Chapter 13
Interference—Diffraction

Abstract This chapter is based on two extremely important properties of waves. One is diffraction, the property whereby a wave confined to a part of space gradually spreads out to nearby spatial regions if there are no barriers to prevent its spread; the other is interference, which arises from the fact that, in linear media, waves add to each other at amplitude level when they meet. Two waves of equal amplitude may at one extreme add so that they completely cancel each other or in the opposite extreme add constructively so that the resultant wave (with twice the amplitude) carries four times as much energy as does each individual wave. This chapter shows how these general principles can be applied when there are particular geometric constraints for wave motion, like a single slit, a double slit, a diffraction grating, a circular aperture or a spherical obstacle. Arago's spot is mentioned since it provides a historical perspective. Both analytical mathematics and numerical methods are used. Particular topics like Huygens's principle, Airy's disc, Rayleigh's resolution criterion, diffraction-limited optics and Babinet's principle are also presented.

13.1 The Nature of Waves—At Its Purest

In this chapter, we will describe some of the most *wave*-specific phenomena found in physics! These are phenomena that can be displayed for all waves in space, such as sound waves, waves in water and electromagnetic waves, including visible light. In many contexts, the experiments are simple and transparent, and the extension of the waves in time and space becomes a central and almost inevitable ingredient of any explanation model.

There are a number of phenomena that can be observed when two or more waves work together. Sometimes, the results are surprising—and often beautiful! In this chapter, we will primarily discuss interference and diffraction. Historically, we may say that the word "interference" was primarily used when two separate waves interacted, while the word "diffraction" was most commonly used when some parts of a wave interacted with other parts of the same wave. It is almost impossible to keep these two concepts apart in every situation, with the result that sometimes we are confronted with an illogical use of these words.

A. I. Vistnes, *Physics of Oscillations and Waves*, Undergraduate Texts in Physics,
https://doi.org/10.1007/978-3-319-72314-3_13

Whatever the names, diffraction and interference are, as already mentioned, some of the most wave-specific phenomena are known to us. Thomas Young's double-slit experiment is one of the most discussed topics in physics today, and interference is the main reason why one could not overlook the wave nature of light a hundred years ago when Einstein and others found support for the view that the light sometimes appears to behave like particles.

In this chapter, we will first and foremost illustrate interference and diffraction through phenomena related to light, but sometimes it is useful to resort to concrete water waves to better understand the mechanisms behind the phenomena.

When two or more waves work together, we need to know how to add or combine waves.

The basis for all interference and diffraction is the *superposition principle*:

The response to two or more concurrent stimuli s_i will at a given time and place be equal the sum of the response the system would have on each of the stimuli individually.

Superposition implies, in other words, additivity, which is expressed mathematically as:

$$F(s_1 + s_2 + \cdots + s_n) = F(s_1) + F(s_2) + \cdots + F(s_n) \ .$$

This means that F is a linear mapping. In other words, F must be a linear function!

In physics, we know that many phenomena behave approximately linearly. The most familiar examples are probably Ohm's law for resistors and Hooke's law for the extension/compression of a spring. As long as the "amplitudes" are small, an (approximately) linear relation applies. But we know that this law does not give a good description for larger "amplitudes". Then, the "higher-order terms" must be taken into account (the expression can be understood as referring to a Taylor expansion). We mention this to remind you that the superposition principle does not apply in every situation. In this chapter, however, we still limit almost exclusively to linear systems where superposition applies.

In this chapter, phenomena will sometimes be presented qualitatively, usually with a simple formula. In addition, we will provide a more "complete" mathematical description of three basic situations:

- Interference from a double slit,
- Interference from a grating (many parallel slits),
- Diffraction from a single slit.

The mathematical details of the actual derivation are of limited value (especially for gratings and single slit), but the main idea that lies at their root is of paramount importance, so be sure to get a firm hold on it!

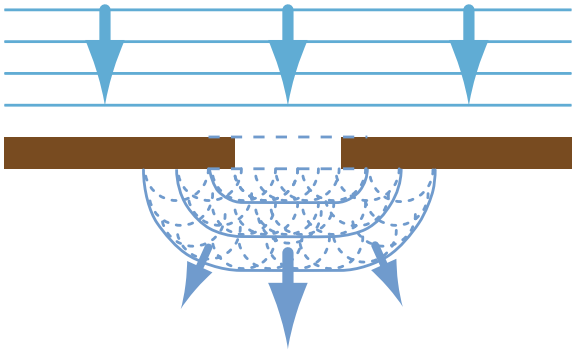

Fig. 13.1 According to Huygens's principle, we may think of any point on a wavefront as the source of elementary waves

13.2 Huygens's Principle

Our description of interference and diffraction is based on Huygens's principle, which states that:

> *Any point in a wave can be viewed as a source of a new wave, called the elementary wave, which expands in all directions. For following a wave motion, we can start from, for example, a wavefront and construct all conceivable elementary waves. If we go one wavelength along these elementary waves, their envelope curve will describe the next wavefront (Fig. 3.1).*

Fresnel modified the above view by saying that if we are to find the wave amplitude somewhere in space (also well away from an original wavefront), we can sum up all conceivable waves provided that we take into account both amplitude and phase (and whether or not something obstructs the wave).

The Dutchman Christiaan Huygens[1] lived from 1629 to 1695 and the Frenchman August-Jean Fresnel from 1788 to 1827, and we might wonder if such an old viewpoint has any relevance today, when we have Maxwell's equations, relativity theory and quantum physics. Remarkably enough, the Huygens–Fresnel principle is still applicable, and it is in a way a leading principle in quantum electrodynamics (QED), the most accurate theory available today. True enough, we do not use the vocabulary of Huygens and Fresnel for describing what we do in QED, but mathematically speaking the main idea is quite equivalent. In quantum electrodynamics, it is said that we must follow all possible ways that a wave can go from a source to the place where the wave (or probability density) is to be evaluated. If a particle description is used, the phase information lies at the bottom even in the quantum field. In other words, the Huygens–Fresnel principle is hard-wearing (Fig. 13.1).

[1] Unusual pronunciation, see Wikipedia.

Throughout the chapter, we assume that the light is "sufficiently coherent". We will return to coherence in Chap. 15, and here we will only state that the light we start with (e.g. emerging from a slit) can be described mathematically as an almost perfect sinusoidal wave without any changes in amplitude or frequency as time flows. In other words, we assume complete predictability in the phase of the Huygens–Fresnel elementary waves in relation to the phase of the waves we start with.

13.3 Interference: Double-Slit Pattern

In 1801, when the Englishman Thomas Young (1773–1829) conducted his famous double-slit experiment, Newton's corpuscular (particle) model for light was the motivation. The corpuscular model seemed appropriate in so far as it accounted for the fact that light rays travel in straight lines and for the observed laws of reflection. And Newton's red, green and blue corpuscles provided an excellent starting point for explaining additive colour mixing.

If Newton's light particles pass through two narrow parallel slits, we would expect to see two strips on a screen placed behind a double slit. But what did Young observe? He saw *several* parallel strips! These strips are called interference fringes. This fringe pattern was almost impossible to explain on the basis of Newton's particle model. Young, and subsequently Fresnel and others, could easily explain this phenomenon, and we shall presently look at the mathematics (Fig. 13.2).

The two slits are assumed to be narrow (often 1–1000 times the wavelength), but "infinitely" long so that we can look at the whole problem as two-dimensional (in a plane perpendicular to the screens and slits).

We assume that light enters with a wavefront parallel to the slits so that the light starts with identical phase throughout the "exit plane" in both slits. We assume further

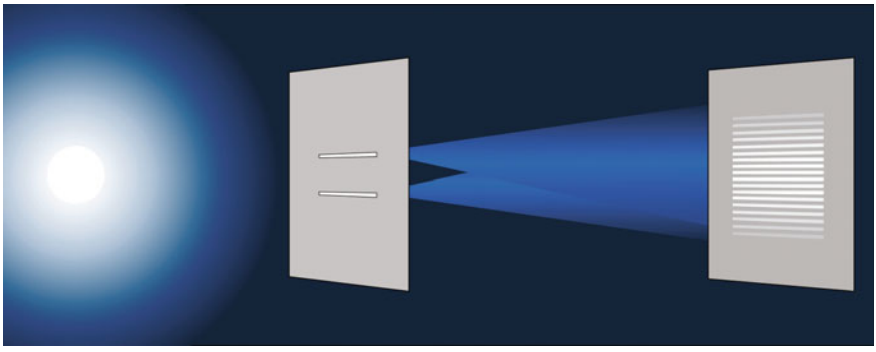

Fig. 13.2 Experimental set-up for Young's double-slit experiment. Slit sizes and strip patterns are greatly exaggerated compared to the distance between light source, slits and screen

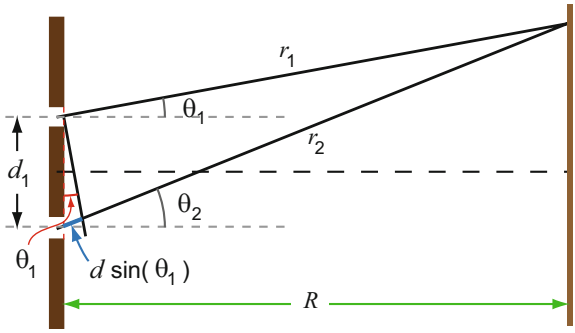

Fig. 13.3 Schematic light path from the double slit to a given point on the screen at the back. In reality, the distance R from slits to the screen is much greater than the gap d_1 between the slits. See text for details

that each of the slits emits elementary waves, and for reasons just mentioned, these waves will have a wavefront that is shaped as part of a cylindrical surface with the slit as the cylinder axis. In a plane perpendicular to the columns, we will then get a purely two-dimensional description (see Fig. 13.3).

We are dealing with light, that is, with an electromagnetic wave. The wave is transverse and is described by an electric and a magnetic field, each of which has a certain direction in space. We assume that we are considering the interference phenomenon so far away from the slits that we can ignore the *difference in the direction* in space for electrical fields originating from slit 1 compared to the field originating from slit 2. It will be sufficient for us therefore to add the two electrical fields as scalar quantities with the correct intensity and phase.

We want to find the electric field on a screen parallel to the plate with the slits, in a direction θ relative to the normal vector between the slits (see Fig. 13.2). The contributions from the two slits are then:

$$E_1(\theta_1) = E_{1,0}(r_1, \theta_1) \cos(kr_1 - \omega t - \phi)$$

$$E_2(\theta_2) = E_{2,0}(r_2, \theta_2) \cos(kr_2 - \omega t - \phi)$$

where ϕ is an arbitrary phase angle when space and time are given. Since the screen with the slits and the screen where we capture the image are very far apart compared to the gap between the slits, the angles θ_1 and θ_2 will be almost identical, and we replace them both with θ:

$$\theta_1 \approx \theta_2 = \theta \ .$$

For the same reason, we will assume that the two amplitudes are identical, and write:

$$E_{1,0}(r_1, \theta_1) = E_{2,0}(r_2, \theta_2) = E_0(r, \theta) \ .$$

The total amplitude in the direction θ is therefore (according to the superposition principle):

$$E_{tot}(\theta) = E_0(r,\theta)[\cos(kr_1 - \omega t - \phi) + \cos(kr_2 - \omega t - \phi)] \,.$$

Using the trigonometric identity

$$\cos a + \cos b = 2\cos\left(\frac{a+b}{2}\right)\cos\left(\frac{a-b}{2}\right)$$

and get:

$$E_{\text{tot}}(\theta) = 2E_0(r,\theta)\ \cos\left(k\frac{r_1+r_2}{2} - \omega t - \phi\right)\cos\left(k\frac{r_1-r_2}{2}\right) \,.$$

Superposition always operates on amplitudes (i.e. to say, a real physical quantity, not an abstract quantity such as energy or intensity). Be that as it may, physical measurements are often based on intensity. When we view light on a screen with our eyes, the light intensity we sense is proportional to the intensity of the wave.

The intensity of a plane electromagnetic wave in the far-field zone is given by the Poynting vector, but the scalar value is given by:

$$I = cED = c\varepsilon E^2$$

where c is the velocity of light, E the electric field, D the electric flux density (electric displacement), and ε the electric permittivity. Hence:

$$I(\theta,t) = c\varepsilon E_{\text{tot}}^2(\theta,t) = 4c\varepsilon E_0^2(r,\theta)\ \cos^2\left(k\frac{r_1+r_2}{2} - \omega t - \phi\right)\cos^2\left(k\frac{r_1-r_2}{2}\right) \,.$$

This is the so-called instantaneous intensity that varies over time within a period. We are most interested in time-averaged intensity. The first $\cos^2$ term varies with time, and the time average of $\cos^2$ is 1/2. Accordingly:

$$I(\theta) = 2c\varepsilon E_0^2(r,\theta)\cos^2\left(k\frac{r_1-r_2}{2}\right) \,.$$

We define

$$r_1 - r_2 = \Delta r = d\sin\theta$$

where d is the distance between the slits. Furthermore, we bring in the wavelength through the relationship $k = 2\pi/\lambda$.

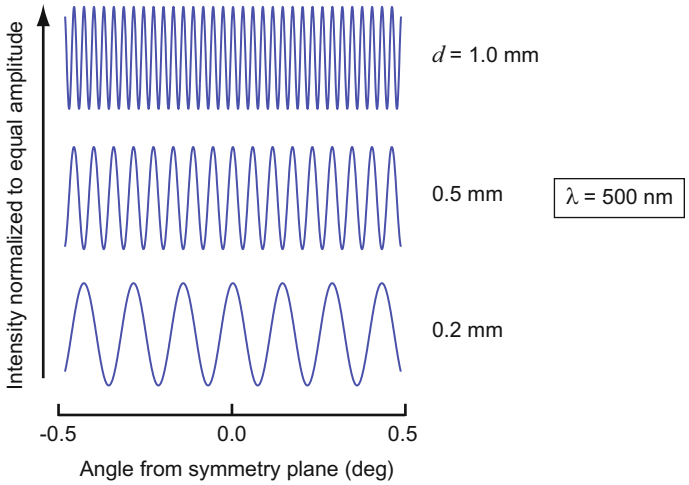

Fig. 13.4 Strip pattern on a screen behind the double slit. The distance between the slits is indicated

Whence follow the intensity distribution of the light that has passed a double slit (Fig. 13.4):

$$\overline{I}(\theta) = 2c\varepsilon E_0^2(r,\theta)\cos^2\left(\frac{d\sin\theta}{\lambda}\pi\right) . \tag{13.1}$$

When $\theta = 0$, we get maximal intensity. The minima are obtained when the argument of the cosine function is an odd multiple of $\pi/2$:

$$\frac{d\sin\theta}{\lambda}\pi = (2n+1)\frac{\pi}{2}, \qquad (n = 0, 1, 2, \ldots) .$$

The condition for a minimum is thus found to be:

$$\sin\theta = \frac{\lambda}{d}\left(n + \frac{1}{2}\right) .$$

The maxima occur when:

$$\sin\theta = \frac{n\lambda}{d} \quad \text{approximately} .$$

The word "approximately" has been added because the exact expression for the maxima depends also on how $E_0^2(r,\theta)$ varies with θ.

