,\mathrm{e}^{\mathrm{i}(kx-\omega t-\delta)}\right] \;=\; |u_0|\cos(kx-\omega t-\delta); \tag{58.3}$$

however this is a simplification: as $\mathbf{E}$ itself can be at any direction perpendicular to $\mathbf{k}$, and therefore can be thought as *any* vector lying on a plane, the scalar representation must be insufficient as multiple $\mathbf{E}$s may correspond to the same u. Let us then consider the wave travelling in the x-direction, where $\mathbf{E}$ is decomposed into two directions y and z, and therefore

$$E_y \;=\; E_{0y}\cos(kx-\omega t-\delta_y); \tag{58.4}$$
$$E_z \;=\; E_{0z}\cos(kx-\omega t-\delta_z). \tag{58.5}$$

Re-casting this into a simple vector form and removing a constant phase, we end up with

$$\begin{pmatrix}E_y\\E_z\end{pmatrix} \;=\; \operatorname{Re}\begin{pmatrix}E_{0y}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\\E_{0z}\,\mathrm{e}^{\mathrm{i}(kx-\omega t-\delta)}\end{pmatrix}, \tag{58.6}$$

where $\delta \;=\; \delta_z-\delta_y$. We define the light to be **polarised** if the phase difference between the two directions of the field, δ, is constant.

Linear Polarisation

The simplest case that we can look into is **linear polarisation**, which has $\delta \;=\; 0$ or $\delta \;=\; \pi$. In the case where $\delta \;=\; 0$, or equivalently $\delta_y \;=\; \delta_z$, we have

$$\begin{pmatrix}E_y\\E_z\end{pmatrix} \;=\; \operatorname{Re}\begin{pmatrix}E_{0y}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\\E_{0z}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}\end{pmatrix}, \tag{58.7}$$

which suggests that the electric field vector $\mathbf{E}$ fixed along a certain plane that is at an angle α from the y axis, with

$$\tan\alpha \;=\; \frac{E_{0z}}{E_{0y}}. \tag{58.8}$$

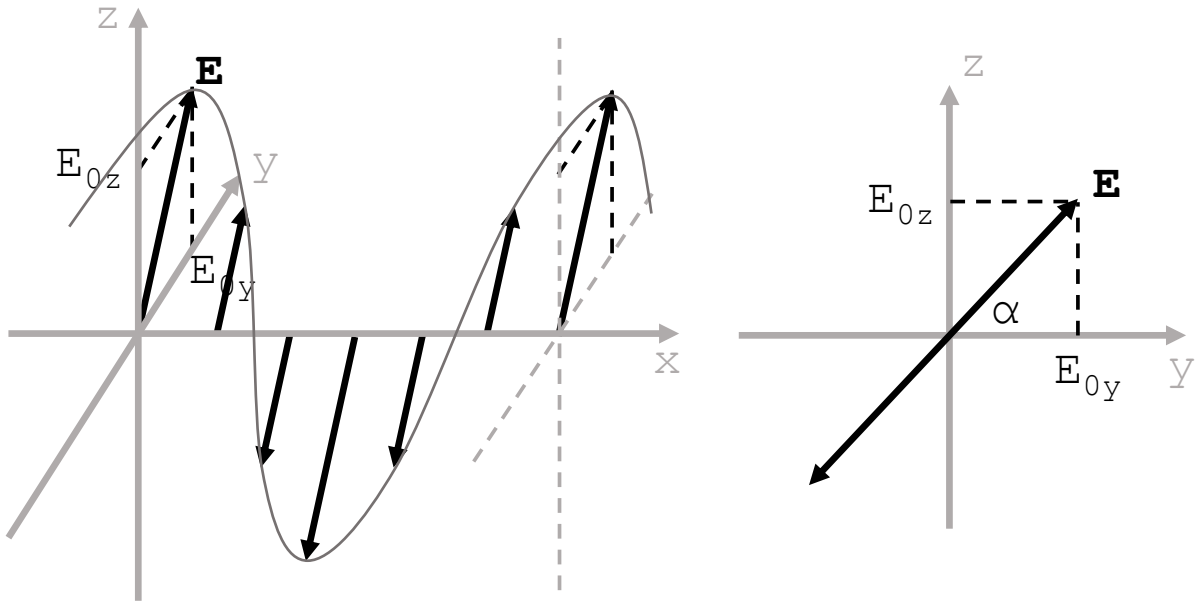

Figure 58.1: Linearly polarised light with $\delta = 0$.

This is shown in figure 58.1. An alternative option for linearly polarised light is $\alpha = \pi$, where we simply have

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} = \mathrm{Re}\begin{pmatrix} E_{0y}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)} \\ -E_{0z}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)} \end{pmatrix}. \tag{58.9}$$

This gives the exact same picture of the previous case, but simply with the two components of the electric field with opposite signs. This is demonstrated in figure 58.2.

So what is the difference between linearly polarised light with unpolarised light? The difference lies in the fact that unpolarised light has the phase difference between the y and z components of the electric field

$$\delta = \delta_z - \delta_y \tag{58.10}$$

completely random at any time. Such light is therefore also **incoherent**.

Summary

1. The two components of the electric field vector **E** perpendicular to the direction of wave travel can have a phase difference δ between them.

2. If δ is random then the light is unpolarised. If δ is either 0 or π then the light is linearly polarised.

§59. Elliptically Polarised Light

Circularly Polarised Light

We have previously looked at the case where δ is either 0 or π. Another special case is $\delta = \pi/2$

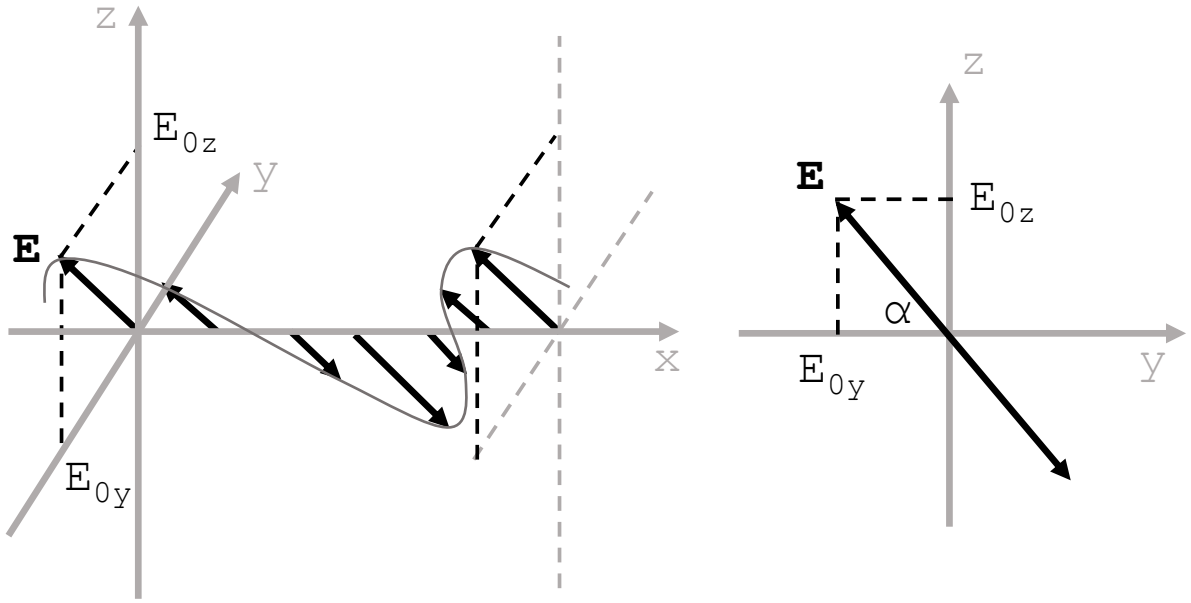

Figure 58.2: Linearly polarised light with $\delta = \pi$.

and $E_{0y} = E_{0z} = E_0$. In that case we have the electric field vector $\mathbf{E}$ as

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} = E_0 \,\mathrm{Re}\begin{pmatrix} \mathrm{e}^{\mathrm{i}(kx-\omega t)} \\ \mathrm{e}^{\mathrm{i}(kx-\omega t-\pi/2)} \end{pmatrix} = E_0 \begin{pmatrix} \cos(kx-\omega t) \\ \sin(kx-\omega t) \end{pmatrix}, \tag{59.1}$$

which means that if we fix time and look at the different values of $\mathbf{E}$ along the direction of wave propagation, it traces out a helix with a circular cross-section. If we then point our thumb of our right hand into the direction of wave propagation, we then have the four fingers curling to the direction that the $\mathbf{E}$ field is next going into. As a result, the polarisation of this wave is called **right-hand circular polarisation**. This is illustrated in figure 59.1. An alternative method is to look at the source, i. e. fix $x = 0$, and then look at how $\mathbf{E}$ rotates as time increases. We then find the electric field on a circle that is rotating clockwise. In this case, if we curl the four fingers of our right hand around the electric field vector, then the thumb points towards the source — an alternative take to the name "right-hand circular polarisation".

Another type of circular polarsiation is the case where $\delta = -\pi/2$ and $E_{0y} = E_{0z} = E_0$. In this case, we have

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} = E_0 \,\mathrm{Re}\begin{pmatrix} \mathrm{e}^{\mathrm{i}(kx-\omega t)} \\ \mathrm{e}^{\mathrm{i}(kx-\omega t+\pi/2)} \end{pmatrix} = E_0 \begin{pmatrix} \cos(kx-\omega t) \\ -\sin(kx-\omega t) \end{pmatrix}. \tag{59.2}$$

Analysed analogously to the above way for right-hand circularly polarised light, it is clear that this is given the name **left-hand circular polarisation**. This is also illustrated in figure 59.2. Sometimes left-handed circularly polarised light is described as "positive helicity".

