2.12. Wave optics: Diffraction

Building upon our understanding of interference and polarization, which highlighted the wave nature of light, we now turn to another key wave phenomenon: **diffraction**. While interference involves the superposition of waves from multiple sources, diffraction describes the bending of waves as they pass through an aperture or around an obstacle. This bending is a direct consequence of the wave nature of light and becomes particularly noticeable when the size of the aperture or obstacle is comparable to the wavelength of the light.

2.12.1 Revisiting the double-slit experiment: Diffraction

In the previous chapter, we studied the double-slit experiment by using interference between a discrete number of waves, i.e. two waves. Waves "bend", i.e. **diffract**, at a interface which are comparable in size to the wavelength. We described resulting pattern of bright and dark fringes from the double-slit experiment by interference:

- **Constructive interference**, where path differences are integer multiples of the wavelength, leads to bright fringes.
- **Destructive interference**, with half-integer multiples, results in dark fringes.

But, there is a catch we did not consider so far: Simple **interference predicts equally bright fringes, in reality, the intensity of these fringes is modulated**, with the central ones being the brightest, and their intensity decreasing as you move away from the center. This suggests that each slit itself is acting as a source of waves (compare Huygens' principle) that interfere with each other in a more complex way. This phenomenon, responsible for the intensity variations, is called **diffraction**.

Interference vs. diffraction

Interference and diffraction are **fundamentally the same phenomenon**, both stemming from the **superposition of coherent waves**. The distinction often lies in the conceptualization and source arrangement:

- **Interference** typically involves superposition from a few discrete sources, like the two rays.
- **Diffraction** involves the superposition from a continuous distribution of sources or a large number of closely spaced sources.

Think of water waves spreading out after passing through a narrow opening in a barrier. Light behaves similarly, bending around obstacles or spreading out after passing through narrow apertures. Essentially,

diffraction can be seen as the interference of a wave with itself, where each point on a wavefront acts as a source of secondary wavelets (Huygens' principle).

2.12.2 Diffraction at a single-slit

Before mathematically describing the intensity variation of the double-slit experiment, let's start with an easier set up: the **diffraction patterns at a single-slit**.

Consider **monochromatic light from a coherent source** (all waves have same wavelength λ and phase) passing through a single narrow slit of **width** D (D similar to λ). This results in a diffraction pattern on a distant screen, characterized by a central bright maximum flanked by minima and weaker secondary maxima.

The diffraction pattern arises from the interference of waves originating from different parts of the slit. Following the logic from the interference at the double-slit, we can see that the path difference is $\Delta = D \sin\theta$ (assuming screen is far away from slit), and, therefore the relation to the wavelength λ can be described as:

$$D \sin\theta = m\lambda$$

with m as the order. Note that n can be a rational number and, not as in previous section, an integer.

For the zero-th order, we obtain a maximum, i.e. the **central maximum**. **Higher-order maxima** can also be obtained. Not that the intensity of these maxima will be lower than of the central maximum, because we will always have constructive and destructive interference (consider not only two but many rays equidistantly placed across the slit). Let's pick θ such that the **path difference between the outer most rays is** $\frac{3}{2}\lambda$. In that case the rays from the central third will destructively interfere with rays from the upper or lower third for negative and positive angles, respectively (see simulation). Therefore, higher-order maxima occur for $m \approx \pm\frac{3}{2}, \pm\frac{5}{2}, \ldots$ **Minima** occur if all rays destructively interfere. That is the case, if the path difference of the outer most rays is a multiple of the wavelength, i.e. $m = \pm 1, \pm 2, \ldots$.

To summarize:

- **Central maximum:** Rays passing straight through the slit are in phase, creating a central bright region at an angle $\theta = 0$.
- **Minima:** Minima occur at angles θ where the path difference between rays from the top and bottom of the slit is an integer multiple of the wavelength λ:

$$D \sin\theta = m\lambda, \quad m = \pm 1, \pm 2, \ldots$$

- **Higher-order maxima:** Between the minima, weaker maxima appear approximately where the path difference is a half-integer multiple of the wavelength:

$$D \sin\theta \approx (m + \frac{1}{2})\lambda, \quad m \approx \pm\frac{3}{2}, \pm\frac{5}{2}, \ldots$$

```
interactive(children=(FloatSlider(value=0.0, description='Angle (degrees)', max=45.0, min=-4
5.0), FloatSlider(…
```

```
<function __main__.simulate_single_slit_diffraction(diffraction_angle_deg, slit_width, n_rays
=3)>
```

Intensity in single-slit diffraction pattern

Now that we know the position of the minima and maxima, let's consider their intensity, i.e. how bright they appear on the screen.

Similar to the simulation, we assume the slit to be split into N thin strips, each with a thickness of Δy. Remember each Huygens' principle, each point will emit a wavelet. Thus, each thin strips emits light in all direction towards the screen. Again, as in the simulation, we will consider only rays from the strips that are parallel to each other, i.e. have the same angle θ. As we consider the strips to be much thinner than the wavelength λ, **each strip acts as a coherent source** (wave from an individual strip are in phase). As before, there will be a **path difference** $\Delta = \Delta y sin\theta$, which we can convert into a **phase difference** $\Delta\beta$:

$$\Delta\beta = \frac{2\pi}{\lambda}\Delta y \sin\theta$$

Under the assumption that the slit is uniformly illuminated, each strip is associated with is own electric field, each with the amplitude ΔE_0. While the amplitude is the same for each strip, the phase differs between the strips. Thus, the **electric field of each strip** has a magnitude/amplitude and a orientation/phase, i.e. **is a vector**. Therefore, we obtain the **intensity on the screen** as the **vector sum of all strips**.