We should notice that usually, at least for light, the gap between the slits is large relative to the wavelength. That is, the angle between two minima (or between two

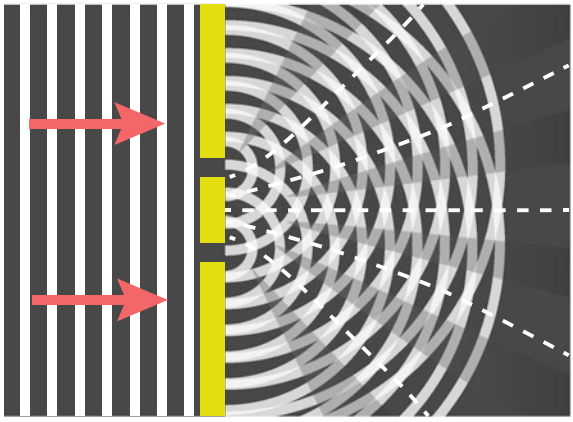

Fig. 13.5 Direction of the interference lines can be demonstrated by placing two sets of concentric circles, with the centre of each set in the middle of a slit

maxima) is usually quite small. This means that we can in principle get an interference pattern consisting of very many parallel bright strips on the screen with dark spaces in between. Thus, we will not get just *two* strips, as a particle model of light would predict.

How many strips do we really get? Well, it depends on $E_0^2(r, \theta)$. If we use Huygens's principle and only use one elementary wave, it should have the same intensity in all directions (where the wave can expand). But the gap (between the slits) cannot be infinitesimally narrow, for in that case virtually no light would pass through. When the slit has a finite width, we should actually let elementary waves start at any point in the slit. These elementary waves will set up a total wave for slit 1 and a total wave for slit 2, which will *not* have the same electric field in all directions θ. We will address this problem below (diffraction from one slit).

Since $E_0^2(r, \theta)$ will be large only for a relatively narrow angular range, we get a limited number of fringes on the screen when we collect the light from the double slit. This will be treated later in this chapter.

In Fig. 13.5, we finally show a fairly common way of illustrating interference by a double slit. With the centre in each of the two slits (and in a plane normal to and in the middle of the slits), wavefronts are drawn, characterized by the property that electric fields is, for example, maximum in a direction normal to the plane under consideration. At all places where the crest (top, peak) of a wave from one slit overlaps the crest of a wave from the other slit, there will be a constructive interference and we will get maximum electric field. These are places where the circles cross each other.

The places where a wave crest from a slit overlaps a wave trough (valley, minimum) from the second slit (i.e. in the middle of two circles from this slit), there will be a destructive interference and we will have almost a negligible electric field.

We can see from the figure that positions with constructive interference lie along lines that radiate approximately midway between the two slits. It is in these directions

that we get the bright strips in the interference pattern from a double slit. In the midst of these, there is destructive interference and little or no light.

It is instructive to demonstrate how the angles between the directions of constructive interference change as we vary the distance between the centres in the circular patterns.

13.3.1 Interference Filters, Interference from a Thin Film

We have previously seen that when we send light towards a flat interface between air and glass, about 5% of the light is reflected at the surface (even more at larger angles of incidence). Such reflection deteriorates the contrast and image quality in general if lenses in a binocular or a camera are not given an anti-reflection treatment. But how can we deposit such a coating on a lens?

Figure 13.6 shows schematically how we can proceed. We put a thin layer of some transparent substance on the outside of the glass and choose a material that has a refractive index about halfway between the refractive indices of air and glass. We will then reflect about as much light from the air–coating interface as from coating–glass interface. If we ignore yet another reflection (in the return beam), we see that light reflected from the upper and lower layers will have the same direction when they return to the air. The two "rays" will superpose. If the two have opposite phase, they will extinguish each other. This means that the light *actually* reflected will (on the whole) be significantly less intense than if the coating was not present.

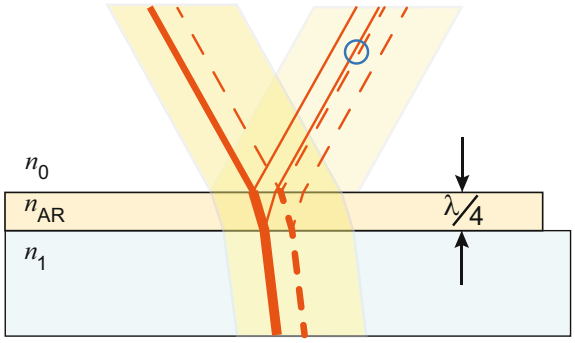

Fig. 13.6 An anti-reflection treatment of a lens or spectacle consists of a thin transparent layer with a refractive index roughly halfway between the refractive indices of air and glass. The layer must be about a quarter wavelength thick for the wavelengths where the filter has to give the best performance. A beam of light that is slightly inclined towards the surface is drawn to produce the sum of a part of the wave reflected on the surface of the anti-reflection layer (dashed) and a part of the wave reflected from the surface of the glass itself (solid line). The overlap between these is marked with a circle

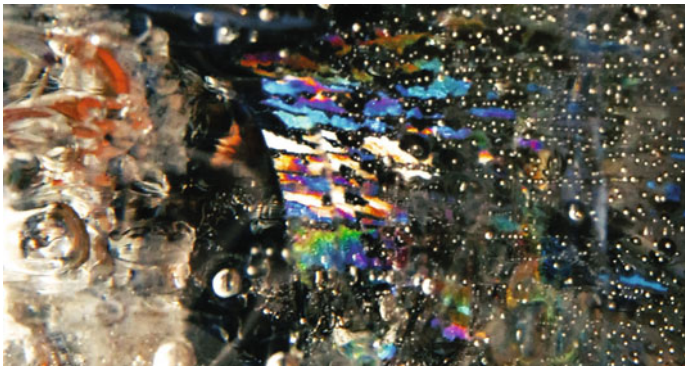

Fig. 13.7 Play of colours in a cracked ice chunk

By carefully selecting all parameters, we can determine whether there will be destructive or constructive interference. In the first case, we get an anti-reflective layer as already shown. In the second case, we get more reflection. In this case, a coating consisting of several layers on top of each other is often used and the parameters are chosen so that light reflected everywhere comes in phase with other reflections and that the light transmitted from different layers is always out of phase with other transmission contributions. In this way, it is possible to make mirrors that can have more than 99.9% reflection for a particular wavelength and for a particular direction of a beam of light towards the mirror, while at other wavelengths we can look across the mirror! It is quite nice to experience such mirrors!

In nature and everyday life, thin films form spontaneously, for example in thin cracks or thin air layers between two glass plates. For example, if we put a "watch glass" (slightly curved glass for covering the dial of a pocket watch) on top of a flat glass surface, we get constructive and destructive interference between light reflected at the interfaces between air and the curved and flat glass surfaces. Since the effect is wavelength dependent, the circles are coloured and they are called Newton's rings.

In Fig. 13.7, one finds another example of the same effect. There is a chunk of ice in which a slight crack has occurred after a blow against the piece, and the play of colours is evident.

13.4 Many Parallel Slits (Grating)

If we have many parallel slits with the same mutual distance d, and if we collect the light on a screen far from the slits (compared to d), we get a situation that can be analysed in much the same way as the double slit. The difference is that we must sum up contributions from all N slits (see Fig. 13.8).

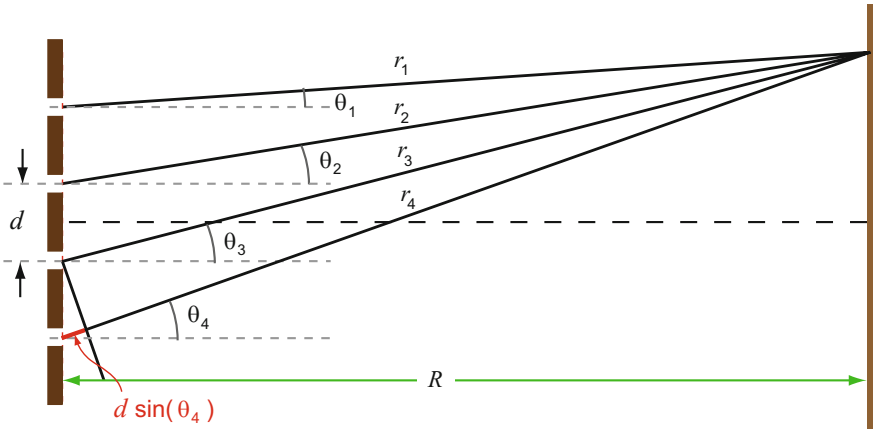

Fig. 13.8 Starting point for the mathematics for many slits is likewise that for the double slit. In practice, we often have several hundred slits per mm, illuminated by a light beam with a diameter of the order 1 mm. The screen is often 1 m away or more. Thus, all r_n are nearly equal and likewise for the θ_n. This simplifies the calculations

The resultant field will be:

$$E_{\text{tot}}(\theta) = E_1 + E_2 + \cdots + E_N$$

$$= E_0(r,\theta)\big[\cos(kr_1 - \omega t - \psi) + \cos(kr_2 - \omega t - \psi) + \cdots + \cos(kr_N - \omega t - \psi)\big] .$$

In order to simplify the calculations further, we note that the absolute phase ψ relative to the selected position and time is uninteresting. When we only look at the time-averaged intensity, only phase differences due to different path lengths of the various elementary waves will count. For a given angle θ, the difference between two adjacent elementary waves will be given by $d \sin\theta$. This path difference represents a phase difference ϕ, and we have already shown above that this phase difference is given by $\phi = 2\pi d \sin\theta/\lambda$.

Notice:
The following page (slightly more) is a pure mathematical treatment and adds nothing to the physics. You may skip this and jump directly to the next figure and/or the grey marked text starting with "The intensity pattern …".

Notice that if we start from one slit, the phase difference to the next will be ϕ, then 2ϕ, the next 3ϕ, etc. Then, we can write the resultant field in this simplified way:

$$E_{\text{tot}}(\theta) = E_0(r,\theta)\big\{\cos\omega t + \cos(\omega t + \phi) + \cos(\omega t + 2\phi) + \cdots + \cos[\omega t + (N-1)\phi]\big\} ,$$

$$E_{\text{tot}}(\theta) = E_0(r,\theta) \sum_{n=0}^{N-1} \cos(\omega t + n\phi) \ .$$

We use now Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$ and the symbol $\Re$ as before for taking the real part of a complex expression and get:

$$\sum_{n=0}^{N-1} \cos(\omega t + n\phi) = \Re \sum_{n=0}^{N-1} \mathrm{e}^{\mathrm{i}(\omega t + n\phi)} = \Re \left(\mathrm{e}^{\mathrm{i}\omega t} \sum_{n=0}^{N-1} \mathrm{e}^{\mathrm{i}n\phi} \right) \ .$$

From the mathematics, we know that the sum of a geometric series with common ratio k can be written as:

$$1 + k + k^2 + \cdots + k^{N-1} = \sum_{n=0}^{N-1} k^n = \frac{k^N - 1}{k - 1} \ .$$

Applying this relation to the sum $\sum_{n=0}^{N-1} e^{in\phi}$ (k standing for $e^{i\phi}$), we get:

$$\sum_{n=0}^{N-1} \cos(\omega t + n\phi) = \Re \left(\mathrm{e}^{\mathrm{i}\omega t} \sum_{n=0}^{N-1} \mathrm{e}^{\mathrm{i}n\phi} \right) = \Re \left(\mathrm{e}^{\mathrm{i}\omega t} \frac{\mathrm{e}^{\mathrm{i}N\phi} - 1}{\mathrm{e}^{\mathrm{i}\phi} - 1} \right)$$

$$= \Re \left(\mathrm{e}^{\mathrm{i}\omega t} \ \frac{\mathrm{e}^{\mathrm{i}N\phi/2}}{\mathrm{e}^{\mathrm{i}\phi/2}} \ \frac{\mathrm{e}^{\mathrm{i}N\phi/2} - \mathrm{e}^{-\mathrm{i}N\phi/2}}{\mathrm{e}^{\mathrm{i}\phi/2} - \mathrm{e}^{-\mathrm{i}\phi/2}} \right)$$

$$= \Re \left(\mathrm{e}^{\mathrm{i}\omega t} \ \mathrm{e}^{\mathrm{i}N\phi/2 - \mathrm{i}\phi/2} \ \frac{2\mathrm{i}\sin\frac{N\phi}{2}}{2\mathrm{i}\sin\frac{\phi}{2}} \right)$$

$$= \Re \left(\mathrm{e}^{\mathrm{i}(\omega t + N\phi/2 - \phi/2)} \ \frac{\sin\frac{N\phi}{2}}{\sin\frac{\phi}{2}} \right)$$

$$= \cos(\omega t + N\phi/2 - \phi/2) \ \frac{\sin\frac{N\phi}{2}}{\sin\frac{\phi}{2}} \ .$$

Combining this with earlier expressions, the electric field in the direction of θ will be:

$$E_{\text{tot}}(\theta) = E_0(r,\theta) \cos\left(\omega t + \frac{N\phi}{2} - \frac{\phi}{2} \right) \ \frac{\sin\frac{N\phi}{2}}{\sin\frac{\phi}{2}} \ .$$

In the same way as for the double slit, we are interested in the intensity of the interference pattern we can observe. Again we have:

$$I(\theta, t) = c\varepsilon E_{\text{tot}}^2(\theta, t) \ .$$

When the time average is calculated, $\overline{\cos^2(\omega t + \frac{N\phi}{2} - \frac{\phi}{2})} = \frac{1}{2}$ as before. Accordingly:

$$I(\theta) = \frac{1}{2} c\varepsilon E_0^2(r, \theta) \left[\frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \right]^2$$

$$I(\theta) = \frac{1}{2} c\varepsilon E_0^2(r, \theta) \left[\frac{\sin \frac{N\phi}{2}}{\sin \frac{\phi}{2}} \right]^2 \ . \tag{13.2}$$

The intensity pattern is then described by:

$$I(\theta) = I_0(r, \theta) \left[\frac{\sin \dfrac{N\phi}{2}}{\sin \dfrac{\phi}{2}} \right]^2 \ . \tag{13.3}$$

where $I_0(r, \theta)$ is the intensity contribution to the light passing *one* of the N slits and $\phi = 2\pi d \sin\theta/\lambda$ is the phase difference between two neighbour slits for the actual θ.

We can show (using L'Hôpital's rule) that when ϕ goes to zero, the expression inside the square brackets goes to N. That is, the intensity of the strip found at $\phi = 0$ becomes N^2 times the intensity we had from one slit only. The other maxima we find for $\sin \frac{\phi}{2} = 0$ (assuming we ignore the angular dependence of $E_0^2(r, \theta)$). It follows that maxima will occur when:

$$\sin(\pi d \sin\theta/\lambda) = 0$$

or, equivalently, when:

$$m\pi = \pi d \sin\theta/\lambda, \quad (m = \ldots, -2, -1, 0, 1, 2, \ldots)$$

$$\sin\theta = \frac{m\lambda}{d} \ . \tag{13.4}$$

These are the same directions as for interference maxima for a double slit.

We see that the positions of the intensity maxima are independent of the N, the number of slits.