We note that although circularly polarised light also has its intensity decreased by a factor of two when passed through a linear polariser — the same behaviour as unpolarised light, since

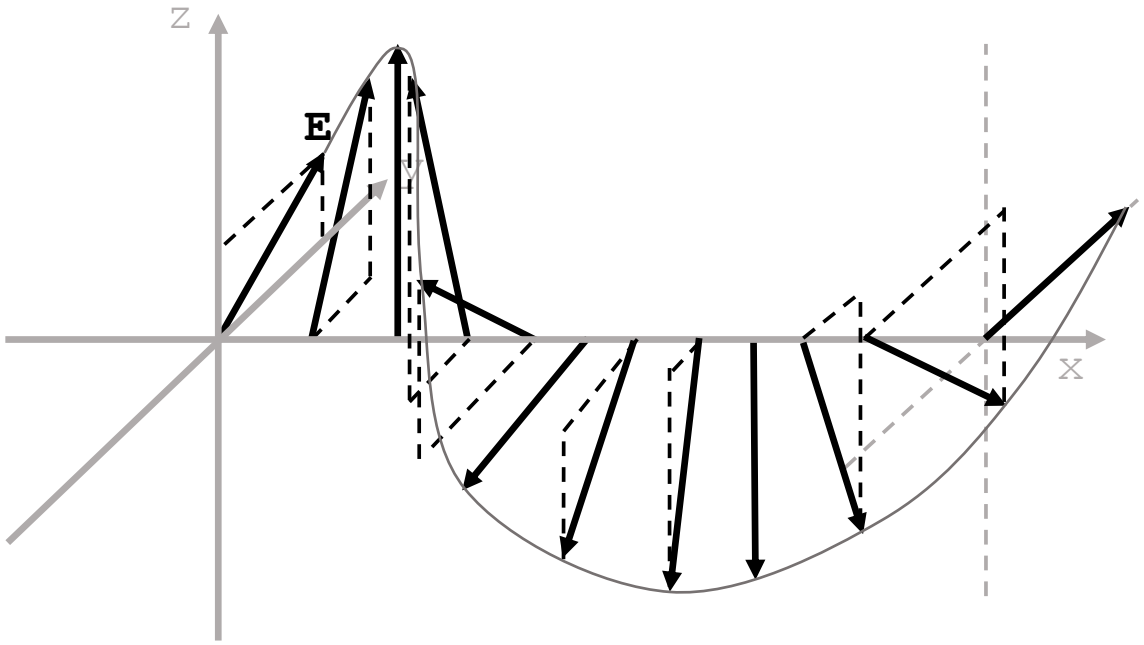

Figure 59.1: The electric field **E** in right-handed circularly polarised light.

it has a constant δ, it cannot be classified as "unpolarised radiation" — they are completely different animals.

Elliptically Polarised Light

We shall then take the next step of generalisation, which is to release the condition of $E_{0y} = E_{0z}$, which then gives

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} = \begin{pmatrix} E_{0y}\cos(kx-\omega t) \\ \pm E_{0z}\sin(kx-\omega t) \end{pmatrix}, \tag{59.3}$$

where when looking at the source, the electric field traces out an ellipse with the two semi-axes as E_{0y} and E_{0z} respectively. These two semi-axes lies on the y and z axes.

If we then release our constraint further, this time with a completely general value of δ, then we have elliptically polarised light with the two semi-axes aligned in an oblique angle Θ to the y- and z- axes, which link is discussed in §60. Both of these cases are demonstrated in figure 59.3. To show this algebraically, consider the real form of E_y and E_z, i. e.

$$E_y = E_{0y}\cos(kx-\omega t) \tag{59.4}$$

and

$$E_z = E_{0z}\cos(kx-\omega t-\delta) = E_{0z}[\cos(kx-\omega t)\cos\delta + \sin(kx-\omega t)\sin\delta]. \tag{59.5}$$

Rearranging equation 59.5, we have

$$\begin{aligned} \sin^2(kx-\omega t)\sin^2\delta &= \left[\frac{E_z}{E_{0z}} - \cos(kx-\omega t)\cos\delta\right]^2 \\ &= \frac{E_z^2}{{E_{0z}}^2} - 2\frac{E_z}{E_{0z}}\cos(kx-\omega t)\cos\delta + \cos^2(kx-\omega t)\cos^2\delta. \end{aligned} \tag{59.6}$$

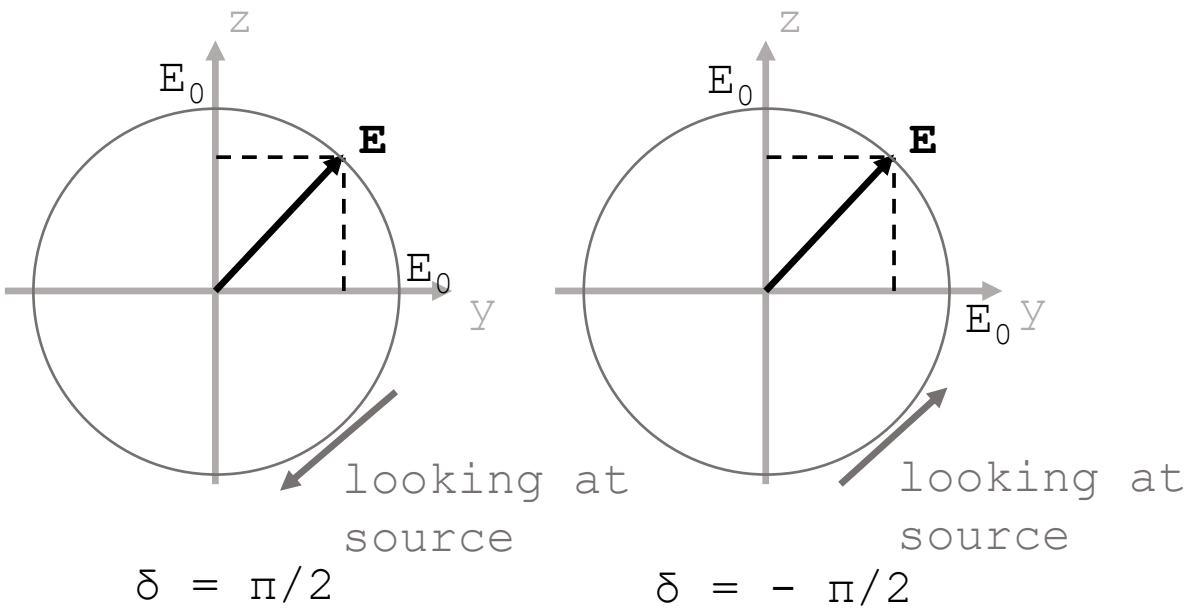

Figure 59.2: The electric field **E** in right- ($\delta = \pi/2$) and left- ($\delta = -\pi/2$) handed circularly polarised light.

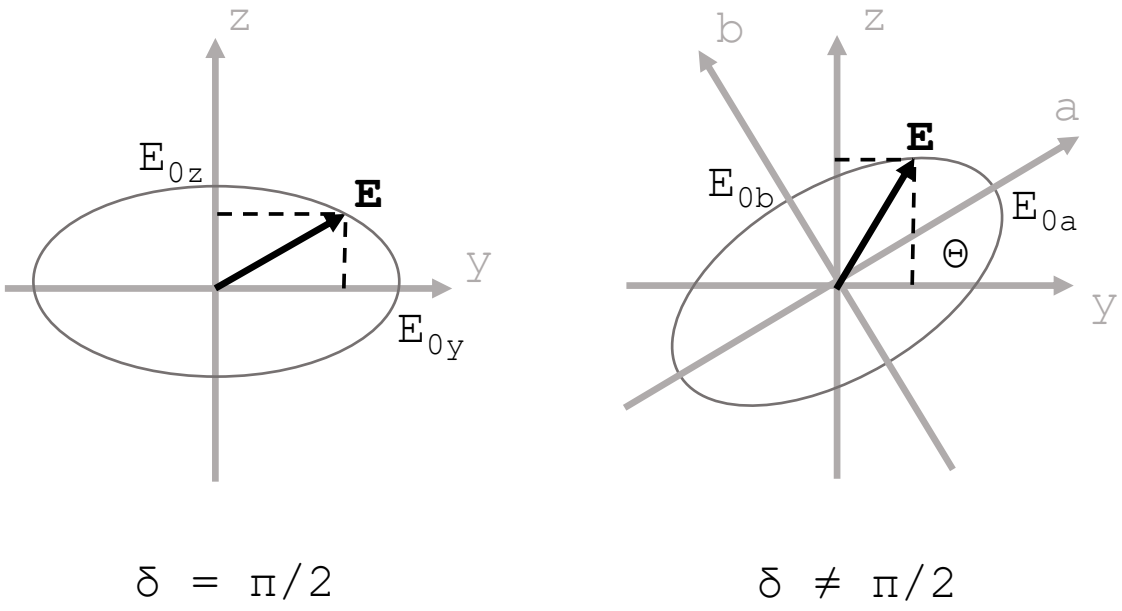

Figure 59.3: The electric field **E** in elliptically polarised light.

We may then rearrange equation 59.4 to get

$$\cos(kx - \omega t) \;=\; E_y/E_{0y}, \qquad \sin^2(kx - \omega t) \;=\; \left[1 - (E_y/E_{0y})\right]^2, \tag{59.7}$$

which we may substitute into equation 59.6 to obtain

$$\frac{E_y^2}{{E_{0y}}^2} + \frac{E_z^2}{{E_{0z}}^2} - \frac{2}{E_{0y}E_{0z}} E_y E_z \cos\delta \;=\; \sin^2\delta, \tag{59.8}$$

the equation of an ellipse. When $\delta \;=\; \pm\pi/2$ we then have

$$E_y^2/{E_{0y}}^2 + E_z^2/{E_{0z}}^2 \;=\; 1, \tag{59.9}$$

i. e. the ellipse lines up with the y- and z- axes.