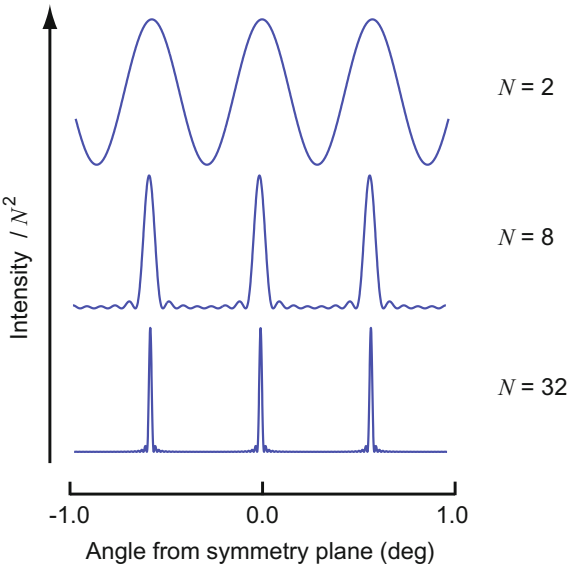

Fig. 13.9 Intensity distribution versus angle for 2, 8 and 32 slits

Figure 13.9 shows how the intensity distribution is for slightly different numbers of slits. We see that the most distinctive feature is that the peaks are becoming more pronounced when the number of slits increases.

Difference among the various interference lines

Equation (13.4) can be rewritten like this: $d \sin\theta = m\lambda$.

Thus, light from the same wavefront contribute equally with identical phase to the centre line ($m = 0$) of the diffraction pattern.

However, there is one wavelength difference between light passing one slit and light passing a neighbour slit, for light-contributions to the first line on each side of the central one ($m = \pm 1$) in the diffraction pattern. Thus, if 100 slits are illuminated, it will be 99 wavelengths difference between the contribution from slit 1 and slit 100.

For the second line on each side of the central one ($m = \pm 2$), there will be two wavelengths difference between contributions from one slit and the neighbour slit.

Thus, in order to have a result in agreement with our derivation in practice, it requires a very regular wave. We will discuss this a bit more when we talk about temporal coherence in Chap. 15.

It can be shown that the half-width of the peaks are given by:

$$\Delta\theta_{1/2} = \frac{1}{N\sqrt{\left(\frac{d}{\lambda}\right)^2 - m^2}} \tag{13.5}$$

where m gives, as before, the order. We see that the central line $m = 0$ has the smallest line width and that the line width increases when we examine lines further and further from the centre (question: Can $(d/\lambda)^2 - m^2$ be negative?).

In Fig. 13.10, we have drawn the same curves as in Fig. 13.9, but now with logarithmic y-axis. The purpose is to show details of the small peaks between the main peaks. We see that the small peak nearest to a main peak is about three log units (close to a factor 1000) less than the main peak. There are no dramatic deviations from this rule even though we change the number of slits significantly. However, we see that the width of each principal peak decreases with the number of slits, even if we include a few small peaks on each side of the principal peak. Furthermore, the logarithmic plot shows that the intensity of the main peaks relative to the minor peak approximately midway increases dramatically with the number of slits.

13.4.1 Examples of Interference from a Grating

We looked in Eq. (13.4) that the angle between the fringes in the interference pattern from a grating depends on the relationship between the wavelength and the slit separation in the grating. It is a blessedly simple relationship. An angle is easy to

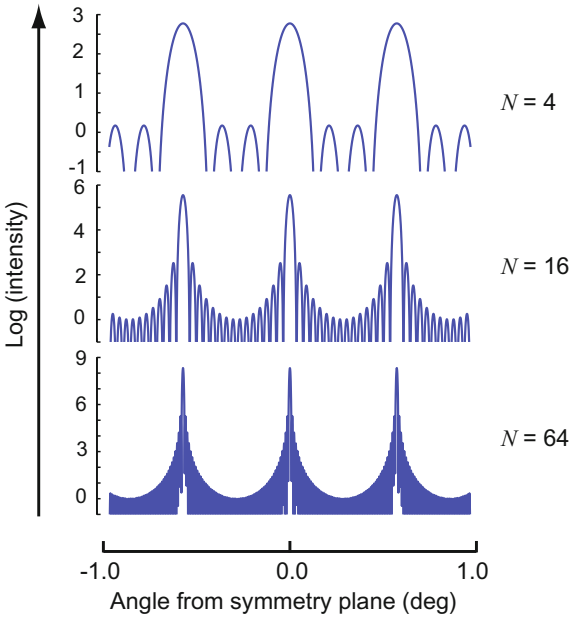

Fig. 13.10 Intensity distribution versus angle of 4, 16 and 64 slits but now drawn in logarithmic scale along the y-axis to study details near the zero line between the main peaks

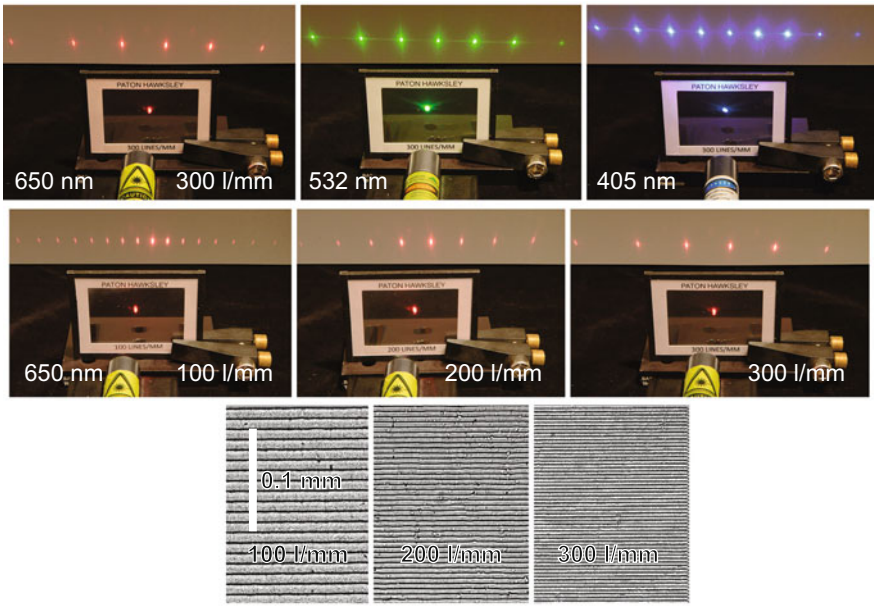

Fig. 13.11 Experimental images showing how the distance between lines (dots) in an interference pattern varies with wavelength and distance between the slits in the gratings (indicated as the number of lines per mm in the grating)

measure, and the distance between the slits in a grating is quite easy to measure (but the slits are tightly spaced, and we need a good microscope). And then, the wavelength of light is the last parameter.

In Fig. 13.11, we show examples of how the fringes look like (almost like dots, since we have used lasers with a rather narrow beam). The pictures in the top row show how the distance between the dots changes when the wavelength of the light changes. The gap between the slits in the grating is always the same (the "grating constant" is 300 lines per mm). We notice that red light gives the *largest* angles. This is in a way the opposite of what we saw in dispersion. Red light, on being dispersed by a glass prism, suffered the *smallest* deviation.

In the middle row of pictures in Fig. 13.11, we use red light all the time, but changed the distance between the slits in the grating. We see that the distance between the dots on the screen increases when the gap between the slits becomes smaller (when the number of lines per mm increases), completely in accord with Eq. (13.4). The bottom row shows photographs, taken through a microscope, of the three gratings used.

Experiments like these show that, in one way or another, wavelength must be a central part in the description of light and that wavelength must have a link with real distances in space since there are only distances in space in the gratings that vary in the experiments when we switch from one grating to another.

13.5 Diffraction from One Slit

Suppose that we now have a single slit illuminated from one side with plane polarized waves with wavefront parallel to the "surface" of the slit. We can *model* the slit as a grating where the slits lie so close and are so wide that they completely overlap one another. If the single slit has an opening of width a, we can imagine that it consists of N parallel sub-slits with a centre-to-centre distance (from the centre of a sub-slit to the centre of the neighbouring sub-slit) $d = a/N$.

There are two different methods for calculating the light intensity on a screen after the slit. The simplest method is based on an approach where the screen is thought to be very far away from the slit, compared to both the width of the slit and the wavelength. This case is called Fraunhofer diffraction and is characterized by the supposition that the amplitude of the electric field from each of the sub-slits is approximately identical on the screen and that the angle from a sub-slit to a given position on the screen is approximately equal to the angle from another subdivision to the same position.

If the distance between the slit and the screen is not very large relative to the slit width and/or wavelength, we must use more accurate expressions for contributions to the amplitude and angles. This case is called Fresnel diffraction and is more difficult to handle than Fraunhofer diffraction. The difficulties can be surmounted by using numerical methods, and we will return to this topic later in the chapter.

Let us now go back to the simple Fraunhofer diffraction, where we consider a slit composed of N narrow parallel slots that lie edge to edge. We can now use a similar expression as for the grating, Eqs. (13.2) and (13.3), if we replace d with a/N and correct for amplitudes at the different slits. In the expression of the phase difference ϕ, we now get the following relation:

$$\phi = 2\pi \frac{d \sin\theta}{\lambda} = 2\pi \frac{a \sin\theta}{N\lambda} = \frac{2\alpha}{N}$$

where

$$\alpha = \pi \frac{a \sin\theta}{\lambda} \,. \tag{13.6}$$

For the interference pattern for N equal slits in Eq. (13.3), we found that the intensity peaks were N^2 times the intensity evolving from each slit. Since intensities are proportional to electric field amplitudes squared, this corresponds to an efficient amplitude of the sum signal that is N times the amplitude due to each slit alone. This is the case since the contributions of light to a peak in the interference pattern are exactly in phase with each other, whichever slit it went through (assuming high coherence).

For the diffraction from one slit, *the contributions to the diffraction pattern from all fictitious sub-slits will never add with the same phase*. We therefore have to assign

an effective amplitude E_{ss} for each sub-slit like

$$E_{ss} = E_{tot}/N \tag{13.7}$$

where E_{tot} is an "effective amplitude" attributed to the entire slit. This is by no means a strict mathematical description, but it provides a qualitative explanation why we need to use the $1/N$ factor for the amplitudes when we divide the slit into N sub-slits.

By applying the variables of Eqs. (13.6) and (13.7) in the expression for the total intensity distribution, according to Eq. (13.2), the result is:

$$I(r,\theta) = \frac{1}{2}c\varepsilon E_{ss}^2(r,\theta)\left[\frac{\sin\frac{N\phi}{2}}{\sin\frac{\phi}{2}}\right]^2 = \frac{1}{2}c\varepsilon\frac{E_{tot}^2(r,\theta)}{N^2}\left[\frac{\sin\alpha}{\sin\frac{\alpha}{N}}\right]^2 .$$

When N is chosen to be very large, the angle α/N will be so small that $\sin\frac{\alpha}{N} \approx \frac{\alpha}{N}$. The intensity distribution can then be written as:

$$I(r,\theta) = \frac{1}{2}c\varepsilon\frac{E_{tot}^2(r,\theta)}{N^2}\left[\frac{\sin\alpha}{\frac{\alpha}{N}}\right]^2 = \frac{1}{2}c\varepsilon E_{tot}^2(r,\theta)\left[\frac{\sin\alpha}{\alpha}\right]^2 .$$

When $\theta \to 0$, α also goes to zero, and $\sin\alpha/\alpha$ approaches unity, so that:

$$I(r,0) = \frac{1}{2}c\varepsilon E_{tot}^2(r,\theta) \equiv I_0(r) .$$

The intensity distribution in the single-slit diffraction pattern thus takes the form (see Fig. 13.12):

$$I(r,\theta) = I_0(r)\left[\frac{\sin\alpha}{\alpha}\right]^2 \tag{13.8}$$

where

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

The intensity vanishes when

$$\alpha = n\pi$$

where n is a positive or negative integer. This happens when

$$\sin\theta = n\frac{\lambda}{a} . \tag{13.9}$$

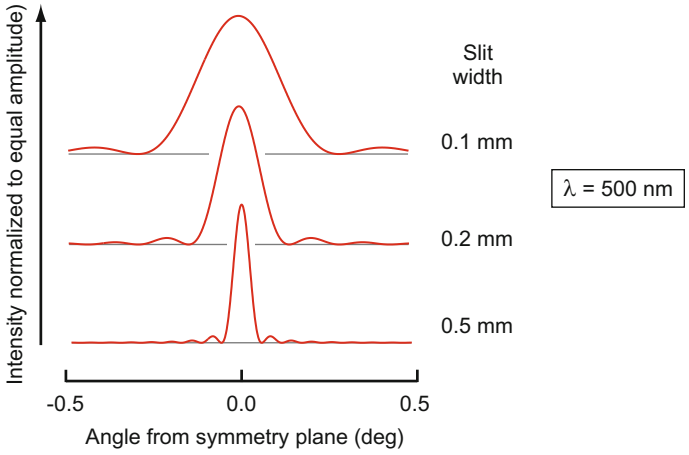

Fig. 13.12 Intensity distribution for strips after a single slit

It can be shown that the maxima lie approximately midway between the angles where the intensity is zero.

At first sight, Eq. (13.8) might look quite similar to the intensity distribution from a diffraction grating. However, it is a considerable difference.

The angle between the central top and the first minimum for a grating is given by:

$$\phi = 2\pi \frac{d \sin\theta}{\lambda} = \frac{2\pi}{N} ,$$

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

Thus, the width of the central top is becoming narrower as the number of slits increases, and it does not depend on this approximation of the width of each slit.

For the single slit, however, we thought that the slit was divided into N close-lying sub-slits, the separation between which is a/N. However, the width does not depend on this fictitious division of sub-slits. The width depends on the width of the single slit only.

The angle that gives the first zero of the intensity distribution for a single slit can be determined through a different and simple argument. Figure 13.13 shows how we can think that a pair of fictitious sub-slits a distance $a/2$ apart from each other work together to get destructive interference for *all* the light passing through the slit.

We also see from the figure that the minimum for diffraction from one slit must always exist at a larger angle than that for diffraction from two or more separate slits (since the distance d between the slits must necessarily be greater than or equal to the slit width in a grating). In other words, the angular distance of the first minimum of

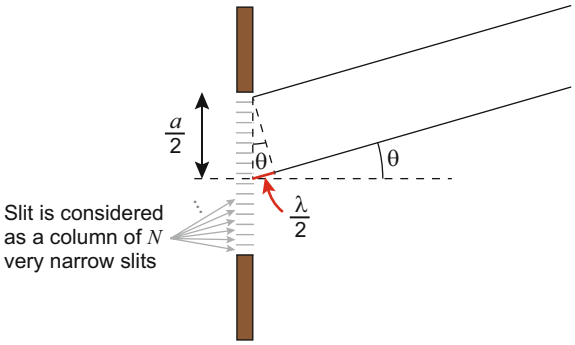

Fig. 13.13 Geometric conditions showing the direction for which the intensity of diffraction from a single slit will be zero. For any choice of a pair of fictitious sub-slits a distance $a/2$ apart, the difference in the light path will be equal to half a wavelength (which yields destructive interference)

a grating can easily be much less than for the angular distance to the first minimum in the diffraction pattern.

We can calculate the half-value for the intensity distribution from the single-slit pattern using Eq. (13.5) for a grating, but again replace the slit separation d with our fictitious gap a/N. Thus, we get:

$$\Delta\theta_{1/2} = \frac{1}{N\sqrt{\left(\frac{a}{N\lambda}\right)^2 - m^2}} = \frac{1}{\sqrt{\left(\frac{a}{\lambda}\right)^2 - (Nm)^2}} .$$

The half-width for the central peak in the single-slit pattern comes out to be ($m = 0$):

$$\Delta\theta_{1/2} = \frac{\lambda}{a} .$$

We find, of course, that the expression does not depend on N.

A typical intensity distribution in the single-slit pattern looks approximately as shown in Figs. 13.12 and 13.14. There is a distinctive central peak with weak stripes on the side. It can easily be shown that we do not get more marked peaks than the central top (since the denominator never gets zero except for the central top).

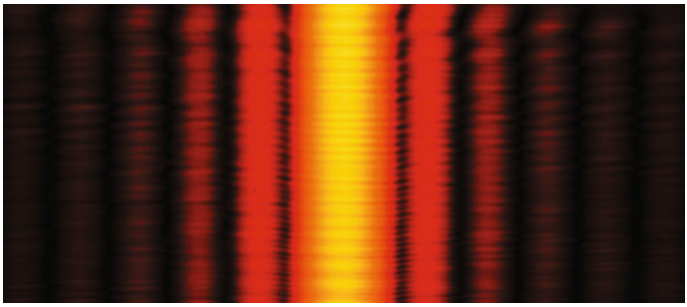

Fig. 13.14 An example of the observed intensity distribution in a single-slit diffraction pattern. The central band is overexposed to make the side bands come out well

13.6 Combined Effect

In the development of the expression of the intensity distribution from a single slit, we did not pay particular attention to the fact that the strength of the electric field will vary with the angle θ. In treating the double slit and grating, we placed more emphasis on this. The reason is that it is actually the underlying diffraction from each slit that forms the envelope for $E_0^2(r, \theta)$! We do not get the clearest fringe pattern from a double slit or from a grating to extend beyond the central peak of the diffraction image from each single slit.

In practice, therefore, we will always have a combined effect of diffraction from a single slit and interference from two or more simultaneous slits. Figure 13.15 shows the combined effect of diffraction from each of the two parallel slits and interference due to the fact that we have two slits. The example is chosen to match an optimal double-slit experiment where there are a significant number of clearly visible fringes within the central diffraction peak.

13.7 Physical Mechanisms Behind Diffraction

So far, we have used the Huygens–Fresnel principle to calculate mathematically what intensity distributions we get from interference and diffraction. But what are the physical mechanisms behind diffraction? There are several different descriptions; among other things, it is popular to invoke Heisenberg's uncertainty relationship for this purpose. It is an "explanation" that does not really go back to physical

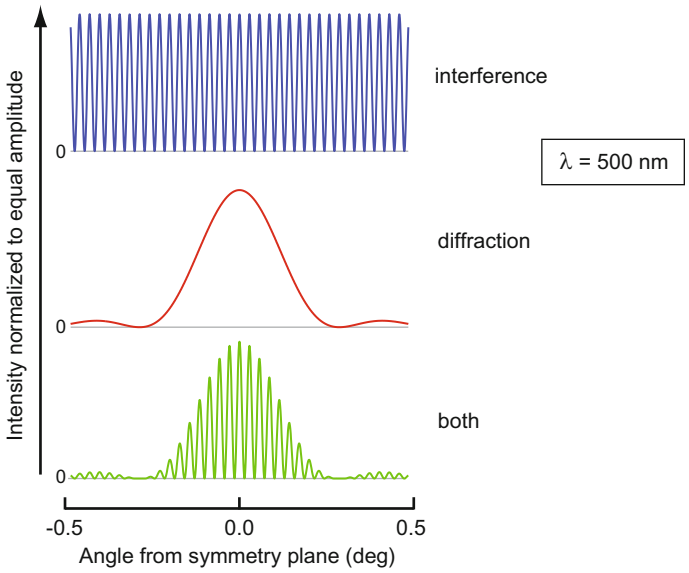

Fig. 13.15 Calculated intensity distribution for the fringe pattern from a double slit when each slit is 200 wavelengths wide and the gap between the slit centres is 2000 wavelengths

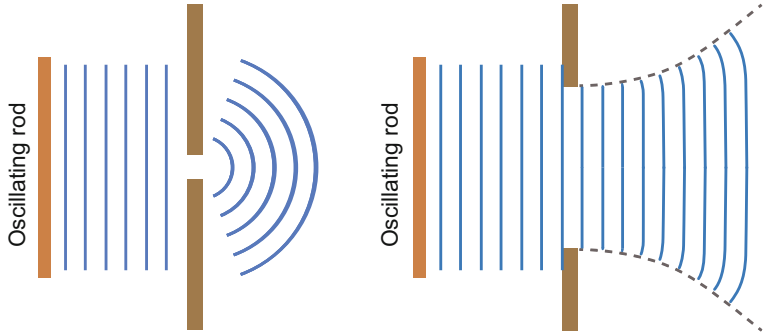

Fig. 13.16 Diffraction may appear for surface waves on water. Waves can be generated by a rod oscillating up and down the water surface. When waves are sent to a wall with a gap, the waves behind the wall get a form dictated by diffraction

mechanisms, and it is just a mathematical game. We will try to figure out more physical mechanisms for the phenomenon.

We choose to look at plane waves created by letting an oscillating rod go in and out of a water surface (see Fig. 13.16). The waves so generated move towards a vertical wall with a vertical opening (slit). The waves are parallel to the wall. The waves are only (approximately) plane over a limited length, but are at least so long that they cover the entire opening in the wall.

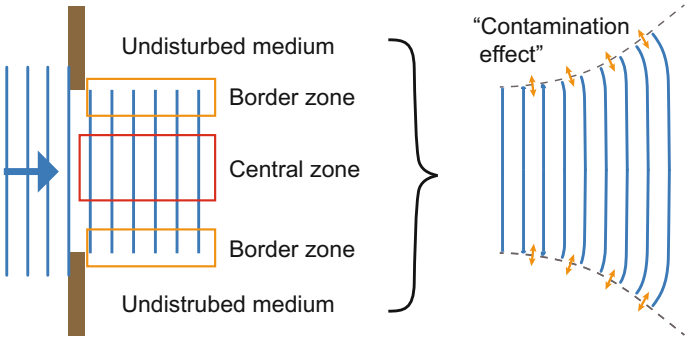

Fig. 13.17 With the starting point in the previous figure, we have tried to illustrate the situation when diffraction does not occur. See the text for further discussion of this hypothetical case

Diffraction causes the waves on the opposite side of the wall to adopt a fan shape if the slit is narrow (e.g. with a width approximately equal to the wavelength). If the gap is much wider, that is, a few wavelengths wide, the diffraction will lead to waves approximately as shown in the right part of the figure. The question is then: What are the mechanisms behind this diffraction?

In Fig. 13.17, we have shown in the left part how the waves would go after the wide gap if there is no diffraction. Then, the waves would continue as a train of waves having the same length as the width of the slit.

In the central part of the waves, the wave will initially continue as before. This is because the neighbouring area of every part of a wave in the central zone consists of waves that move alike. There are no possibilities for leakage sideways, and the wave continues as before.

In the border regions, the situation is very different. Try to visualize yourself a water wave that is sliced laterally and moves the water surface with a wave on one side and perfectly plane water surface on the other side of the partition. It would just not work out! Water from the wave edge would affect water that was originally thought to be outside the wave. There must be a continuous water surface also along the demarcation. This situation would give rise to a "contamination process" where energy is stolen from the peripheral areas of the waves and fed into the area where there would have been a flat, calm water surface without diffraction.

The contamination process will continue all the time, and the wave will therefore become wider and wider. The very same mechanisms lie behind the contamination process as those which propagate the wave and give it a definite speed. As a result, the diffraction pattern becomes almost the same for any diffraction situation as long as we scale the slit width with the wavelength. The waves will eventually be curved at the edges. Also, the region we called the central zone will eventually feel the influence of the edge, causing the wavefront to take the form of an almost perfect arc when it is far from the slit relative to the width of the slit. The radius of curvature will eventually equal the distance from the slit to the wavefront we consider

(i.e. the waves far from the gap look as if they come from a point in the middle of the opening).

A physical explanation model completely analogous to that used for water waves can be applied to electromagnetic waves. It is impossible to have an electromagnetic field in a continuous medium (or vacuum) where there is a sharp separation between an area with a significant electromagnetic field and an adjacent area (all the way up to the previous) where there is no electromagnetic field. Maxwell's equations will ensure that the electromagnetic field will contaminate the area that without diffraction would be without fields and we have continuity requirements just like surface waves on water.

The key point is that a wave entails an energy transfer from a region in space to the neighbouring area, and such an energy transfer will always take place if there are physical differences between the regions, provided that there is actually a connection between the two areas.

Any situation where we create side effects between regions with waves and adjacent regions without waves (where the two are in contact with each other) is a source of contamination and thus diffraction. Contamination can propagate and appear even after the wave has moved far from the spatial constraints that created the unevenness in wave intensity.

13.8 Diffraction, Other Considerations

We have derived above the intensity distribution of the light falling on a screen after it has passed a narrow slit. The intensity distribution just after the slit can be considered a spatial square pulse. However, the intensity captured on a screen at a large distance shows an intense bell-shaped central peak with fainter lines on either side (see Fig. 13.12). The two closest side peaks have the intensity 4.72 and 1.65% of the intensity of the central maximum. Is there something magical about this change from a square to a bell-shaped intensity distribution? In a way, there *is*.

Figure 13.18 shows the square of the Fourier transform of the product of a sine curve and a square function. The Fourier transformed curve has the exact same shape as the intensity distribution we calculated for diffraction from a single slit. This is an example of a part of the optics called "Fourier optics".

If we multiply a sine function with a Gaussian curve instead of a square function, the square of the Fourier transformed becomes a pure Gaussian curve. If we start experimentally with a Gaussian intensity distribution in a beam, the beam can be made either narrower or wider, using lenses and diffraction, and still retain its Gaussian intensity distribution. In other words, diffraction will not cause any peaks beyond the centre line when the beam has a Gaussian intensity distribution.

It can be shown more generally that the intensity distribution for diffraction from a slit is closely related to the intensity distribution of the beam of light we start with. In other words, intensity distribution can be regarded as a form of "boundary conditions" when a wave spreads out after hitting materials that limit its motion.

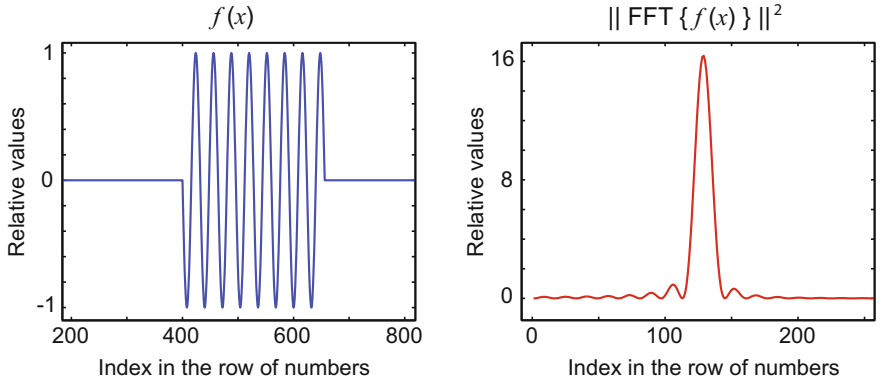

Fig. 13.18 If a sinusoidal signal is multiplied by a square pulse, we get a signal as shown in the left part of the figure (only the interesting area is included). Here, 4096 points are used in the description, the sine signal has 32 points per period, and the square pulse is chosen so that we get eight full periods within the square. If this signal is Fourier transformed and we calculate the square of the absolute value of the Fourier coefficients, we get the curve shown to the right of the figure (only the interesting area is included). The curve has the exact same shape as the curve we calculated for diffraction from a single slit

Modern optics often uses laser beams with Gaussian intensity distribution across the beam. Then, the beam shape will be retained even after the beam is subjected to diffraction.

A beautiful formalism based on matrices (called the ABCD method) has been developed that can be used to calculate how diffraction changes the size of a laser beam (assuming the intensity profile is Gaussian). In this formalism, first and foremost two quantities are included, which are of prime importance for the development of such a beam. One is the diameter of the beam (diameter between points where intensity has fallen to $1/e^2$ of the peak value). The second parameter is the radius of curvature of the wavefront as a function of position along the beam. The formalism is based on "small angles". This is for your orientation.

Test yourself:
The information given in the caption to Fig. 13.18 is associated with the figure itself. If you want to test how much you remember from Fourier transformation, try to answer the following questions:

1. Can you explain why the top of the right part of the figure ends where it is?
2. Is there any connection between the position where the square pulse occurred in the left part of the figure and the position/intensity in the right part of the figure? Explain as usually the answer!
3. If the left-hand square pulse was only half as wide as in our case, how would you expect the right figure to look like?

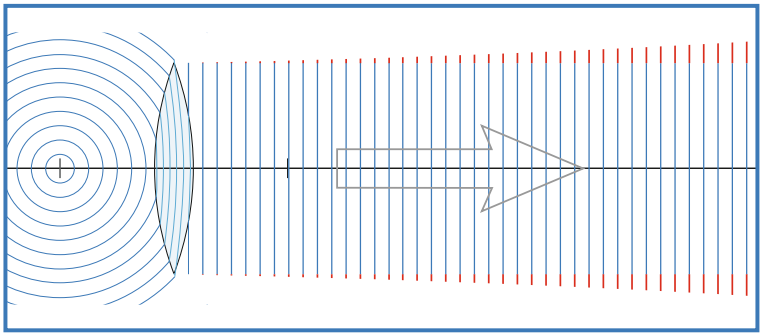

Fig. 13.19 When light from a light source is sent through a lens as shown in this wavefront diagram, diffraction will affect the edge of the light beam (highlighted in red). It is very easy to create this situation, while it is almost impossible to create the reverse process in practice (which corresponds to the reversal of the time flow)

13.8.1 The Arrow of Time

The laws of physics are often such that a process, in principle, works equally well both when time runs forwards and in the reverse direction. When we used light ray diagrams in the previous chapter, the diagram would have remained valid if we had switched the object and the image. The lens formula is also symmetrical in this sense.

The wavefront diagrams in the previous chapter will also be used (when ignoring some shadow effects) both forwards and backwards in time.

The conditions are different when we consider diffraction. In Fig. 13.19, we have examined how diffraction affects the beam shape after light has passed a lens. It is *in principle* possible to run the process backwards in time also in this case, but *in practice* it is impossible. It would require us to reproduce the wave conditions in the smallest detail.

Diffraction is therefore an example of physical processes in which the time arrow in practice cannot be reversed. Perhaps, the best-known example of a similar process is the diffusion of, e.g. molecules in a gas or liquid.

13.9 Numerical Calculation of Diffraction

The derivations we have carried out so far are based on analytical mathematics, which has given us closed-form expressions for intensity distributions in various diffraction patterns. These expressions, though absolutely priceless, are based on approaches that represent only some limiting cases of a far more complex reality.

We will now see how numerical methods can help us calculate diffraction patterns for a much larger range of variation in the parameters that enter a problem.