For elliptically polarised light with $\delta \;\neq\; \pm\pi/2$, we can simplify the mathematics by looking at them obliquely, by setting up a set of axes a and b parallel and perpendicular to the semi- and major- axes. When the radiation is viewed in that way, it is polarised with $\tilde{\delta} \;=\; \delta_a - \delta_b \;=\; \pm\pi/2$. Then, we define

$$\mathbf{E}_a \;=\; E_{0a} \begin{pmatrix} \cos\Theta \\ \sin\Theta \end{pmatrix} \cos(kx - \omega t); \tag{59.10}$$

$$\mathbf{E}_b \;=\; \pm E_{0b} \begin{pmatrix} -\sin\Theta \\ \cos\Theta \end{pmatrix} \sin(kx - \omega t), \tag{59.11}$$

where $(\cos\Theta, \sin\Theta)^{\mathrm{T}}$ and $(-\sin\Theta, \cos\Theta)^{\mathrm{T}}$ are the unit vectors along the a and b axes respectively, and describe the electric field vector $\mathbf{E}$ with

$$\mathbf{E} \;=\; \mathbf{E}_a + \mathbf{E}_b. \tag{59.12}$$

Now, with this introduced parameter Θ, we can say that all polarised radiation, after a rotation of axes, have $\tilde{\delta} \;=\; 0, \pi$, or $\pi/2$.

Summary

1. Circularly polarised light has the electric field vector $\mathbf{E}$ describing a circle. It has the phase difference between y and z as $\delta \;=\; \pm\pi/2$, corresponding to right- and left-handed circular polarisation, and also the amplitudes $E_{0y} \;=\; E_{0z} \;=\; E_0$.

2. Elliptically polarised light with $\delta \;=\; \pm\pi/2$ has the electric field vector $\mathbf{E}$ describing an ellipse with the semi-axes aligned with y and z directions and with lengths E_{0y} and E_{0z}. If $\delta \neq \pi/2$, then the light is still elliptically polarised, and we are able to find an angle Θ such that if we rotate our coordinate axes by Θ we have $\tilde{\delta} \;=\; \delta_a - \delta_b \;=\; \pm\pi/2$.

§60. More on Circularly and Elliptically Polarised Light

Relation between δ and Θ

We shall first explore the relation between the phase difference δ and the angle Θ. Our starting point is equation 59.8

$$\frac{E_y^2}{{E_{0y}}^2} + \frac{E_z^2}{{E_{0z}}^2} - \frac{2}{E_{0y}E_{0z}} E_y E_z \cos\delta \;=\; \sin^2\delta. \tag{60.1}$$

We recognise this equation as a quadratic form

$$\epsilon^{\mathrm{T}}\Omega\epsilon \;=\; \sin^2\delta, \qquad \epsilon \;=\; \begin{pmatrix} E_y \\ E_z \end{pmatrix}, \qquad \Omega \;=\; \begin{pmatrix} \dfrac{1}{{E_{0y}}^2} & -\dfrac{\cos\delta}{E_{0y}E_{0z}} \\ -\dfrac{\cos\delta}{E_{0y}E_{0z}} & \dfrac{1}{{E_{0z}}^2} \end{pmatrix}, \tag{60.2}$$

for which the orientations of the eigenvectors of the matrix Ω points to the a- and b- axes. We should have one possible eigenvector along a-axis to be

$$\mathbf{a} \;=\; \left(\frac{\cos\delta}{E_{0y}E_{0z}}, \quad \frac{1}{2{E_{0y}}^2} - \frac{1}{2{E_{0z}}^2} + \frac{\sqrt{{E_{0y}}^4 + {E_{0z}}^4 + 2{E_{0y}}^2{E_{0z}}^2\cos(2\delta)}}{2{E_{0y}}^2{E_{0z}}^2} \right)^{\mathrm{T}}, \tag{60.3}$$

as a result we have

$$\Theta \;=\; \arctan\left(\frac{{E_{0z}}^2 - {E_{0y}}^2 + \sqrt{{E_{0y}}^4 + {E_{0z}}^4 + 2{E_{0y}}^2{E_{0z}}^2\cos(2\delta)}}{2E_{0y}E_{0z}\cos\delta} \right), \tag{60.4}$$

an explicit relation between δ and Θ.

Superposition of circularly polarised radiation

We note that if we superpose left- and right-hand circularly polarised radiation with equal amplitudes, we have

$$E_0 \begin{pmatrix} \cos(kx-\omega t) \\ \sin(kx-\omega t) \end{pmatrix} + E_0 \begin{pmatrix} \cos(kx-\omega t) \\ -\sin(kx-\omega t) \end{pmatrix} \;=\; 2E_0\cos(kx-\omega t)\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{60.5}$$

linearly polarised radiation in the y direction. Sometimes we depict the quantum mechanical state of the photon as $|R\rangle$ for a right-handed circularly polarised photon, and $|L\rangle$ for a left-handed circularly polarised photon. Then, we can write down

$$\frac{1}{\sqrt{2}}\Big(|R\rangle + |L\rangle \Big) \;=\; |H\rangle\,, \tag{60.6}$$

where $|H\rangle$ is a horizontally polarised photon. If the two amplitudes are unequal, then we have the resulting radiation elliptically polarised. An illustration of this is shown in figure 60.1.

Summary

1. By expressing equation 59.8 as a quadratic form and finding its eigenvectors, we obtain the angle Θ as

$$\Theta \;=\; \arctan\left(\frac{{E_{0z}}^2 - {E_{0y}}^2 + \sqrt{{E_{0y}}^4 + {E_{0z}}^4 + 2{E_{0y}}^2{E_{0z}}^2\cos(2\delta)}}{2E_{0y}E_{0z}\cos\delta} \right). \tag{60.7}$$

2. If we superpose left- and right-handed circularly polarised light, we get linearly polarised light if the amplitudes are equal; and elliptically polarised light if not.

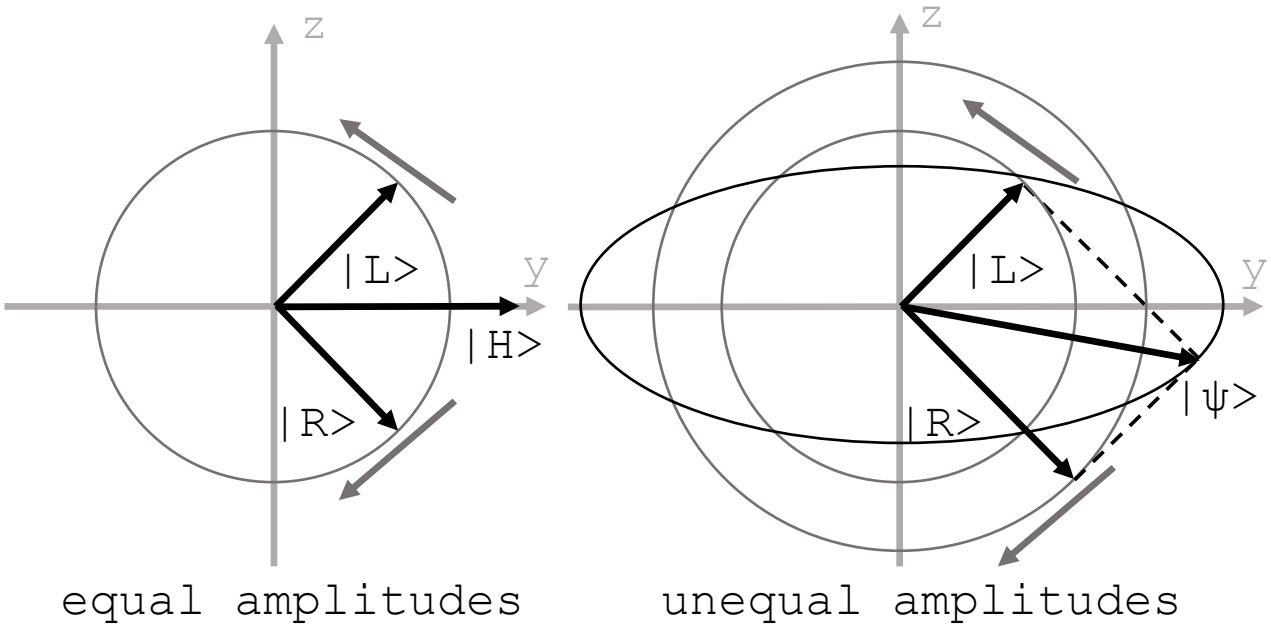

Figure 60.1: The superposition of radiation with left- and right-circular polarisations.