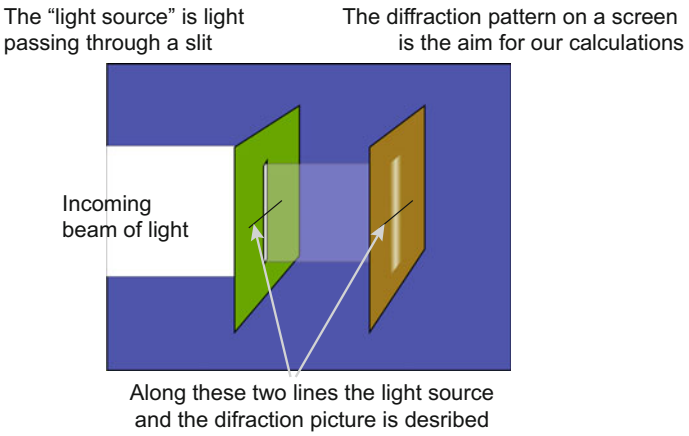

Fig. 13.20 Sketch showing where we describe the light source and the diffraction image by calculating diffraction at a slit

For simplicity, we consider the calculation of diffraction at a slit. Light is incident normally on a flat surface with a rectangular opening, a slit whose length is much larger than the width. The light distribution after the slit has then approximately a cylindrical symmetry, and we therefore consider watching electric fields and intensities along a one-dimensional line across the slit (see Fig. 13.20).

13.9.1 The Basic Model

The model for our numerical calculation is the same as that used for deriving the analytical solutions, except that we do not have to make such drastic assumptions as were introduced earlier. Figure 13.21 shows how we are going to proceed.

We will base our analysis on electromagnetic waves originating from N source points along a line across the slit. The points have positions x_n ranging from $-a/2$ to $a/2$ since the width of the slit is a (see Fig. 13.21). The amplitude of the electric field is A_n, so that the electromagnetic wave at the point x_n is

$$\vec{\boldsymbol{E}}_n = A_n \mathrm{e}^{\mathrm{i}(kz-\omega t+\theta_n)} \vec{\boldsymbol{u}}_n$$

where the symbols have their usual meanings, except $\vec{\boldsymbol{u}}_n$, which is just a unit vector that indicates the direction of the electric field (perpendicular to the direction of motion of the wave at the specified location, assuming a plane polarized wave). θ_n is an angle that gives relative phase from one point to another across the slit. If the wavefront of the incoming light beam is parallel to the plane of the slit, all θ_n are identical, and the parameter can then be dropped.

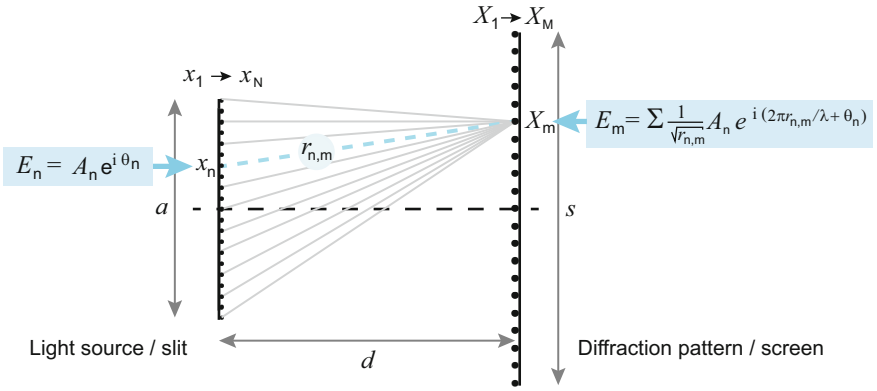

Fig. 13.21 Sketch that indicates how the Huygens–Fresnel principle is used in calculating diffraction from a slit

In our modelling of diffraction, we will take the starting point of electric fields at the same time in the entire slit. Then, $e^{-i\omega t}$ will be a constant phase factor that will disappear when intensities are to be calculated in the end. We therefore drop it at this point. Similarly, we will drop the corresponding factor when calculating the field on the screen where the diffraction image is captured.

If we put the slit in the xy-plane ($z = 0$), we end up with a simplified expression for the electric fields at different points *across the slit*:

$$\vec{\boldsymbol{E}}_n = A_n e^{i\theta_n} \vec{\boldsymbol{u}}_n \qquad (z = 0)\ . \tag{13.10}$$

Let us look at the diffraction pattern captured on a screen parallel to the slit, at a distance d from the slit. Numerically, we calculate the diffraction image in M points symmetrically positioned relative to the centre of the slit. The calculations span a width of s so that the position of the selected points X_m is from $-s/2$ to $s/2$. We must choose a suitable value for s to capture the interesting parts of the diffraction pattern (but not much more).

The electric field at any point X_m will be the sum of contributions from electromagnetic waves coming from all the points x_n in the slit. Since the distance $r_{n,m}$ between the relevant points changes as we pass all x_n, the contributions will have different phases at the screen. In addition, the distance differences make the amplitude of the electric field reduced. In total, we then get the following expression for summation of all contributions to the electric field at point X_m:

$$\vec{\boldsymbol{E}}_m = \sum_n \frac{A_n}{\sqrt{r_{n,m}}} e^{i(2\pi r_{n,m}/\lambda + \theta_n)} \vec{\boldsymbol{u}}_{n,m} \qquad (\text{since } kr = 2\pi r/\lambda)\ .$$

The expression is problematic, because there is no easy way to find the $\vec{\boldsymbol{u}}_{n,m}$ on each electric field contribution (unless the light is polarized in a special way).

We are therefore more or less forced to process electric fields as scalar quantities in such formalism. As already mentioned earlier in this chapter, this is not a big problem when we consider the diffraction image far from the slit. However, very close to the slit, the scalar approach will be a clear source of error in our calculations.

The basic expression for numerical calculation of diffraction from a slit is then:

$$E_m = \sum_n \frac{A_n}{\sqrt{r_{n,m}}} \mathrm{e}^{\mathrm{i}(2\pi r_{n,m}/\lambda + \theta_n)} \tag{13.11}$$

where

$$r_{n,m} = \sqrt{d^2 + (X_m - x_n)^2}\ . \tag{13.12}$$

The intensity at any point is proportional to the square of the electric field.

Note that we have used the square root of the distance when calculating reduced electric field strength. This is because we have cylindrical symmetry. If we send out light along a line, the intensity through any cylindrical surface with the centre of the line will be the same. The area of the cylindrical surface is $2\pi r L$, where L is the length of the cylinder. Since the intensity is proportional to electric field strength squared, then the electric field itself must decrease as $1/\sqrt{r}$. Had we had spherical geometry, the intensity would have been distributed on spherical surfaces with an area of $4\pi r^2$, and the electric field would decrease as $1/r$.

13.9.2 Different Solutions

Calculations based on the expressions (13.11) and (13.12) may be demanding in some contexts, as calculations of sines, cosines, squares and square roots are included in each term. In addition, it needs $N \times M$ calculations. For modern computers, this is very affordable for straightforward calculations of diffraction. Nonetheless, if the diffraction calculations are included in more comprehensive calculations of image formation based on Fourier optics and more, the above expressions are in fact a bit too computer-intensive even today.

Historically, therefore, different simplifications have been made in relation to the above expressions in order to reduce the calculation time. In many current situations where we study diffraction images of light, $a \ll d$ and $s \ll d$ are in Fig. 13.21. We can then use a Taylor expansion in the expression of $r_{n,m}$ instead of Eq. (13.12). The result is (you may try to deduce the expression for yourself):

$$r_{n,m} = \sqrt{d^2 + (X_m - x_n)^2} \approx d\left(1 + \frac{1}{2}\frac{(X_m - x_n)^2}{d^2} - \frac{1}{8}\frac{(X_m - x_n)^4}{d^4}\right)\ . \tag{13.13}$$

In Eq. (13.11), the most important term, namely $r_{n,m}$, occurs in the factor $e^{i2\pi r_{n,m}/\lambda}$. If we substitute the approximate expression for $r_{n,m}$, we get:

$$\mathrm{e}^{\mathrm{i}2\pi r_{n,m}/\lambda} \approx \mathrm{e}^{\mathrm{i}2\pi d/\lambda}\ \mathrm{e}^{\mathrm{i}\pi \frac{(X_m - x_n)^2}{d}/\lambda}\ \mathrm{e}^{-\mathrm{i}\pi \frac{1}{4}\frac{(X_m - x_n)^4}{d^3}/\lambda}\ . \tag{13.14}$$

or with a more readable way to write exponentials:

$$\exp\left[\mathrm{i}2\pi r_{n,m}/\lambda\right] \approx \exp[\mathrm{i}2\pi d/\lambda]\ \exp\left[\mathrm{i}\pi \frac{(X_m - x_n)^2}{d}/\lambda\right]\ \exp\left[-\mathrm{i}\pi \frac{1}{4}\frac{(X_m - x_n)^4}{d^3}/\lambda\right]$$

In different situations, some of these terms will be practically constant, and this is precisely the basis of some historical classifications of diffraction.

We will now try to provide an overview of different variants of computational accuracy:

1. **Less than a few wavelengths away from the edges of the slit**. Here, we must use Maxwell's equations and bring polarization and surface currents in the material surrounding the slit. "Evanescent waves" are part of the solution. (This is a complicated calculation!)
2. **For** $d^3 \leq 2\pi a^4/\lambda$. This is a problematic area where Maxwell's equations can be used for the smallest d, while the expressions (13.11) and (13.12) begin to work reasonably well for the largest d which satisfies the stated limit.
3. **For** $d^3 \gg 2\pi a^4/\lambda$, **we have Huygens–Fresnel diffraction**. The expressions (13.11) and (13.12) work. Even if we put $1/\sqrt{r_{n,m}} = 1/\sqrt{d}$ and we skip the last term of the Taylor expansion in Eq. (13.13), the result will be satisfactory.
4. **For** $d \gg \pi a^2/\lambda$, **we have Fraunhofer diffraction**. The expressions (13.11) and (13.12) work. Although we use the same approaches as for Huygens–Fresnel diffraction, and then $(X_m - x_n)^2 \approx X_m^2 + 2X_m x_n$ in the middle of the series of Eq. (13.13), the results will be satisfactory.

Figure 13.22 displays numeric calculations based on the expressions (13.11) and (13.12) directly. In the first case, we are relatively close to the slit (Huygens–Fresnel zone), while in the second case we are in the transition between the Huygens–Fresnel and Fraunhofer zones.

Note that when we are near the slit (Huygens–Fresnel zone), the diffraction image on the screen will have approximately the same size as the slit. However, some of the intensity at the edge of the slit leaks into the shadow section (marked with arrow in the figure), resulting in a continuous intensity distribution between shadow and full light intensity. We get characteristic fringe patterns in the image of the slit. There are larger "spatial wavelengths" on these fringes near the edge of the slit than towards the centre. There are only faint fringes in the shadow section on each side of the image of the slit.

Figure 13.23 shows a photograph of two diffraction patterns that have features similar to those used in the numerical calculation.

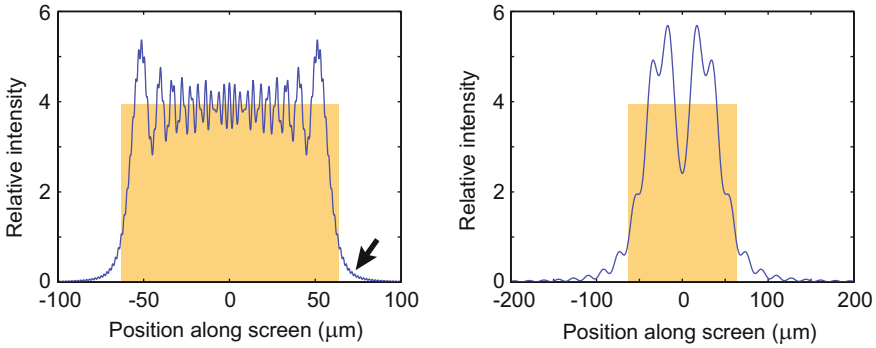

Fig. 13.22 Diffraction from a slit calculated from the Huygens–Fresnel principle. The left part of the figure corresponds to the screen being fairly close to the slit. The right part depicts the situation a little farther away from the slit, yet not quite as far as in Fraunhofer diffraction, which was treated analytically earlier in the chapter. The width of the slit is marked with a yellow rectangle

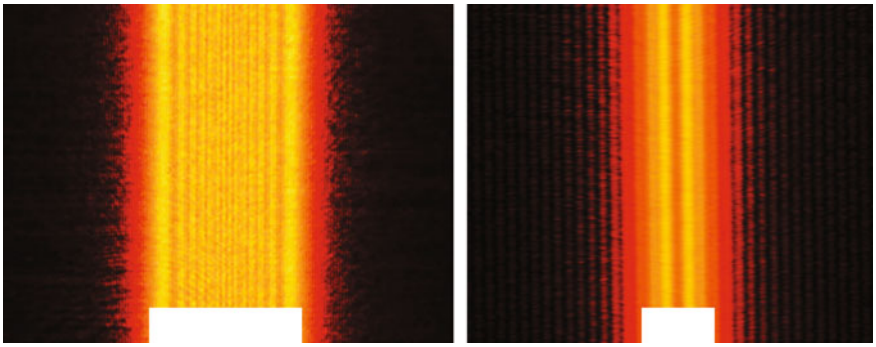

Fig. 13.23 Photograph of diffraction image of a slit with approximately the distances that were used in the calculations in Fig. 13.22 correspond to. The size of the slit is marked at the bottom

Similarly, we have shown calculations and an example of diffraction pattern in the border region between the Huygens–Fresnel and Fraunhofer zones in the right part of Figs. 13.22 and 13.23. We see some wavy features here both in the image of the slit and in the light falling on the shadow zone.

The Fraunhofer zone diffraction pattern is exactly the same as that derived analytically, and illustrative results are already given in Fig. 13.12 and a photograph in Fig. 13.14. In that case, we only have wavy features in the zone outside the central peak.

13.10 Diffraction from a Circular Hole

When a plane wave is sent to a circular hole, we also get diffraction (see Figs. 13.24 and 13.26), but it is more difficult to set up a mathematical analysis of that problem than for slits. As a result, the image that can be collected on a screen shows a distinctly

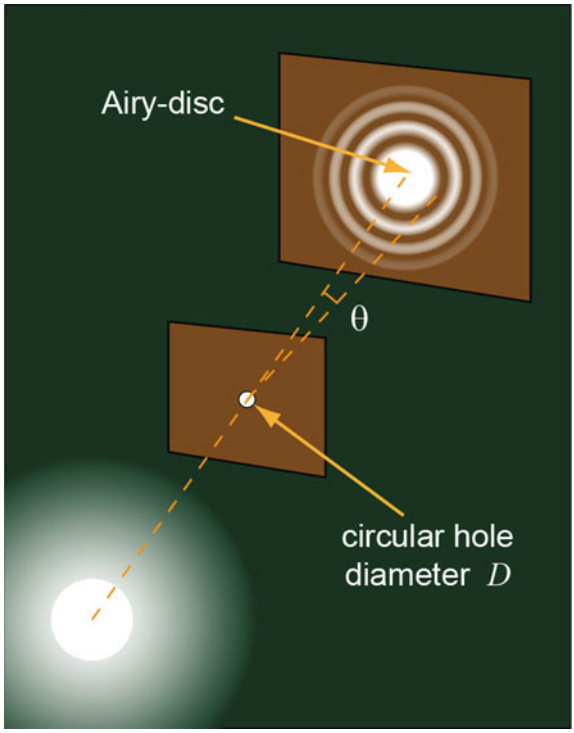

Fig. 13.24 Experimental set-up for observing diffraction from a circular hole

central bell-shaped peak, with weak circles. The central peak seems to form a circular disc, called the "Airy disc".