§61. Crystal Optics

Electromagnetic Waves in Anisotropic Media

Now we move to Maxwell's equation in a medium that is anisotropic, i. e. that has different behaviours when light is radiated into the medium at different directions. To briefly motivate how radiation will change in an anisotropic medium, we consider Gauß's law in a medium

$$\operatorname{div} \mathbf{D} \;=\; 0 \quad \Rightarrow \quad \mathrm{i}\mathbf{k} \cdot \mathbf{D} \;=\; 0. \tag{61.1}$$

If the material is anisotropic, then $\mathbf{D}$ and $\mathbf{E}$ will no longer be parallel, and therefore $\mathbf{E}$ and $\mathbf{k}$ will not longer be perpendicular, which is an immediate consequence of the anisotropy of the medium. More precisely, if we consider Faraday's law and Ampère's law in a material with $\mu_r \;=\; 1$

$$\operatorname{curl} \mathbf{E} \;=\; -\partial_t \mathbf{B} \;=\; -\mu_0 \partial_t \mathbf{H} \quad \Rightarrow \quad \mathrm{i}\mathbf{k} \wedge \mathbf{E} \;=\; \mathrm{i}\omega\mu_0 \mathbf{H}; \tag{61.2}$$

$$\operatorname{curl} \mathbf{H} \;=\; \partial_t \mathbf{D} \quad \Rightarrow \quad \mathrm{i}\mathbf{k} \wedge \mathbf{H} \;=\; -\mathrm{i}\omega \mathbf{D}, \tag{61.3}$$

we end up with

$$\mathbf{k} \wedge (\mathbf{k} \wedge \mathbf{E}) \;=\; \mu_0 \omega \mathbf{k} \wedge \mathbf{H} \;=\; -\mu_0 \omega^2 \mathbf{D}, \tag{61.4}$$

a more precise relation between $\mathbf{E}$ and $\mathbf{D}$.

After a second look on Ampère's law, we find that $\mathbf{H}$ is perpendicular to both $\mathbf{k}$ and $\mathbf{D}$. Since we are looking at the case where $\mathbf{D}$ and $\mathbf{E}$ are unparallel, we note that the Poynting vector $\mathbf{S} \;=\; \mathbf{E} \wedge \mathbf{H}$ is not parallel to the $\mathbf{k}$. To summarise, in an anisotropic medium,

- $\mathbf{E}$ is perpendicular to $\mathbf{S}$,
- $\mathbf{D}$ is perpendicular to $\mathbf{k}$,

- **E**, **S**, **D**, and **k** are on the same plane perpendicular to **H**.

After analysing the waves that are in an anisotropic medium in its full generality just based on Maxwell's equations, now let us switch gear to actually analyse the problem from the microscopic perspective, by looking into a crystal.

Crystals

To describe the relation between **D** and **E**, in an anisotropic media, we simply turn the relative permittivity into a 3×3 matrix (or, for the mathematically inclined students, a tensor of second rank), i. e. we write down

$$\mathbf{D} = \varepsilon_0\varepsilon\mathbf{E} = \varepsilon_0 \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix} \mathbf{E}. \tag{61.5}$$

This is the most general case but here we take a very simplified approach. First we shall take the all the off-diagonal elements to be 0. Note that, doing so means that we are choosing three very special axis in space — if we were to rotate this set of coordinate axes, then this simplification will not work. Applying this simplification leaves us with three options only:

- In the case of $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} = \varepsilon_r$, we have an **isotropic** material. For materials with $\mu_r = 1$, we have $\varepsilon = n^2$, where n is the refractive index of the material. To account for this we simply need to change the path length into the optical path length by multiplying by n, which was already well-discussed in the first year optics course.
- In the case of $\varepsilon_{11} \neq \varepsilon_{22} = \varepsilon_{33}$, we have an anisotropic "**uni-axial**" material. We then define $n_o^2 = \varepsilon_{22} = \varepsilon_{33}$ and call the axis corresponding to them (here it is the y- and z-axes) the **ordinary axes**. Also we have $n_e^2 = \varepsilon_{11}$ which we call the axis corresponding to it (here it is the x axis) the **extra-ordinary axis**. This is the case of interest for us.
- In the case of $\varepsilon_{11} \neq \varepsilon_{22} \neq \varepsilon_{33}$, we have an anisotropic "**bi-axial**" material. This is beyond the level of our discussion.

From the next section onwards, we shall focus on the uni-axial crystal only.

Summary

1. In an anisotropic media, in general **E** is perpendicular to **S**, **D** is perpendicular to **k**, and **E**, **S**, **D**, and **k** are all perpendicular to **H**; but **E** is neither parallel to **D** nor perpendicular to **k**, instead they are linked by the relation
$$\mathbf{k}\wedge(\mathbf{k}\wedge\mathbf{E}) = -\mu_0\omega^2\mathbf{D}. \tag{61.6}$$
2. Crystals with $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33}$ are isotropic and crystals with $\varepsilon_{11} \neq \varepsilon_{22} = \varepsilon_{33}$ are uni-axial. For uni-axial crystals, along the ordinary axis $n_o = \sqrt{\varepsilon_{22}} = \sqrt{\varepsilon_{33}}$ and along the extra-ordinary axis $n_e = \sqrt{\varepsilon_{11}}$.

§62. Uni-Axial Crystals

Light Propagation in Uni-Axial Crystals

We shall confine ourselves to an uni-axial crystal with $\varepsilon_{11} = n_e^2$ and $\varepsilon_{22} = \varepsilon_{33} = n_o^2$. Then we can split our discussion into three different cases:

- **E** is in the y-z plane, and we have the refractive index as n_o. In this case we have **D** parallel to **E**, and the speed of light propagation is given by $v_o = c/n_o$. To achieve this we need **k** in the x-direction, and therefore suggesting **k** and **S** are parallel.
- **E** is in the x-direction, and we have the refractive index as n_e. In this case we have **E** parallel to **E** and $v_e = c/n_e$. Again **k** and **S** are parallel. Unpolarised light simply cannot achieve this as **E** can be placed in two different directions.
- **E** is not parallel to any of the axes. In this case, for example, **E** has components along x and z axes,

$$\mathbf{D} = \varepsilon_0 \begin{pmatrix} \varepsilon_{11} & 0 & 0 \\ 0 & \varepsilon_{33} & 0 \\ 0 & 0 & \varepsilon_{33} \end{pmatrix} \begin{pmatrix} E_x \\ 0 \\ E_z \end{pmatrix}, \tag{62.1}$$

  In this case the phase velocity of the light is given by $v = c/n_{\text{eff}}$, where the effective refractive index n_{eff} is between n_o and n_e. More quantitative analysis shows that n_{eff} forms an ellipsoid in space, with semi-axes n_e, n_o, n_o respectively along the x, y, and z axes.

Next we shall look into the final case in more details.

Linearly Polarised Light in Uni-Axial Crystals

We now focus on an arbitrarily directed light ray with **S** pointing in a general direction, and we shall assume that it is linearly polarised with a well defined **E** vector. We decompose the electric field **E** into two linearly polarised rays with electric field vectors $\mathbf{E}_o$ and $\mathbf{E}_e$, demonstrated by figure 62.1. The electric field component that is parallel to the y-z plane, $\mathbf{E}_o$, is electric field vector that gives rise to the **ordinary ray**, and the field component that is parallel to both **E** and $\mathbf{E}_o$, which we denote by $\mathbf{E}_e$, gives rise to the **extra-ordinary ray**. We take special notice that the extra-ordinary ray, in general, does not travel parallel to the extra-ordinary axis.

Since the ordinary wave is confined within the y-z plane, it travels with **k** and **S** parallel to each other. However the extra-ordinary wave is not parallel to any of the three special axes, and as a result it travels with **k** and **S** not parallel to each other. It turns out that, this leads to the consequence that the extra-ordinary ray will not experience Snell's law. This suggests that a single beam of linearly polarised light whose electric field vector **E** is able to travel in two different directions in a crystal, corresponding to the ordinary and extra-ordinary components of the wave, even with the wave incident on the crystal along the normal. This effect is demonstrated in figure 62.2. As a result, uni-axial crystals are sometimes called **birefringent crystals**. Note that this property can also be generalised into an unpolarised beam, as we are also able to decompose the unpolarised beam into two different components with electric field **E** parallel and perpendicular to the direction of travel.

Summary

1. When the electric field vector **E** is along one of the axes of the special coordinate system such that the matrix ε is diagonal, **k** and **S** are parallel, and **D** and **E** are parallel. However if that is not the case then the vectors suggested are not parallel.
2. When light passes inside the uni-axial crystal it will split into an ordinary ray and an extra-ordinary ray. The extra-ordinary ray will not follow Snell's law in general, and therefore the two rays will split — and therefore we say the crystal is "birefringent".

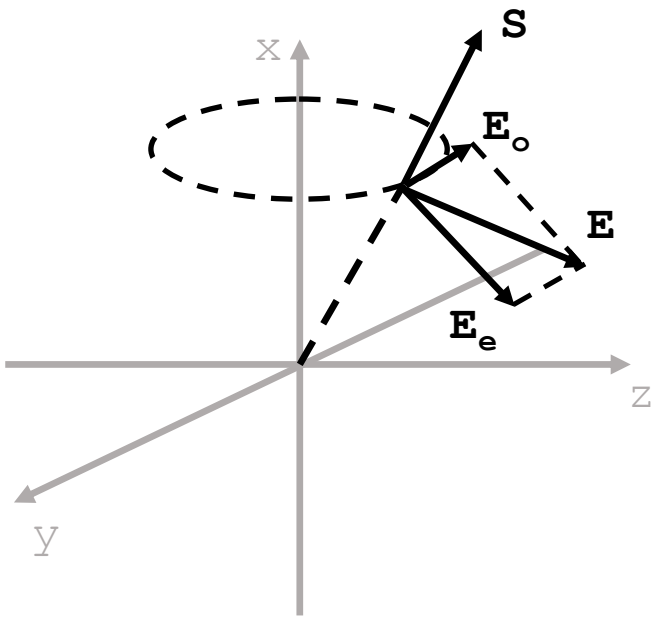

Figure 62.1: The electric field vector **E** decomposed along the ordinary and extra-ordinary axes.

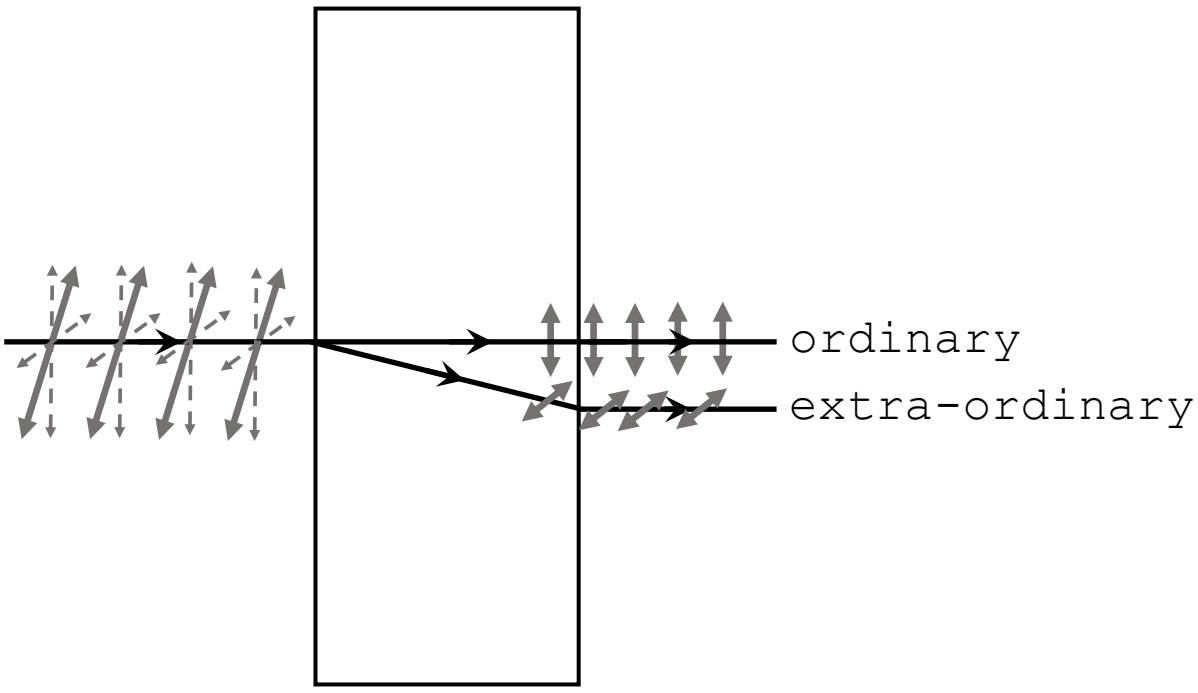

Figure 62.2: A linearly polarised ray passing through an uni-axial crystal.

§63. Retardation of Polarisation

When do extra-ordinary rays satisfy Snell's law?

In the previous section we have stated that extra-ordinary rays do not, in general, satisfy Snell's law, because the extra-ordinary rays has two components that are travelling in different speeds in the crystal, and as a result $\mathbf{k}$ and $\mathbf{S}$ are not parallel. However here are two special cases where both the ordinary ray and extra-ordinary rays satisfy Snell's law.

- When $\mathbf{S}$ is along the extra-ordinary axes, then $\mathbf{E}$ will be located in the plane spanned by vectors on the ordinary axes. In such case, the wave will have $\mathbf{S}$ parallel to $\mathbf{k}$, with a refractive index $n_o = \sqrt{\varepsilon_{22}} = \sqrt{\varepsilon_{33}}$.
- When $\mathbf{S}$ is in the plane spanned by vectors on the ordinary axes, then we must have $\mathbf{E}_o$ along the ordinary axis *and* $\mathbf{E}_e$ along the extra-ordinary axis. The latter statement is special and only applies in this case. Now both the ordinary and extra-ordinary rays travels with $\mathbf{k}$ and $\mathbf{S}$ parallel, with the ordinary ray with $n_o = \sqrt{\varepsilon_{22}} = \sqrt{\varepsilon_{33}}$ and the extra-ordinary ray with $n_e = \sqrt{\varepsilon_{11}}$.

Apparently the first case is rather boring as it would require the same correction as a wave in an isotropic medium. We shall therefore spend some time looking at the second case.

Waveplates

Now we have the two rays incident normally into the crystal and therefore the ordinary and extra-ordinary ray will travel along the same direction, as they both satisfy Snell's law. However, the ordinary ray will "see" a refractive index n_o, and the extra-ordinary ray will "see" a refractive index n_e, which are not the same. If $n_e < n_o$, then we say that the crystal has **negative anisotropy**, and if $n_e > n_o$, then we say that the crystal has **positive anisotropy**. We then call the axis of the $\mathbf{E}$ vector for the ray that travels quicker in the crystal the **fast axis** with refractive index n_f, and the axis of the $\mathbf{E}$ vector for the ray that travels slower in the crystal the **slow axis** with refractive index n_s. Therefore the slow ray is retarded relative to the fast ray. This leads to a shift in phase difference between the fast and slow rays.

The apparatus that exploits this property is called a **waveplate**, which is just a cylindrical sheet of uni-axial crystal with the axis aligned such that if the incident light ray is normal to the surface of the bottom of the cylinder, then $\mathbf{S}$ is in the plane spanned by vectors on the ordinary axes. An example of this is given in figure 63.1. It is clear that in the crystal the difference between optical path lengths ΔOPL of the slow and fast rays are given by

$$\Delta\text{OPL} = \Delta n \times \Delta y, \tag{63.1}$$

where $\Delta n = n_s - n_f$ is the difference between the refractive indices, and Δy is the height of the cylinder, demonstrated by figure 63.1. There are two most common waveplates.

- $\lambda/2$ waveplates have $\Delta s = \left(p + \frac{1}{2}\right)\lambda$, where p is an integer. The phase retardation of the slow ray is given by

$$\Delta\delta = k \times \Delta\text{OPL} = \frac{2\pi}{\lambda} \times \frac{\lambda}{2} = \pi, \tag{63.2}$$

  where we have neglected the phase shift by an integer number of 2π as that simply does not change the wave. This type of plates can flip one of the polarisation components in linearly polarised light, as depicted in figure 63.1. This type of plate can shift the handiness of elliptically polarised light if the semi-axis of the polarisation exactly matches the fast and slow axes.

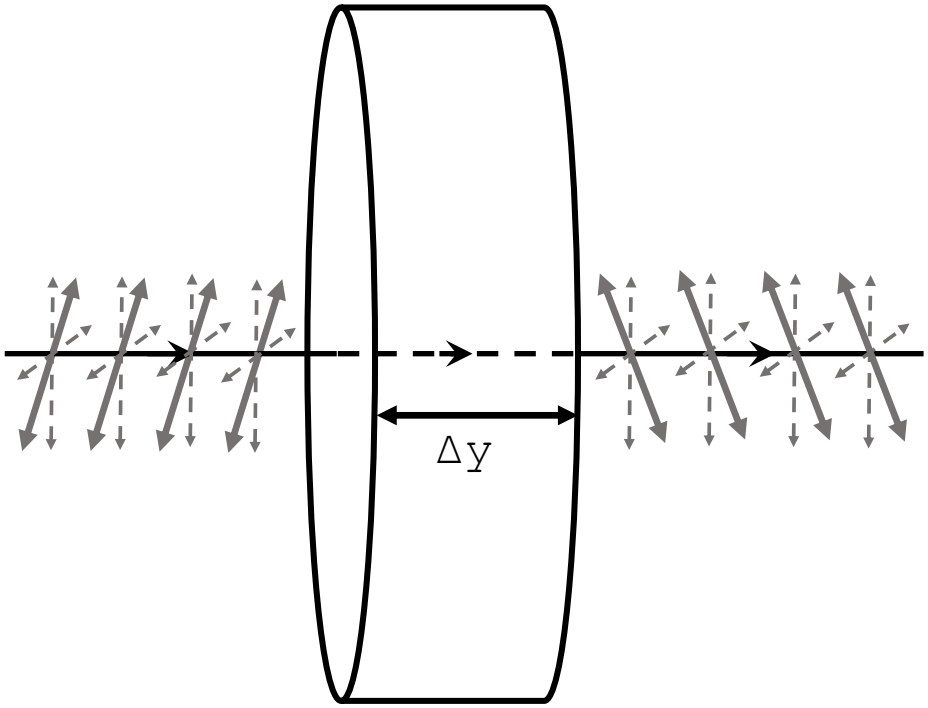

Figure 63.1: An example of a $\lambda/2$ waveplate.

- $\lambda/4$ waveplates have $\Delta\text{OPL} = \left(p + \frac{1}{4}\right)\lambda$, where p is an integer. The phase retardation of the slow ray is given by

$$\Delta\delta = k \times \Delta\text{OPL} = \frac{2\pi}{\lambda} \times \frac{\lambda}{4} = \frac{\pi}{2}. \tag{63.3}$$

  This type of plates can change linearly polarised light into elliptically polarised light with the semi axes aligned with the fast and slow axes of the waveplate, and the length of the semi-axes of the elliptically polarised light exactly matching the components of the electric field vector **E** on the fast and slow axes for the incident ray. Also, if we match the fast and slow axes of the crystal to the semi-axes of an incident elliptically polarised ray, it will turn the ray into a linearly polarised ray. This is demonstrated in figure 63.2. In practise, the waveplate is usually mounted on a steel frame which allows the waveplate to be freely rotated, such that we can align the axes onto any direction we want it to be.