Mathematically, the intensity at an angular distance θ away from the centre line is given by:

$$I(\theta) = I_{\max}\left[1 - J_0^2(\tfrac{1}{2}kD\sin\theta) - J_1^2(\tfrac{1}{2}kD\sin\theta)\right]$$

where J_n denotes the Bessel function of the first kind of order n, D is the diameter of the hole, and the k is the wavenumber. When the distance to the screen is much larger than the diameter of the hole, the intensity distribution becomes:

$$I(\theta) = I_{\max}\left[\frac{2J_1(\frac{1}{2}kD\sin\theta)}{\frac{1}{2}kD\sin\theta}\right]^2$$

where the values of Bessel functions can easily be calculated numerically from the expression:

$$J_n(x) = \frac{1}{\pi}\int_0^{\pi}\cos(n\tau - x\sin\tau)\mathrm{d}\tau\ .$$

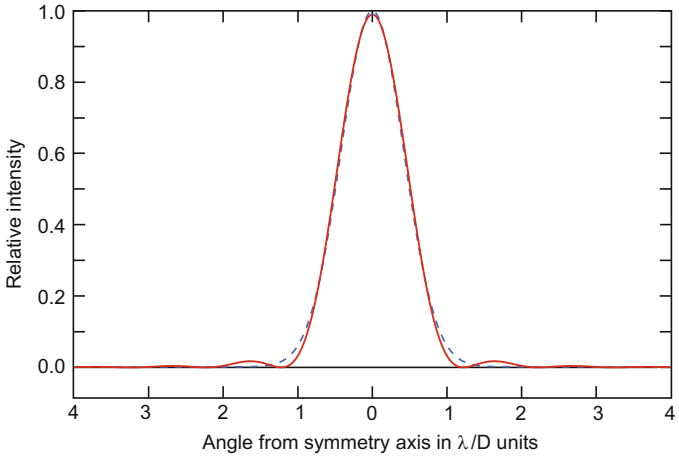

Fig. 13.25 Diffraction pattern for a circular hole far from the hole (*red*). The central peak that makes up the Airy disc has an intensity profile shape very close to a Gaussian (*blue dashed line*)

The angle for the first minimum is given by:

$$\sin\theta = \frac{1.22\,\lambda}{D}$$

where D is, as stated above, the diameter of the hole. Since the angle is usually very small, we can use the approximation:

$$\theta = \frac{1.22\,\lambda}{D}\,. \tag{13.15}$$

The next pair of dark rings have a radius of $2.232\,\lambda/D$ and $3.238\,\lambda/D$, and the intensities of the first three rings are 1.75, 0.42 and 0.16% of the intensity at the centre of the central disc. See Figs. 13.25 and 13.26.

It is of interest to look into some details within the central peak both for a single slit and for an Airy disc. From Fig. 13.13, we can deduce that within the central peak of the single-slit diffraction pattern, the wavefront is very close to plane (given that the slit was illuminated by a plane wavefront). The deviation from the perfect plane wavefront is $\lambda/2$ or less. This is remarkable, since the central peak at a screen easily can be several thousand times as wide as the slit itself (for narrow slits).

Similarly, reasoning along the same lines, it can be shown that within the central peak (the Airy disc) of the diffraction pattern from a circular hole, the wavefront is very close to plane. The maximum deviation from the perfect plane wavefront is slightly larger than $\lambda/2$. Since the intensity of the rings around the Airy disc is much

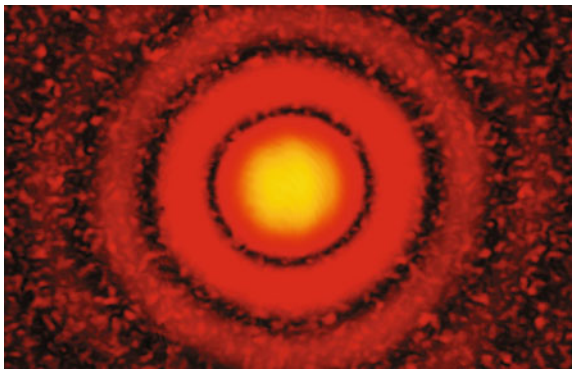

Fig. 13.26 Airy disc as it looks with some overexposure in the central portion to get the surrounding circles. Overexposure is difficult to avoid since the maximum intensity in the first ring is only 1.75% of the maximum intensity in the central disc. There are some "speckles" (bright spots), probably due to scattered laser light in the room

less than for the central peak, the diffraction leads to a kind of transformation from a narrow circular beam with constant intensity and flat wavefront throughout the cross section, to a much wider beam close to Gaussian intensity profile and a much larger cross section, but even so, with an almost flat wavefront.

The expression in Eq. (13.15) and the diffraction-of-light-through-a-circular-hole phenomenon has far-reaching consequences, and we shall mention some.

13.10.1 The Image of Stars in a Telescope

Light from a star comes towards a telescope. The light can be considered a plane wave when it reaches the objective, and the light is focused by a lens or a mirror. In geometric optics, we get the impression that we can collect all the light rays from a distant object at one point, the focal point, as indicated in the left part of Fig. 13.27. At the very least, it should be possible if the angular diameter of the object is very small, such as when we look at the stars in the sky. That is wrong!

The light beam from a star will follow a shape similar to that shown in the right part of Fig. 13.27. The light bundle has a minimum diameter of d which is significantly larger than what we would expect from the angular diameter of the object (the star). The reason is diffraction.

It is actually diffraction from a circular hole we witness. However, it is now a large hole with diameter equal to the diameter D of the objective of the telescope. According to Eq. (13.15), the angle to the first minimum will be very small. It will not be observable if we did not focus on the beam by the telescope objective.

We cannot use Eq. (13.15) directly to calculate the Airy disc size when we focus on the beam the way we do. We therefore use another kind of reasoning and refer to Fig. 13.28.

The light passes in the real life through the objective with diameter D, is focused and forms a diffraction picture where the Airy disc diameter is d in the waist and then continues to the right. However, waves can in principle move backwards as well

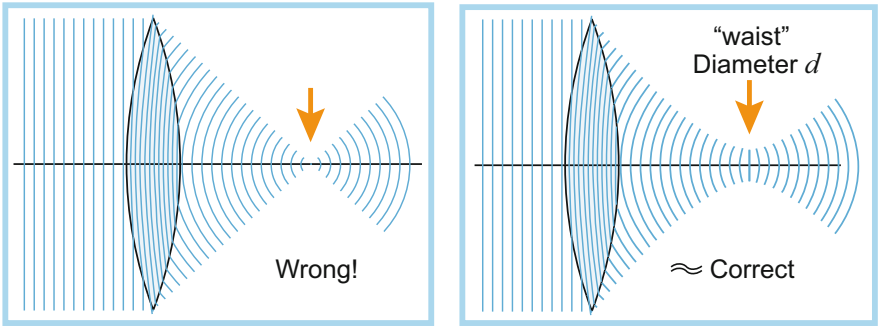

Fig. 13.27 According to ray optics, light from a tiny distant star should be focused on a point one focal length behind a telescope objective (*left*). However, due to diffraction the light beam will have a finite diameter d at its minimum before the beam size increases again (*right*). The part of the beam with minimum diameter is denoted the "waist". The Airy disc will in practice be *much* smaller than shown in the right part of this figure, but *in principle* it will have this pattern. Note also the difference in wavefronts in the two descriptions

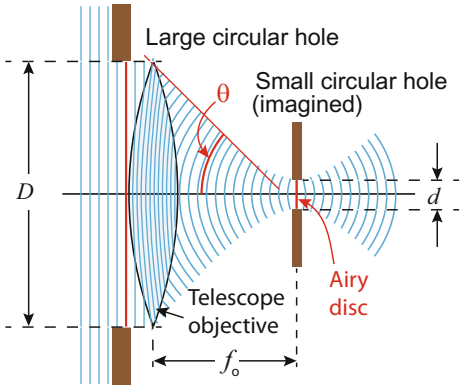

Fig. 13.28 To get an estimate of the size of the Airy disc in the case where light passing through the telescope objective is focused, we can imagine that the waves are moving backwards *from* the Airy disc in the waist *towards the objective*. See text for details

as forwards and follow the same pattern. Thus, we may imagine that we actually start with a reasonably flat wavefront passing through a small circular hole of the same size as the Airy disc, and let the wave move *to the left*. We should then have a situation not very differently from the starting point which lead to Eq. (13.15). The diameter of the beam should increase in size so that the diameter is roughly equal to D when the beam reaches the objective's position.

Based on the wave-moving-backwards argument, diffraction from the tiny Airy disc at the waist will cause the light beam to spread out at an angle θ (relative to the centre line of the beam)

$$\theta \approx \frac{1.22\,\lambda}{d}$$

This diverging conical light bundle due to diffraction must match in its extension the converging conical light beam from the convex lens. Thus:

$$\frac{1.22\,\lambda}{d} = \theta \approx \sin\theta \approx \tan\theta = \frac{D/2}{f_o}$$

where f_o is the focal length of the objective.

The radius of an Airy disc in the focal plane will then be:

$$\frac{d}{2} = \frac{1.22\,\lambda f_o}{D}\ .$$

The Airy disc of a star will have this extension, even though the angular extension in the sky is vanishing small. All stars will make equally sized luminous discs in the focal plane, but the intensity of the disc will reflect the brightness of the star under observation.

You will certainly have noticed that our argument with backward waves has obvious difficulties. We mix perfectly plane wavefronts with a wavefront not exactly flat, and we neglect the difference between a constant intensity across a hole and a more Gaussian intensity profile. We also neglect the rings around the Airy disc. Even so, a more rigorous treatment leads to roughly the same conclusion as we have arrived here.

The diffraction has important consequences. Two stars close to each other in the sky will form partially overlapping discs in the focal plane. If the overlap is very large, we will fail to notice that there are two discs, and will consider them as one. If the overlap is small, we will conclude that there are two discs, representing two stars.

Lord Rayleigh addressed this problem in the following manner:

When two objects (or details in objects) are viewed in a telescope, the ability to separate the two objects will reach its limit when the central maximum in the Airy disc of one object coincides with the first diffraction minimum of the other object. This description is known as **Rayleigh's Resolution Criterion**.

The minimum angle ψ where we can see that there are two Airy discs is then given by:

$$\psi \approx \frac{d/2}{f_o} = \frac{1.22\,\lambda}{D} \,. \tag{13.16}$$

In other words, with a lens of diameter D and focal length f_o, we can distinguish two stars (or other point-like objects) from each other if the angular distance between the stars is at least ψ.

Examples:
As we have just seen, we are not able to distinguish detail that subtends an angle of less than $1.22\lambda/D$, no matter how much we enlarge the image. For a prism binocular with an objective of about 5 cm diameter, the smallest angular distance we can resolve with 500 nm light becomes

$$\frac{1.22 \times 500 \times 10^{-9}}{0.05}$$

which corresponds to 0.00069°. For the Mount Palomar telescope, with a mirror of 5 m diameter, the best resolution is 1/100 of this angle. The Mount Palomar telescope can resolve details that are approximately 50 m apart from each other on the moon, while a prism binocular will only be able to resolve details located 5 km from each other.

The diameter of the pupil in our eye is about 5–10 mm in the dark. This means that without the help of aids we can only distinguish details on the moon which is at least 25–50 km apart (the moon's diameter is 3474 km).

In a prism binocular, the magnification is almost always so small that we cannot see the Airy disc. In a telescope where we can change eyepieces and the magnification can be quite large, it is common to see the Airy discs. A star does not look like a point when viewed with a large magnification through a telescope. The star looks exactly like the diffraction image from a small circular opening on a screen, with a central disc (Airy disc) surrounded by weak rings. The rings are often so faint that it is hard to spot them.

The optical quality of many binoculars and telescopes are so poor, that e.g. spherical aberration, chromatic aberration or other imperfections so that we do not get a nice Airy disc if we enlarge the image of a star. Instead, we get a more or less irregularly illuminated surface that covers an even greater angular range than the Airy disc would have done. For such telescopes, we fail to resolve the fine details that the Rayleigh criterion indicates.

A telescope so perfect that its resolution is limited by the Airy disc is said to have *diffraction-limited optics*. This is a mark of excellence!

Today, it is possible to use numerical image processing in a smart way so that we can reduce the effect of diffraction. We theoretically know what intensity distribution we will get when light from a fictitious point source goes through the optical system we use (telescope or microscope). By an extensive iterative method, one can then slowly but surely generate an image with more details than the original. The image so can get close to represent what we would observe in the

absence of diffraction. In this way, today, in favourable situations, we can attain about ten times better resolution in the images than can be achieved without the extensive digital image processing.

13.10.2 Divergence of a Light Beam

At the Alomar Observatory on Andøya, an ozone detector has been installed where a laser beam is sent 8–90 km up in the atmosphere to observe the composition and movements of molecules up there. The beam of light should be as narrow as possible far up there, and we can wonder how this may be achieved.

The first choice might be to apply a narrow laser beam directly from a laser. The beam is typically 1–2 mm in diameter. How wide would this beam be, for example, at a height of 30 km?

We use the relationship of diffraction from a circular hole and find the divergence angle θ:

$$\sin\theta = \frac{1.22 \times \lambda}{D} \ .$$

For light with wavelength 500 nm and an initial beam diameter of 2.0 mm, at the start, we get:

$$\sin\theta = \frac{1.22 \times 500 \times 10^{-9}}{0.002} = 3.05 \times 10^{-4} \ .$$

The angle is small, and if the radius of the beam at 30 km height is called $D_{30\,\mathrm{km}}$, we find:

$$\frac{D_{30\,\mathrm{km}}/2}{30\,\mathrm{km}} = \tan\theta \approx \sin\theta = 3.05 \times 10^{-4}$$

$$D_{30\,\mathrm{km}} = 18.3\,\mathrm{m} \ .$$

In other words, the laser beam that was 2 mm in diameter at the ground has grown to 18 m in diameter at 30 km altitude!

An alternative is to expand the laser beam so that it starts out much wider than the 2 mm. Suppose we expand the beam so that it is actually $D = 50$ cm in diameter at the ground. Suppose the wavefront is flat at the ground so that the beam at the beginning is parallel (so-called waist) and eventually diverges.

How big will the diameter be at $R = 30$ km height?

We must be meticulous in stating the divergence angle:

$$\frac{D_{30\,\mathrm{km}}/2 - D/2}{R} \approx \tan\theta \approx \sin\theta = \frac{1.22 \times \lambda}{D} \ .$$

On solving this equation for $D_{30\,\mathrm{km}}$, we get 57.3 cm. In other words, a beam that starts out as 50 cm wide only becomes 57.3 cm wide at 30 km height! This is significantly better than if we start with a 2 mm thin beam.

We can, however, make things *even* better! We can choose not to place the laser (light source) exactly at the focal point of the 50 cm mirror we used in town (as a part of making the beam wide). If we place the laser slightly beyond the focal point, the beam will actually converge before it reaches the "waist" (corresponding to the Airy disc) and then diverges again. See Fig. 13.27. How small can we make the waist (Airy disc) at 30 km altitude?