Note that, if we were to create a waveplate with a required wavelength with $\Delta\text{OPL} = \lambda/2$ or $\lambda/4$, i. e. with $p = 0$, the thickness required would be at the order of the wavelength of the material, which is too thin such that the waveplate would be mechanically unstable. This means that either we make the waveplate thicker by adding an integer number of full wavelengths, i. e. make p non-zero, or we can sandwich two waveplates with the fast and slow axes oriented 90° to each other with thicknesses Δy_1 and Δy_2. If we let the first waveplate to have the fast axis oriented in the vertical direction, then we have

$$\text{OPL}_{\text{vertical}} = n_f(\Delta y_1) + n_s(\Delta y_2) \tag{63.4}$$

$$\text{OPL}_{\text{horizontal}} = n_s(\Delta y_1) + n_f(\Delta y_2), \tag{63.5}$$

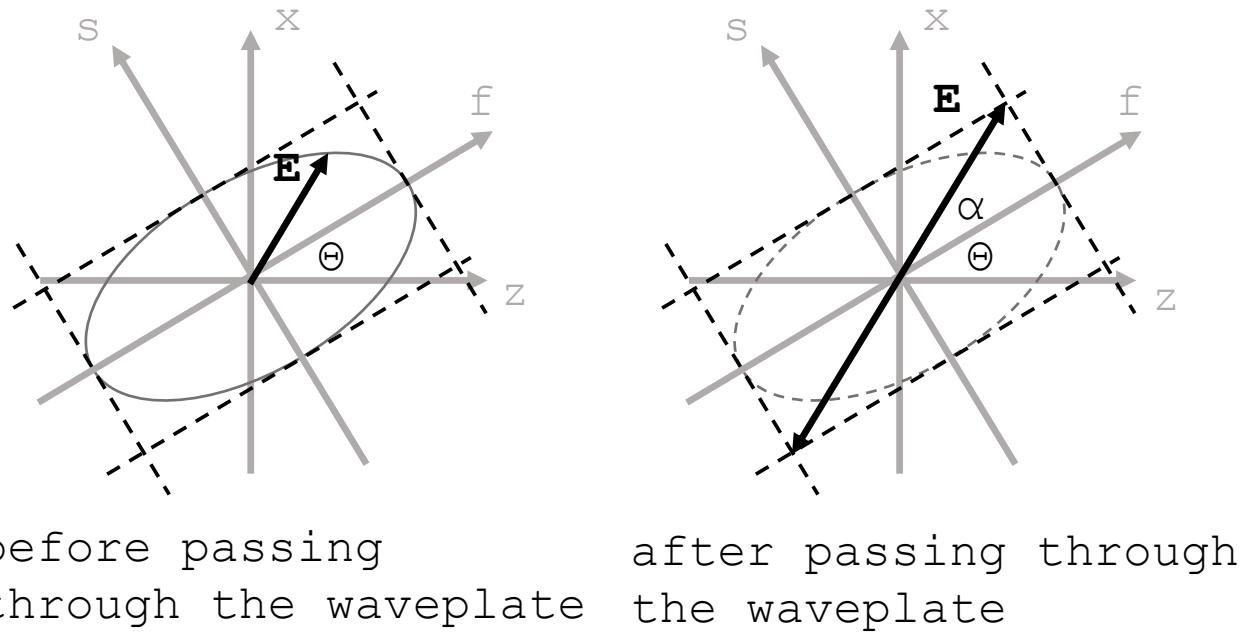

Figure 63.2: An illustration of how a $\lambda/4$ waveplate changes elliptically polarised light into linearly polarised light, when the fast axis of the waveplate aligns with one of the semi-axis of the ellipse described by **E** of the elliptically polarised light. Here the light is travelling along the y-axis.

therefore to make the optical path length difference between the vertical and horizontal components of the light to be $\lambda/4$, for example, we simply need

$$\lambda/4 \;=\; (\Delta y_1 - \Delta y_2) \times (n_f - n_s), \tag{63.6}$$

which allows us to make thick waveplates that are mechanically stable.

Summary

1. When **S** is in the plane spanned by vectors along the two ordinary axes, then we have both rays satisfying Snell's law with refractive indices n_o and n_e for the ordinary and extra-ordinary rays.

2. Either the ordinary ray or the extra-ordinary ray can be the fast ray. We can pass rays through waveplates with thickness Δy, and the optical path difference caused by the retardation of the slow ray is given by

$$\Delta\text{OPL} \;=\; \Delta n \times \Delta y. \tag{63.7}$$

   When linearly polarised rays pass through $\lambda/2$ waveplates one of the components will swap its sign. When they pass through $\lambda/4$ waveplates it will turn into elliptically polarised ray with the semi-axes matching the fast and slow axes of the waveplates. If the incident ray is elliptical, and the axes of the polarised light matches the fast and slow axes of the waveplate, then a $\lambda/2$ waveplate will swap the handiness and a $\lambda/4$ waveplate will make the ray linearly polarised.

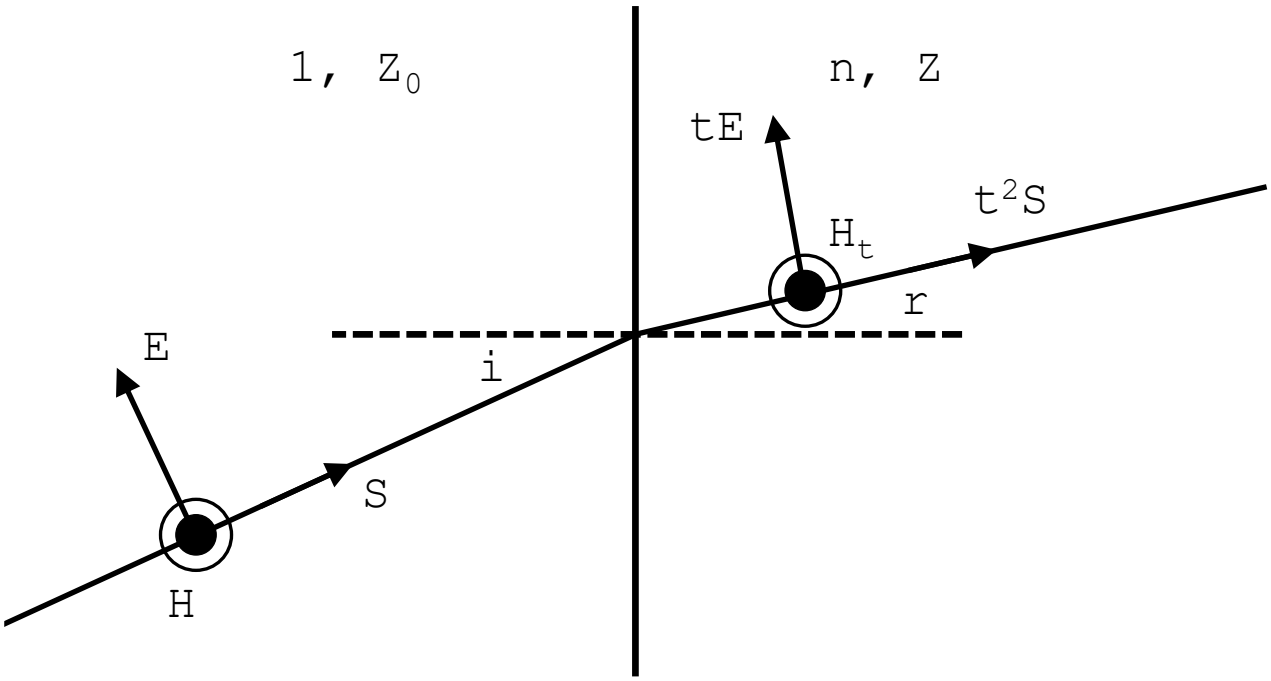

Figure 64.1: An illustration of the incident and transmitted rays upon a surface for a p-polarised light.

§64. Creation of Polarised Light

There are a number of methods for creating polarised light, and we shall go through them one by one.

Creation of Linearly Polarised Light by Reflection

We are able to create polarised light just by reflection the incident unpolarised ray off an isotropic material. When an polarised light approaches the isotropic material, for any given **S** we have two different modes of the electric field vector **E**, s-polarisation and p-polarisation. It is possible to show that we are able to extinguish the reflection of p-polarised ray at the **Brewster angle** as follows.

The incident and transmitted p-polarised light is shown in figure 64.1, with transmission coefficient t and the reflected ray extinguished. They are demonstrated in figure 64.1. We equip the isotropic medium with refractive index n and impedance Z, and air with refractive index 1 and impedance Z_0. Also we let the ray to be incident at the Brewster's angle i and refracted at an angle of refraction r. We note that for $\mu_r = 1$, we have

$$\frac{Z}{Z_0} = \frac{\sqrt{\mu_0/(\varepsilon_0\varepsilon_r)}}{\sqrt{\mu_0/\varepsilon_0}} = \frac{1}{\sqrt{\varepsilon_r}} = \frac{1}{n}. \tag{64.1}$$

According to Maxwell's equations, the parallel components of **E** and **H** must be continuous across the boundary. We set the magnitudes of **E** and **H** before and after hitting the material

to be E, H, tE, and H_t respectively. Equating these components, we yield

$$E\cos i \;=\; tE\cos r \tag{64.2}$$

$$H \;=\; H_t \quad\Rightarrow\quad \frac{E}{Z_0} \;=\; \frac{tE}{Z} \quad\Rightarrow\quad t \;=\; \frac{Z}{Z_0} \;=\; \frac{1}{n}. \tag{64.3}$$

Now substituting this back into the first equation, we obtain

$$\cos i \;=\; \frac{1}{n}\cos r \;=\; \frac{1}{n}\sqrt{1-\sin^2 r} \;=\; \frac{1}{n}\sqrt{1-\frac{1}{n^2}\sin^2 i}, \tag{64.4}$$

using Snell's law $\sin i \;=\; n\sin r$. Squaring this and dividing by $\cos^2 i$, after rearrangement, gives

$$\frac{1}{n^2}\sec^2 i - \frac{1}{n^4}\tan^2 i \;=\; 1. \tag{64.5}$$

Then, using the trigonometric identity $\sec^2 i \;=\; 1+\tan^2 i$ and rearranging the equation, we obtain the relation between the Brewster's angle and the refractive index

$$\tan i \;=\; n. \tag{64.6}$$

For an s-polarised beam, it is not possible to extinguish the reflected ray, and as an result, if we input the unpolarised ray incident to the surface of an isotropic material with refractive index n at the Brewster's angle $i \;=\; \arctan n$, we will find that the reflected beam will be linearly polarised with s-polarisation.

Creation of Linearly Polarised Light by Linear Polarisers

This apparatus is well-explained in any A-level syllabus on physics and therefore will not be re-introduced here — it simply changes any type of light incident on it into linearly polarised light polarised along its transmission axis. An additional comment is that if we have a linearly polarised light incident on a linear polariser, but at an angle θ about the transmission axis, then the magnitude of the electric field E will drop by

$$E \to E\cos\theta. \tag{64.7}$$

As a result, the intensity will drop by

$$I \to I\cos^2\theta, \tag{64.8}$$

commonly known as **Malus's law**.

Creation of Linearly Polarised Light by Polarising Prisms and Beam Splitters

We have stated, in the previous section, that when **S** is in the plane spanned by vectors on the ordinary axes, then the rays satisfies Snell's law, but the ordinary and extra-ordinary rays have different refractive indices. This can be exploited by inputting a ray incident to the crystal normally and making the ray to exit the material at an oblique angle, which makes a **polarising prism**, illustrated by the diagram on the left of figure 64.2. According to Snell's law, the angle of refraction on exiting the prism of the ordinary and extra-ordinary rays r_o and r_e is given by

$$\sin r_o \;=\; n_o \sin i; \tag{64.9}$$

$$\sin r_e \;=\; n_e \sin i, \tag{64.10}$$

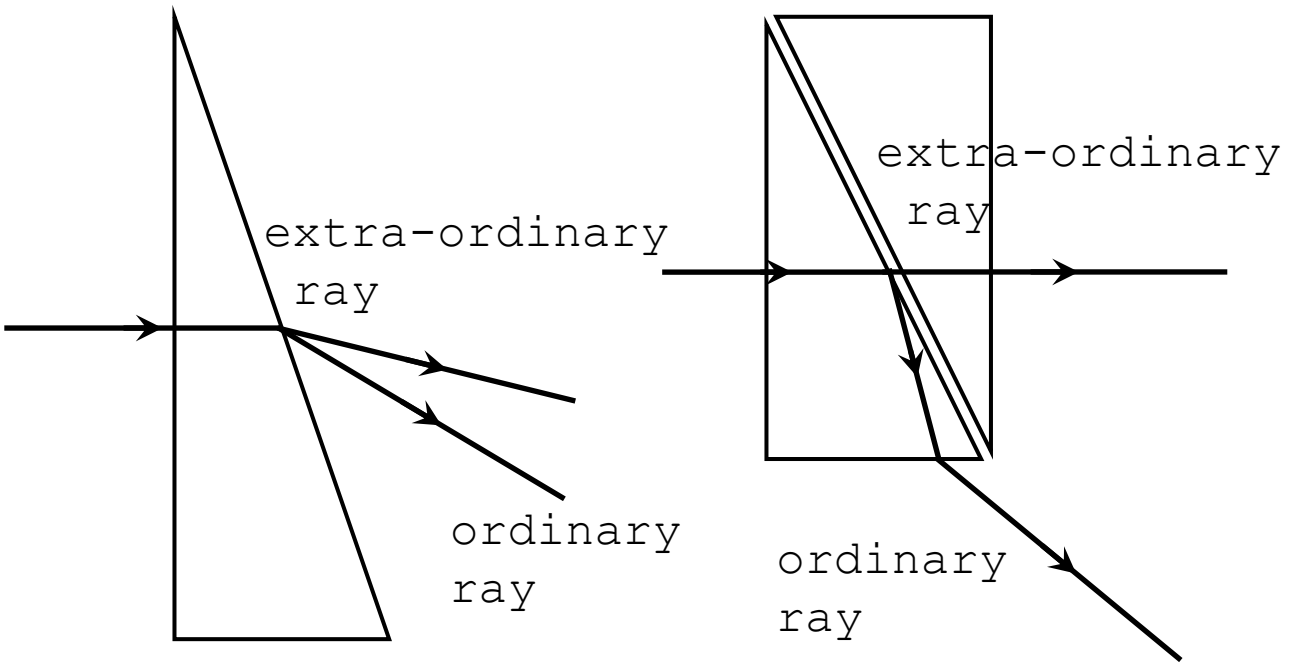

Figure 64.2: Demonstration of setups of a polarising prism and a polarising beam splitter with uni-axial crystals.

where i is the angle of incidence when the rays exit the uni-axial crystal. Alternatively, we can consider a setup where the ordinary ray experiences total internal reflection inside the prism, and the extra-ordinary transmits out of the system, which would make a **polarising beam splitter**. Apparently, this can be achieved by setting

$$\frac{1}{n_0} < \sin i < \frac{1}{n_e}, \tag{64.11}$$

where again i is the angle of incidence when the rays exit the uni-axial crystal.

Creation of Circularly Polarised Light

This can be achieved by the following sequence.

- First pass the unpolarised light through a linear polariser, which changes the light into linearly polarised light.
- Then pass the resulting linearly polarised light through a $\lambda/4$ waveplate, aligning the fast and slow axes of the waveplate 45° to the transmission axis of the linear polariser to make sure that the component electric field strength on the fast and slow axes are equal, so the exiting light is circularly polarised and not elliptically polarised.

Finally we shall focus on examining the polarisation of the light in the next section.

Summary

1. If an incident light is incident on an isotropic material at Brewster's angle $i = \arctan n$, then the reflected beam will be s-polarised only.

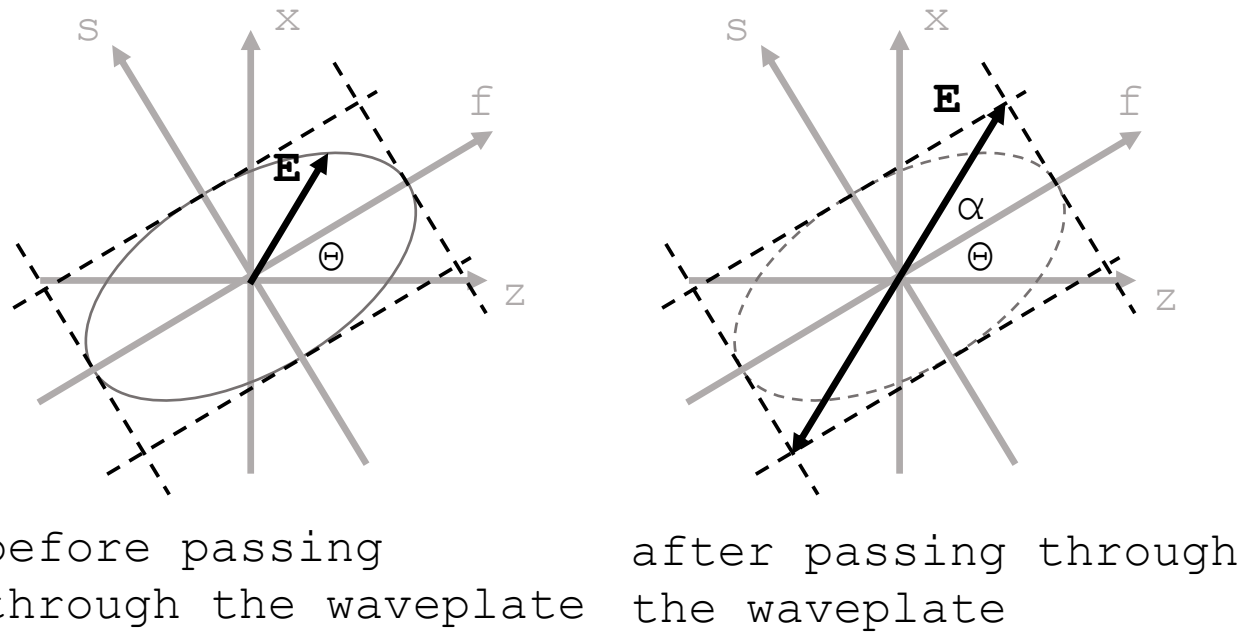

Figure 65.1: A repeat of figure 63.2, which is an illustration of how a $\lambda/4$ waveplate changes elliptically polarised light into linearly polarised light, when the fast axis of the waveplate aligns with one of the semi-axis of the ellipse described by **E** of the elliptically polarised light. Here the light is travelling along the y-axis.

2. A linear polariser polarises any incident light linearly along its transmission axis. If the input light is linearly polarised and the direction polarisation is at an angle θ from the transmission axis, then the intensity will decrease by Malus's law

$$I \;\rightarrow\; I\cos^2\theta. \tag{64.12}$$

3. Using uni-axial crystals we are able to make polarsing prisms and beam splitters, just by splitting the rays using the property that the different polarisations with **E** along the fast and slow axis have different refractive indices $n_f \neq n_s$.

4. To create circularly polarised radiation we can use a combination of linear polariser and a $\lambda/4$ waveplate, with the transmission axis of the linear polarisation at 45° from the fast and slow axes of the waveplate.

§65. Examination of Polarised Light

Finally we would like to find methods to investigate the state of polarisation of a light ray.

Finding α and Θ of Elliptically Polarised Light

We shall remind ourselves of how the $\lambda/4$ plate with the fast axis aligned with the semi-axis of the ellipse described by the **E** vector of an elliptically polarised light changes this elliptically polarised light into linearly polarised light, again noting that the parameters of polarisation for an elliptically polarised light are the angles α and Θ. This is shown in figure 65.1, which is a repeat of figure 63.2. Therefore to examine the nature of elliptically polarised light, we

would like to determine these two parameters. A simple experiment achieving this uses a $\lambda/4$ waveplate followed by a linear polariser as follows.