We can then work *backwards* and consider the "waist" at 30 km height to be the *source* of a diverging beam (on both sides of the waist, since we have symmetry here). In that case, the beam will on its way from the waist to the mirror have diverged to D equal to 50 cm at the location of the mirror (imagining that the beam goes backwards). The calculator will look like this:

$$\frac{D/2 - D_{30\,\mathrm{km}}/2}{R} \approx \tan\theta \approx \sin\theta = \frac{1.22 \times \lambda}{D}$$

$$D_{30\,\mathrm{km}} = 42.7\,\mathrm{cm}\ .$$

In other words, we can even get a smaller beam than the one we started with.

Conclusion: A laser beam that has a 2 mm diameter at the ground becomes 18 m in diameter at 30 km altitude. However, if we start with a beam of 50 cm diameter and focus it so that the waist will be found 30 km above the ground, the beam is "only" 43 cm in diameter at this height. The energy density in the cross section is then over 400 times as great as in the first case.

*13.10.3 Other Examples **

1. Laser beams have often a "Gaussian intensity profile". It may be shown that if you send such a beam through mirrors and lenses, the Gaussian shape will be preserved ("beam optics"), even though the width is changed. We do not get any diffraction rings around the central beam of a Gaussian beam.
2. Diffraction takes place even in our eyes. The pupil's opening is typically 6 mm or less during daily tasks. Thus, diffraction sets the lower limit on the angular distance between two details in our visual field that we can distinguish. Another limitation for the resolving ability of the eye is the size of the light-sensitive cells in the retina. Evolution seems to have chosen an optimal solution since the size of the Airy discs is roughly the same as the effective area of our light-sensitive cells (rods and cones).
3. A camera is not necessarily well adapted. If we choose a sensor chip that gives many pixels per image, it does not necessarily mean that we can *exploit* this resolution. If the Airy disc for the selected lens and aperture (see Chap. 12) is larger than the size of a pixel in the sensor chip, the effective resolution in the image is not as good as the number of pixels indicates. You may test this on your own camera!

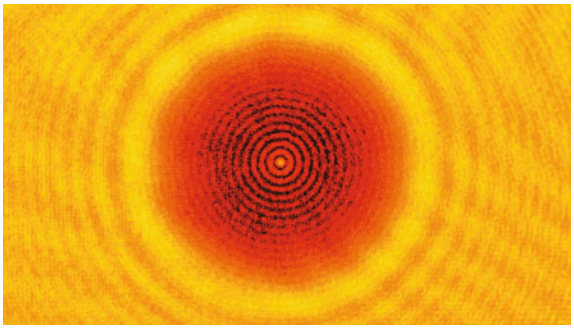

Fig. 13.29 Photograph of Arago's spot in the shadow image of a small ball. The ball was held in place by gluing it up with a small (!) drop of glue on a thin piece of microscope cover glass. In addition to Arago's spot, we see a number of details due to diffraction, both in the shadow section and in the illuminated party. Note that there is no clear boundary between shadow and light

4. As we have seen, the width of the central peak in the diffraction image from a single slit is given by:

$$\Delta\theta_{1/2} = \frac{\lambda}{a} \,. \tag{13.17}$$

   In quantum physics, this result has occasionally been taken as a manifestation of Heisenberg's uncertainty relationship. Rightly enough, the expression in Eq. (13.17) may support this point of view if we threat the phenomena superficially, but there are so many other details in our descriptions of diffraction that the Heisenberg's uncertainty relationship cannot give us.
   Also, in other parts of this book we have had relationships that are reminiscent of Heisenberg's uncertainty relationship. In all these situations, there are fundamental wave features that lie behind.
   It is therefore no wonder that many today perceive Heisenberg's uncertainty relationship as a natural consequence of the wave nature of light and matter and that it has only a secondary link with the uncertainty of measurement.
5. Diffraction has played an important role in our perception of light. At the beginning of the nineteenth century, Poisson showed that if light had a wave nature and behaved according to Huygens's principle, we would expect to see a bright spot in the shadow image of an opaque sphere (or circular disc). Arago conducted the experiment and found that there was indeed a bright spot in the middle (see Fig. 13.29). The phenomenon now goes under the name of Arago's spot (or the Poisson–Arago spot).

Fig. 13.30 Photograph from a military grave field in San Diego, where the graves are placed very regularly. In some directions, we see many gravestones in line. If these tombs emit elemental waves, we would get interference patterns similar to those we have discussed for gratings in this chapter. We would get a series of interference fringe pattern, but the centre of each set would correspond to the different directions to the lines we see in the image. The distance between the lines in each set would depend on the distance between the source points along the direction we consider

13.10.4 Diffraction in Two and Three Dimensions

In our treatment of interference and diffraction, we have so far only considered the summation of waves from elementary wave sources located along a straight line. It is a normal situation for interference and diffraction of light.

For other types of waves, we can find diffracting centres that form two- or three-dimensional patterns. Most well known is perhaps X-ray diffraction. When X-rays are transmitted to a crystal of some substance, single atoms will spread the X-rays so that the elementary waves come from each individual atom in the area of the X-ray.

The atoms in a crystal are in a regular pattern. If we pick out atoms that are on a line in a plane, the elementary waves from these atoms will provide interference lines or interference points that can be calculated with similar equations as those we have been through in this chapter.

Both physics and chemistry provide so-called X-ray diffraction information that can be used to determine the structure of the crystals under investigation. It is this type of research that lies behind almost all the available detailed information about positions of the atoms in relation to each other in different substances.

Figure 13.30 illustrates that points that are regular to each other form lines that can cause interference/diffraction in many different directions.

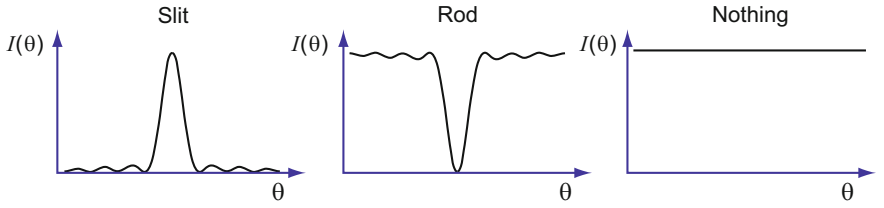

Fig. 13.31 Intensity distribution from a slit and a stick is complementary (really only when we operate at amplitude level and not at intensity level as here). In this case, "Nothing" means light with even intensity everywhere within the given θ interval

13.11 Babinet's Principle

The superposition principle can be used in a very special way where we utilize symmetries.

We have learned the form of the intensity distribution of light diffracted by a narrow slit. How would the interference pattern look at the complementary structure, which is a stick of exactly the same size as the slit? Babinet's principle tells us something about this:

> *Suppose that a wave is sent to a body* **A** *that leads to diffraction (e.g. a long slit a light-tight screen). We send the same wave to a body* $\mathbf{A}'$ *complementary to the first one (e.g. a long stick with the same position and width as the slit on the screen of* **A**), *and we see a different diffraction. If we overlap the diffracted wave in the first case with the diffracted wave in the other, we get a waveform that is identical to the one we would have had if neither* **A** *nor* $\mathbf{A}'$ *existed.*

Figure 13.31 shows the principle. The figure is a simplification, since we specify intensities, but *summation of waves occurs always at amplitude level*.

If we send a relatively narrow laser beam towards a gap and then to a thread of the same thickness as the slit, we can use the principle of Babinet to find out how the two diffraction patterns relate to each other. However, the relationships are quite different from the very wide light-beam/plane-wave situation we show in Fig. 13.31. Outside the narrow laser beam, the intensity is virtually zero when the gap or thread is not found in the light path. But with a slit or thread inside the narrow laser beam, we get diffraction patterns also in the area where there would otherwise be no light. This can be understood by the fact that the superposition principle can only be applied at amplitude level, not at intensity level. A wave $E = E\cos(kz - \omega t)$ will have the very same nonzero intensity as a wave $E = -E\cos(kz - \omega t)$, but the sum of these two waves is zero.

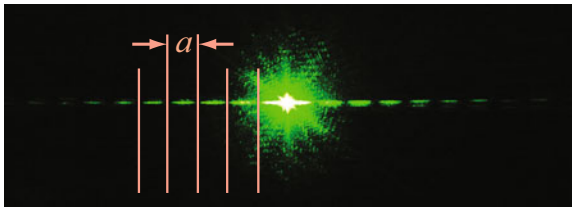

Fig. 13.32 Diffraction of a laser beam from a human hair

Babinet's principle is handy because we can use the theory of diffraction from a slit also by analysing the diffraction image from a thread. Figure 13.32 shows the diffraction image from a single hair placed in the beam from a laser pen. With very simple means, by measuring the distance between the minimum points between the light spots we can determine the thickness of the hair, provided we know the wavelength of the laser. A task at the end of this chapter provides a concrete example of how a boiled measurement may fall.

13.12 Matlab Code for Diverse Cases of Diffraction

Here is a Matlab program that can be used to see examples of how diffraction is affected by different initial light intensity patterns. The program is especially interesting for calculating diffraction when it is not a very long distance between, for example, a single (or double) slit and the screen where we capture the diffraction image. Then, there is a wealth of details that cannot be calculated analytically, but that matches what we can observe in experiments.

For all excitations, we assume that we have coherent light with wavefront in the excitation plane and that we have a form of cylindrical symmetry for each strip in the excitation plane.

The program must be used with a code (a number) as a parameter, such as:

```
diffraction(4)
```

if the intensity distribution on a screen after a double slit is to be calculated. Widths of gap, distance between slots and screens, etc., must be manually entered in the program (in function parameters). A bit of trial and error is required for the calculation area to cover the entire diffraction image we are interested in (but also not much more). Good luck!

Main Program

The code is available at the "supplementary material" Web page for this book at http://www.physics.uio.no/pow.

```
function diffraction(code)

% This program calculates and plots intensity patterns for a
% variety of diffraction and/or interpherence phenomena with
% cylindrical symmetry.
% Functionalities: code = 1: One slit, 2: Gaussian intensity
% profile, 3: Straight edge, 4: Double slit, 5: Read excitation
% data from file (amplitude + phase)
% Program is written by AIV. Version 15. October 2017

% Establishes essential parameters for the calculations.
% Results depend critically on these. See the code for this
% function.

[lambda,a,b,nWavel,N,twopi,Nhalf] = parameters;

% Allocates arrays for the calculations:
[x,x2,x0,x1,sc,r] = allocateArrays(nWavel,N);

% Generate or read in excitation data:
[x0] = generateExcitation(code,lambda,a,N,Nhalf,twopi,x0);

% Calculates sines, cosines, distances and relative phase
% differences for vectors between the plane of excitation and
% the screen for the final pattern:
[sc,r] = generateRelPositionData(N,b,lambda,twopi);

% Sum all contributions to every point on the screen (main
% loop):
[x1] = summation(N,x0,r);

% Plots intensities for diffraction pattern along with a
% marking of the excitation
plotDiffraction(x,x0,x1);

% Calculates and write out linewidths in case the excitation
% was a single slit or Gaussian profile, and write to
% screen the actual linewidth of the intensity profile.
if (code==1) || (code==2)
    linewidth(N,lambda,x1);
end;

% Plots expected theoretical intensity profile for a single
% slit:
if code==1
```

```
    plotTheoreticalSingleSlit(N,a,b,twopi,x,x1);
end;

% Option: Save data to a file (as a string of floating point
% numbers):
%writeToFile(x1);

% Removes all plots when we leave the program (cleans up):
input('Close all figures');
close all
```

Choose Parameters for the Calculations

```
function [lambda,a,b,nWavel,N,twopi,Nhalf] = parameters

% Choose parameters for the calculation.
% Written by AIV. Version 15. October 2017

% Choose resolution, distance to screen, and the width of the
% area on screen the calculation should include. Some constants
% are defined. Results depend critically on the parameters
% set by this function.
% Whether the result will be mainly a Fresnel- og Franuhofer
% diffraction depend on the b parameter. nWavel must be
% increased if b is large to include the full diffraction
% pattern within the calculated area on screen.
% The parameters given in this particular code is suitable
% for a double slit in the Fresnel regime (quite complicated
% pattern).

lambda = 4;         % Four points per wavelength resolution in
% excitation points
a = 20;             % Width of single slit, given in
% # wavelengths
b = 4000 * lambda;  % Distance to screen is b wavelengths
nWavel = 1024*3/2;  % # wavelengths along the screen (an
% integer!)
N = nWavel*lambda;  % Width of excitation area as well as
% screen in # wavelengths
twopi = 2.0*pi;     % Somewhat innecessary, but speeds up
% a bit...
Nhalf = N/2;
return;
```

Allocate Arrays We Need

```
function [x,x2,x0,x1,sc,r] = allocateArrays(nWavel,N);
% Allocates space for various arrays
% Function is written by AIV. Version 15. October 2017

x = linspace(-nWavel/2, nWavel/2, N); % A relative position
% array for plot
x2 = linspace(-N,N,2*N+1);  % Simil, but for plot/test of
% hjelp functions
x0 = zeros(N,2);            % Excitation data, amplitudes
% and phases
x1 = zeros(N,2);            % Amplitudes at screen, amplitudes
% and phases
sc = zeros(2*N + 1,2);     % Store sin/cos for component
% calculations
r = zeros(2*N + 1,2);       % Distance-table: reduction
% factor and phase-correction
                            % based on path length
return;
```

Generates the Various "Excitations" (Single or Double Slit, etc.)

```
function [x0] = generateExcitation(code,lambda,a,N,Nhalf, ...
twopi,x0)
% Generate or read in excitation data. NOTE: There are
% specific requirements for the various excitations that
% can only be changed in the code below.
% Function is written by AIV. Version 15. October 2017

switch code
   case (1)
      disp('Single slit')
      m = a * lambda / 2;   % Slit is a wavelengths wide
      x0(Nhalf-m:Nhalf+m-1,1) = 1.0;
      %x0(:,2)= [1:N].*0.05; % Phases are modifies so that
                             % it mimics a ray is not coming
                             % perpendicular towards the slit.
   case 2
      disp('Gaussian excitation')
      % Intensity
      width = 200*lambda/2.0;
      dummy = ([1:N]-Nhalf)./width;
      dummy = (dummy.*dummy);
```

```
    x0(:,1) = exp(-(dummy));
    % Phase
    R = 1000; % Radius of curvature in # wavelengths
    y = [-Nhalf:Nhalf-1];
    R2 = R*R*lambda*lambda*1.0;
    dist = sqrt((y.*y) + R2);
    fs = mod(dist,lambda);
    x0(:,2) = fs.*(twopi/lambda);
    %figure;  % Plot if wanted
    %plot(x,x0(:,2),'-r');

 case 3
    disp('Straight edge')
    % Excitation is a straight edge, illuminated part: 3/4
    x0(N/4:N) = 1.0;

 case 4
    disp('Double slit')
    % For the double slit, use sufficient large b in
    % 'parameters' in order to get the well known result
x0 = zeros(N,2);
a = 20*4;
d = 200*4;
kx = d/2 + a/2;
ki = d/2 - a/2;
x0(Nhalf-kx+1:Nhalf-kx+a,1) = 1.0;
x0(Nhalf+ki:Nhalf+ki+a-1,1) = 1.0;

 case 5
    disp('Reads excitation data from file')
    % (often earlier calculated results.)
    filename = input('Give name on file with excitation ...
       data: ', 's');
    fid = fopen(filename,'r');
    x0(:,1) = fread(fid,N,'double');  % Need to know #
     % elements
    x0(:,2) = -fread(fid,N,'double');
    status = fclose(fid);
    % figure;  % Testplot to check if data was read properly
    % plot(x,xx0(:,1),'-g');
    % figure;
    % plot(x,xx0(:,2),'-r');
    % aa= xx0(Nhalf);
    % aa  % Test print for one single chosen point
```

```
    otherwise
       disp('Use code 1-5, please.')
end;
return;
```

Calculate Relative Position Data (from Excitation to Screen)

```
function [sc,r] = generateRelPositionData(N,b,lambda,twopi);
% Establish sine and cosine values for vectors from one
% position in x0 to all positions in x1, and find distances
% and relative phase differences between the points.
% Function is written by AIV. Version 15. October 2017

y = [-N:N];
b2 = b*b*1.0;
y2p = (y.*y) + b2;
rnn = sqrt(y2p);
sc(:,1) = b./rnn;
sc(:,2) = y./rnn;
r(:,1) = 1./sqrt(rnn);
fs = mod(rnn,lambda);
r(:,2) = fs.*(twopi/lambda);
% mx = max(r(:,1));  % For testing if field reduction vs
 % distance is correct
% r(:,1) = mx;
%  plot(x2,r(:,2),'-k'); % Test plot of these variables
%  figure;
return;
```

Summation of all Contributions

```
{\footnotesize
\begin{verbatim}
function [x1] = summation(N,x0,r)
% Runs through x1 (screen) from start to end and sum all
% contributions from x0 (the excitation line) with proper
% amplitude and phase.
% Function is written by AIV. Version 15. October 2017

for n = 1:N
    relPos1 = N+2-n;
    relPos2 = relPos1+N-1;
    amplitude = x0(:,1).*r(relPos1:relPos2,1);
    fase = x0(:,2) - r(relPos1:relPos2,2);
```

```
    fasor(:,1) = amplitude .* cos(fase);
    fasor(:,2) = amplitude .* sin(fase);
    fasorx = sum(fasor(:,1));
    fasory = sum(fasor(:,2));
    x1(n,1) = sqrt(fasorx*fasorx + fasory*fasory);
    x1(n,2) = atan2(fasory, fasorx);
end;
return;
```

Plot the Diffraction Pattern

```
function plotDiffraction(x,x0,x1);
% Plots intensities for diffraction picture along with a
% marking of the excitation. Some extra possibilities are
% given, for testing or special purposes.
% Function is written by AIV. Version 15. October 2017

%plot(x,x1(:,1),'-r');   % Plots amplitudes (red) (can
 % often be skipped)
figure;
x12 = x1(:,1).*x1(:,1);  % Calculation of intensities
hold on;
scaling = (max(x12)/8.0);
plot(x,x0(:,1).*scaling,'-r');   % Plot initial excitaion
plot(x,x12(:,1),'-b');  % Plot relative intensities (blue)
xlabel('Position on screen (given as # wavelengths)');
ylabel('Relative intensities in the diffraction pattern');

% figure;
% plot(x,x1(:,2),'-k');   % Plot phases (black) (most
  % often skipped)
return;
```

Calculates Linewidths (FWHM) for Single-Slit and Gaussian Intensity Profile

```
function  linewidth(N,lambda,x1);
% Calculates linewidths (FWHM) for single slit and Gaussian
% intensity profile.
% Function is written by AIV. Version 15. October 2017

x12 = x1(:,1).*x1(:,1);  % Calculation of intensities

mx2 = max(x12(:,1))/2.0;
lower = 1;
```

```
upper = 1;
for k = 1:N-1
        if ((x12(k,1)<=mx2) && (x12(k+1,1)>=mx2))
            lower = k;
        end;
        if ((x12(k,1)>=mx2) && (x12(k+1,1)<=mx2))
            upper = k;
        end;
end;
disp('FWHM: ')
(upper-lower)*1.0/lambda
return;
```

Plot Theoretical Single-Slit Pattern

```
function plotTheoreticalSingleSlit(N,a,b,twopi,x,x1);
% Plots the theoretical intenstity pattern for our single slit.
% Function is written by AIV. Version 15. October 2017

%figure;
theta = atan2(([1:N]-(N/2)),b);
betah = (twopi*a/2).*sin(theta);
sinbetah = sin(betah);
theoretical = (sinbetah./betah).*(sinbetah./betah);
x12 = x1(:,1).*x1(:,1);  % Calculate intensities
scaling = max(x12);
plot(x,theoretical.*scaling,'-g');
return;
```

Write Data to File (for Other Purposes Later)

```
function writeToFile(x1);
% Write data to file (as a string of floating point numbers)
% Function is written by AIV. Version 15. October 2017

filename = input('Give the name of new file for storing ...
results: ', 's');

fid = fopen(filename,'w');
fwrite(fid,x1(:,1),'double');
fwrite(fid,x1(:,2),'double');
status = fclose(fid);
return;
```

13.13 Learning Objectives

After working through this chapter, you should be able to:

- Explain the principle of Huygens–Fresnel.
- Derive the condition of constructive interference from a double slit (when the slits are assumed to be very narrow).
- Describe the interference pattern from a double slit, and indicate why the attempt by Thomas Young had a great historical significance.
- Give the main idea of a regular anti-reflection treatment of optics.
- Specify how the interference image changes qualitatively when using more than two parallel identical slits.
- Explain the qualitative intensity distribution in a diffraction pattern for a narrow single slit when we consider the pattern far from the slit.
- Calculate using numerical methods interference pattern also for Fresnel diffraction.
- Specify how the diffraction pattern looks like for light passing through a circular hole.
- Explain how diffraction sets limits on how close two stars can be on heaven before we can no longer distinguish them when we view them through a telescope.
- Calculate the maximum achievable angular resolution for lenses in many different contexts (eye, camera objective, telescope, etc.).
- Know Babinet's principle.
- Know the so-called Arago spot (also called Poisson's spot) and why this phenomenon has a historical significance.

13.14 Exercises

Suggested concepts for student active learning activities: Superposition, Huygens–Fresnel principle, wavefront, coherent, double slit, single slit, optical grating, grating constant, interference pattern, half-width of peaks, interference filter, thin film, diffraction, border region, Huygens–Fresnel diffraction, Fraunhofer diffraction, Airy disc, beam optics, beam waist, Rayleigh's resolution criterion, Arago's spot, X-ray diffraction, Babinet's principle, amplitude level summation.

Comprehension/discussion questions

1. Is it possible to conduct Young's double-slit experiment with sound? Discuss a possible experimental set-up and whether there is a difference between longitudinal and transverse waves in this context.

2. We use the superposition principle "on amplitude level" instead of "on intensity level". Explain why.
3. We have a telescope and want to check if an object we observe is a double star or not. In other words, we need *slightly* greater resolution, and we assume that the telescope has so-called diffraction-limited optics. What do we mean by this expression? Can we increase the resolution by "shuttering down" so we only use a central part of the lens? Or can we increase the resolution by inserting a filter that transmits light either in the blue area or in the red area?
4. In a diffraction experiment with a single slit and light with wavelength λ, there is no intensity minimum. What can we say about the width of the slit?
5. A regular rainbow we get when the drops are above a certain size. For very small drops, the rainbow becomes almost white. How small do you think the drops must be for it to happen?
6. Good speakers (in stereo systems) are often composed of at least one bass speaker (woofer, low frequencies) and a treble speaker (tweeter, high frequencies). The former often has a relatively large diameter, while the latter is usually only a few inches in diameter. Try to give one explanation of this choice based on what you know about diffraction. Also, come with an explanation that is based on a physical mechanism different from diffraction.
7. Why is a diffraction grating (with many slits) better than a double slit if it is to be used in a spectrometer by means of which we can measure wavelengths?
8. Diffraction from a single slit also affects the interference pattern from a grating. Explain the relation.
9. Try to describe the essence of Fig. 13.27. Pay particular attention to the similarities and inequalities between the left and right parts of the figure.
10. Will the interference intensity pattern depend on the diameter of the laser beam when you send a laser beam through an diffraction grating? Explain.

Problems

11. Two coherent sources (always the same phase) for radio waves are located 5.00 m apart, and the waves have a wavelength of 3.00 m. Find points on a line passing through the two sources where we have constructive and destructive interference (if such points exist).
12. Two slits with a mutual distance of 0.450 mm are placed 7.5 m from a screen and illuminated with coherent light with wavelength 500 nm. What distance is there between the second and third dark lines in the interference strips on the screen?
13. An anti-reflection coating on a lens has the refractive index $n = 1.42$ (and that for the glass is 1.52). What is the minimum thickness the coating must have for red light with a wavelength of 650 nm to have minimal reflection?
14. In a Young double-slit experiment, place a piece of glass with refractive index n and thickness L in front of one of the slots. Describe qualitatively what happens to the interference pattern.
15. We use a 10 cm diameter biconvex lens with focal length 50 cm for focusing the sunlight as a "burning glass". The light does not accumulate at one point, but in a disc with a diameter of d. There are two contributions to the size of the disc,

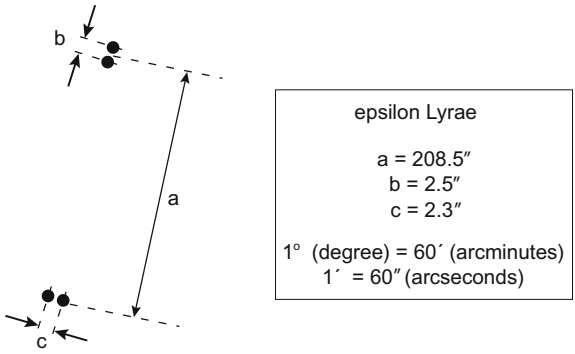

Fig. 13.33 Angular distances between the four stars we perceive with the naked eye as one star, namely Epsilon Lyrae

namely that the sun is imaged by the lens and that the lens causes diffraction. Determine the two contributors to see which one is more important in this case.

16. The "star" Epsilon Lyrae is a double star in the constellation of Lyra, where each of the components is again a double star. The angular distance between the stars is as shown in Fig. 13.33. What demands must we impose on a telescope so that it enables us to distinguish the first pair (observing "two stars"). What requirements must be met to observe all four stars on an evening of calm and clear air?
17. A digital SLR camera has a CMOS chip that is 15.8 × 23.6 mm in size and has 2592 × 3872 pixels. A 35 mm focal length lens is used with f-number 3.3/22 (min/max). What size is the largest and smallest Airy disc from the lens? Enter the answer in both absolute measure and relative to pixel size.
18. We consider the diffraction pattern from a human hair held in the beam of a green laser pen of wavelength 532 nm. It is 16.2 cm between two minimum points with 11 light areas between when the laser pen (hair) is 185 cm from the screen where the measurements were made. How big diameter does the hair have? Is the value you arrive at reasonable based on available information about diameters for human hair?
19. A diffraction grating has its third-order light band at the angle 78.4° for light with wavelength 681 nm. Determine how many lines the grating has per centimetre. Also, determine the angles of the first- and second-order bands. Is there a fourth order of bands?
20. We light with a standard He–Ne laser wavelength 632.8 nm perpendicular to a CD. The "grooves" in a CD are 1.60 μm apart. What are the angles of reflections from the CD?
21. The Hubble Space Telescope has an aperture (opening) of 2.4 m and is used for visible light (400–700 nm). The Arecibo Radio Telescope in Puerto Rico is 305 m in diameter (built in a valley) and is used for radio waves of wavelength 75 cm.

(a) What is the smallest crater size on the moon that can be separated from a neighbouring crater with the two telescopes? (Distance to the moon is about ten times the perimeter of the earth, more specifically 3.84×10^8 m.)
(b) Suppose we want to turn Hubble into a spy satellite that goes into a new orbit around the earth. If we were able to read the number plates for cars with the telescope, what height would the new path to Hubble be?

22. Observe the moon with only the eyes. Try to notice the smallest structure you can distinguish. Find a picture of the moon, and find the structure there. Determine the distance across the structure, and compare it with what you would expect from Rayleigh's resolution criterion.
23. Take a photograph of a distant bright spot with your camera. Analyse the image to see if you can detect the Airy disc. This would require blowing up the image you took until you can see single pixels in the image. Attempt to calculate how large the Airy disc is expected to be.
24. Start with Fig. 13.30. Assume that the distance between the gravestones sideways is a and that they are a distance b behind each other. Determine the angle between each row of gravestones that are behind each other as we see in the photograph. Also, determine the distance between adjacent gravestones along the lines we see (the distance that will match the slit separation distance in a diffraction grating).
25. From a hotel window, interference-like patterns were observed when small light spots were viewed through curtains (made of netting and partially see-through, see Fig. 13.34). Examples of the light phenomenon we observed at night through the curtains are shown in Fig. 13.35). The image does not change if we move closer to or farther away from the curtain while viewing the light coming from outside.

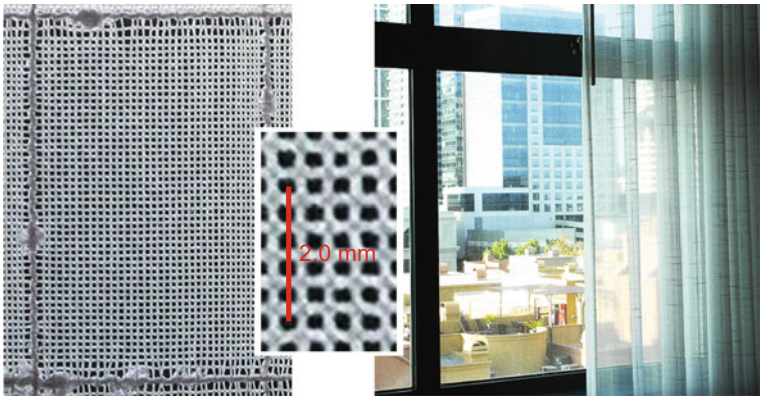

Fig. 13.34 Picture of a light curtain which it was possible to look through. Details show how the fibres in the curtain were in relation to each other. The bar in the middle part is originally 2.0 mm long

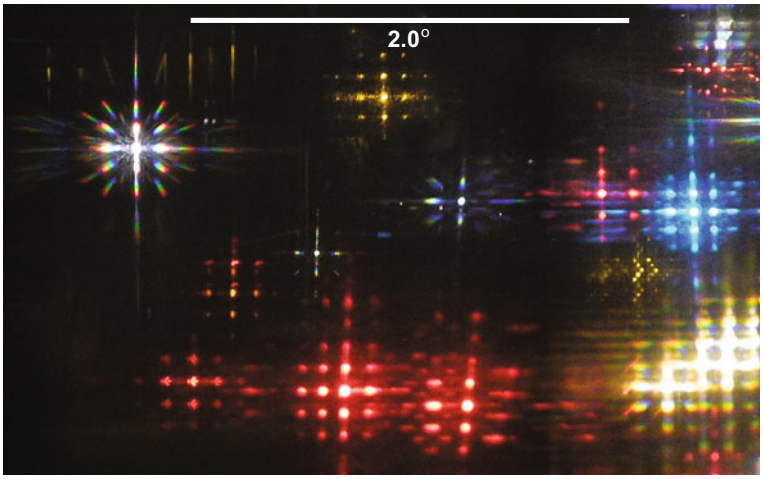

Fig. 13.35 Picture of distant light points observed through the curtain in the previous figure. The bar indicates an angle of 2.0°

(a) Describe which details in the observed light pattern which indicate that diffraction/interference is responsible for what we see.
(b) Carry out calculations that can support such a conclusion (there is probably an estimated 20% uncertainty in the measurements indicated in the figures).