- We first rotate the $\lambda/4$ waveplate and try to change the elliptically polarised light into linearly polarised light. This can only be done if the fast axis of the waveplate is aligned with one of the semi-axes of the elliptically polarised light. To check whether we have achieved this successfully, we place a linear polariser in front of the waveplate and we simply rotate the polariser and see whether we are able to extinguish the output ray. If that is not possible, then we rotate the waveplate again.

- After we find the orientation of the waveplate such that the light is fully extinguishable by the linear polariser, we then rotate the linear polariser such that the output is maximised.

- Then, the parameter Θ is the direction of one of the fast and slow axes of the waveplate, depending on the handedness of the elliptically polarised light. The direction of the transmission axis of the linear polariser is $\Theta + \alpha$, which allows us to find the parameter α.

Apparently the first step would require many trial-and-errors for the correct orientations of the components of the apparatus. To avoid this, we may consider first passing the light through a single polariser. Then we rotate the polariser for maximum transmission, which is an alternative method of determining Θ. Then we can simply orient our $\lambda/4$ waveplate along that direction, which saves us from iteratively adjusting the waveplate and the linear polariser in behind.

Examination of Partially Polarised Light

We note that for purely elliptically polarised light, the above procedure is plausible, but for **partially polarised light**, that is, light that is formed as a mixture of elliptically polarised light and unpolarised light, it is impossible to rotate the linear polariser such that the light passing through the waveplate is totally extinguished. However through the same set of apparatus, we are not only able to find Θ and α for the elliptically polarised light, but we are also able to find the ratio of intensities of the unpolarised and polarised components of the incident light.

To do this, we shall first use a single linear polariser. We record the maximum and minimum intensities recorded by the linear polariser by rotating the linear polariser. We denote the ratio of the two intensities as κ. We shall call the transmission axis of the linear polariser when the maximum intensity is recorded a: this is the semi-major axis of the elliptically polarised light. The orientation of a from the horizontal is exactly the definition of Θ.

Then, we place a $\lambda/4$ waveplate in front of the linear polariser. We align the fast axis of the waveplate with a, such that the output from the waveplate is linearly polarised. Then we rotate the linear polariser again, where this time we denote the axis of maximum intensity as m. The angle between a and m is exactly the parameter α in figure 65.1.

The data collected from the experiment, namely κ and α, can be processed as follows. We denote the length of the semi-axes of the ellipse described by the **E** vector of the elliptically polarised component in the incident radiation as E_{0a} and E_{0b}. We then denote the intensity of the electric field of the unpolarised component of the incident light after passing through a linear polariser as E. From the previous experiments, we have got the following information

about the radiation:

$$\kappa \;=\; \frac{E^2 + {E_{0a}}^2}{E^2 + {E_{0b}}^2}; \qquad \text{(ratio of intensities)} \tag{65.1}$$

$$\frac{E_{0a}}{E_{0b}} \;=\; \tan\alpha. \qquad \text{(determination of } \alpha\text{)} \tag{65.2}$$

Noting that the unpolarised light is incoherent, which is the reason we write the intensity as $E^2 + E_0^2$ instead of $(E + E_0)^2$. Rearranging these equations will allow us to express E^2 in terms of κ, α, and one of ${E_{0a}}^2$ and ${E_{0b}}^2$. After that, we can work out the ratio of intensities of the unpolarised and elliptically polarised components χ by

$$\chi \;=\; \frac{E^2 + E^2}{{E_{0a}}^2 + {E_{0b}}^2} \;=\; \frac{2E^2}{{E_{0a}}^2 + {E_{0b}}^2}, \tag{65.3}$$

which is exactly what we are looking for.

Finally there is a note of caution. When we apply a single linear polariser to a incoming light with no change in the output intensity when we rotate the polariser, we cannot call it a day by claiming the incident light is unpolarised — it might be circularly polarised. If we then add a $\lambda/4$ waveplate before the linear polariser and still find no change in intensity when we rotate the polariser, then we can claim that the incident light is unpolarised, as any circularly polarised component will turn into either linearly polarised light or elliptically polarised light when passed through the waveplate, and therefore should be detectable as a maximum or a minimum when the linear polariser is rotated.

Summary

1. With a combination of a $\lambda/4$ waveplate with a linear polariser, we are able to determine the value of α and Θ of an incident elliptically polarised light, which is a complete set of parameters of describing the light.

2. With the same combination of apparatuses applied to an incident partially polarised light, we are able to determine the ratio of the intensities of the unpolarised and elliptically polarised light. Furthermore we are able to determine α and Θ for the elliptically polarised component for the incident light.

§66. Jones-Vector Formalism

Jones-Vector Formalism

The next topic that we shall look into may look a bit detached but is very useful, which is the Jones-vector formalism.They can describe the polarisation states of polarised light, yet unable to describe partially polarised or unpolarised light. A Jones-vector is simply a part of the electric field vector, hidden inside equation 58.6,

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} \;=\; \mathrm{Re}\begin{pmatrix} E_{0y}\mathrm{e}^{\mathrm{i}(kx-\omega t)} \\ E_{0z}\mathrm{e}^{\mathrm{i}(kx-\omega t-\delta)} \end{pmatrix} \;=\; \sqrt{{E_{0y}}^2 + {E_{0z}}^2} \times \mathrm{Re}\big[\mathrm{e}^{\mathrm{i}(kx-\omega t)}\,|\psi\rangle\big], \tag{66.1}$$

where the Jones-vector $|\psi\rangle$ is

$$|\psi\rangle \;=\; \frac{1}{\sqrt{{E_{0y}}^2 + {E_{0z}}^2}} \times \begin{pmatrix} E_{0y} \\ E_{0z}\mathrm{e}^{-\mathrm{i}\delta} \end{pmatrix}. \tag{66.2}$$

Some examples include polarised light with the y and z components equal, for example, linearly polarised light with polarisation angles 45° have Jones-vectors

$$|+\rangle \;=\; \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \quad \text{and} \quad |-\rangle \;=\; \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}, \tag{66.3}$$

respectively and left- and right-handed circularly polarised light have Jones-vectors

$$|L\rangle \;=\; \frac{1}{\sqrt{2}}\begin{pmatrix}1\\\mathrm{i}\end{pmatrix} \quad \text{and} \quad |R\rangle \;=\; \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-\mathrm{i}\end{pmatrix} \tag{66.4}$$

respectively. Another example would be linearly polarised light with two electric field components unequal, which has Jones-vectors

$$|\psi\rangle \;=\; \frac{1}{\sqrt{{E_{0y}}^2 + {E_{0z}}^2}} \times \begin{pmatrix}E_{0y}\\\pm E_{0z}\end{pmatrix} \tag{66.5}$$

which, taking one of the two electric field components to be 0 gives linearly polarised light along y- and z- directions only as

$$|H\rangle \;=\; \begin{pmatrix}1\\0\end{pmatrix} \quad \text{and} \quad |V\rangle \;=\; \begin{pmatrix}0\\1\end{pmatrix}. \tag{66.6}$$

Note that the three pairs of Jones vectors $|H\rangle$ and $|V\rangle$, $|+\rangle$ and $|-\rangle$, and $|L\rangle$ and $|R\rangle$ are orthogonal, i. e.

$$\langle +|-\rangle \;=\; 0, \quad \langle H|V\rangle \;=\; 0, \quad \langle L|R\rangle \;=\; 0, \tag{66.7}$$

suggesting that either pair would form a basis set of polarisation states of light, and is able to span all the other polarisation states, including the two other basis sets of polarisation states.

Polarisers in the Jones-Vector Formalism

Polarisers, in the Jones-vector formalism, are operators that projects light into the basis defined by the polariser. For example, polarisers with its transmission axis horizontal and vertical are given by

$$\Pi_H \;=\; |H\rangle\langle H| \;=\; \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} \quad \text{and} \quad \Pi_V \;=\; |V\rangle\langle V| \;=\; \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix}. \tag{66.8}$$

Previously we have also looked into circular polarisers, which uses a combination of a linear polariser and a $\lambda/4$ waveplate to change the state of polarisation circular. Such a construction may be useful, but if light is directed through this polariser backwards (i. e. directing light through first through the waveplate then through the linear polariser), the polariser does not behave in the desired way. A polariser that solves the previous problem, called a **homogeneous circular polariser**, are described by the following operators in the Jones-vector formalism,

$$\Pi_L \;=\; |L\rangle\langle R| \;=\; \frac{1}{2}\begin{pmatrix}1 & -\mathrm{i}\\\mathrm{i} & 1\end{pmatrix} \quad \text{and} \quad \Pi_R \;=\; |R\rangle\langle L| \;=\; \frac{1}{2}\begin{pmatrix}1 & \mathrm{i}\\-\mathrm{i} & 1\end{pmatrix}, \tag{66.9}$$

for left- and right-circular polarisers.

Summary

1. A Jones-vector is a part of the electric field vector, defined by

$$|\psi\rangle \;=\; \frac{1}{\sqrt{{E_{0y}}^2 + {E_{0z}}^2}} \times \begin{pmatrix} E_{0y} \\ E_{0z}\mathrm{e}^{-\mathrm{i}\delta} \end{pmatrix}. \tag{66.10}$$

   There are three convenient set of orthogonal basis states of polarised light, which are the pairs $|H\rangle$ and $|V\rangle$, $|+\rangle$ and $|-\rangle$, and $|L\rangle$ and $|R\rangle$.

2. Polarisers are operators in the Jones-vector formalism.