Let consider the total phase difference β across all slits, i.e. from top to bottom slit (distance $D = N\Delta y$):

$$\beta = N\Delta\beta = \frac{2\pi}{\lambda}N\Delta y \sin\theta = \frac{2\pi}{\lambda}D\sin\theta$$

If the total phase difference $\beta = 2\pi$ all vectors will be oriented evenly from 0 to 2π. Hence, effectively the vectors cancel each other (there is always a pair of vectors with same amplitude but opposite orientation). This gives use the first minima and subsequent minima are found for $\beta = \pm 2\pi, 4\pi, \ldots$.

For higher order maxima, the total phase difference needs to be $\beta = \pm 3\pi, \pm 5\pi, \ldots$ as part of the vectors cancel each other (the portion that would form a circle if we attach the vectors att each others ends). With higher order the fraction of vectors canceling each other increases and, therefore, the intensity of the maxima decreases.

The intensity of this diffraction pattern can be analyzed using the **phasor technique**. For the central maximum ($\theta = 0$), all the phasors are in phase, and the resultant electric field E_0 is simply the sum of the amplitudes of all individual phasors (i.e. the vectors):

$$E_0 = N\Delta E_0$$

Now, consider a general angle θ where there is a phase difference $\Delta\beta$ between consecutive phasors. These N phasors, each with magnitude ΔE_0, will form a circular arc when placed end to end. The total

phase difference across all phasors is $\beta = N\Delta\beta$.

Let's denote the radius of this circular arc as r. The length of the arc is equal to the magnitude of the resultant electric field when all phasors are in phase, which is $E_0 = N\Delta E_0$. The angle subtended by this arc at the center of the circle is the total phase difference β. The relationship between the arc length, radius, and angle is:

$$E_0 = r\beta$$

Dividing both sides by 2, we get:

$$\frac{E_0}{2} = r\frac{\beta}{2}$$

Now, let's consider the resultant electric field E_θ at the angle θ. This is represented by the chord that connects the start and end of the circular arc formed by the phasors. We can find the magnitude of this chord by considering the isosceles triangle formed by the two radii to the ends of the arc and the chord itself. The angle between the two radii is β. We can bisect this triangle with a line that is perpendicular to the chord and passes through the center of the circle. This creates two right-angled triangles. The angle opposite to half the chord ($E_\theta/2$) is $\beta/2$, and the hypotenuse is the radius r. Using trigonometry, we have:

$$\sin\left(\frac{\beta}{2}\right) = \frac{E_\theta/2}{r}$$

Rearranging this equation, we get:

$$\frac{E_\theta}{2} = r\sin\left(\frac{\beta}{2}\right)$$

Now we have two expressions involving r:

1. $\frac{E_0}{2} = r\frac{\beta}{2}$
2. $\frac{E_\theta}{2} = r\sin\left(\frac{\beta}{2}\right)$

We can eliminate r by dividing the second equation by the first equation:

$$\frac{E_\theta/2}{E_0/2} = \frac{r\sin(\beta/2)}{r(\beta/2)}$$

$$\frac{E_\theta}{E_0} = \frac{\sin(\beta/2)}{\beta/2}$$

The intensity of the light is proportional to the square of the amplitude of the electric field. If I_0 is the intensity at the central maximum (proportional to E_0^2) and I_θ is the intensity at an angle θ (proportional to E_θ^2), then:

$$\frac{I_\theta}{I_0} = \left(\frac{E_\theta}{E_0}\right)^2 = \left(\frac{\sin(\beta/2)}{\beta/2}\right)^2$$

Substituting the expression for β:

$$I_\theta = I_0 \left(\frac{\sin\left(\frac{\pi D \sin\theta}{\lambda}\right)}{\frac{\pi D \sin\theta}{\lambda}} \right)^2$$

The function $(\frac{\sin(x)}{x})^2$, often called the sinc squared function, has a central maximum at x = 0 and its amplitude decreases for larger values of x, with zeros occurring at multiples of π. This mathematical form explains the central bright maximum and the decreasing intensity of the secondary maxima in the single-slit diffraction pattern. Minima occur when $\sin(\beta/2) = 0$, leading to $D \sin\theta = m\lambda$ for $m = \pm 1, \pm 2, \pm 3, \ldots$. Higher-order maxima, with much lower intensities, appear roughly halfway between these minima.

2.12.3 Diffraction at a double-slit

In the previous chapter, we used interference to determine where the minima and maxima occur in a double-slit experiment. This is still valid, but in reality the pattern on the screen will not have infinite, equally bright peaks but a finite number of peaks with the brightest peak at the center and lower intensity peaks surrounding it. This is due to diffraction.

Let's use our knowledge about the intensity at a single slit and refine it for the double-slit experiment. Consider a double-slit setup where each slit has a finite width D, and the distance between the centers of the two slits is d. We can think of this as two individual single slits, each contributing to the electric field at a point on the screen.

From our previous discussion, the electric field at an angle θ due to a single slit of width D can be represented by:

$$E_{single} = E_{0,single} \frac{\sin(\beta/2)}{\beta/2}$$

where $E_{0,single}$ is the maximum electric field amplitude for the single slit (at $\theta = 0$), and $\frac{\beta}{2} = \frac{\pi D \sin\theta}{\lambda}$.

Now, in a double-slit experiment, we have two such slits separated by a distance d. Let's assume that the electric field from each slit has the same amplitude as the single-slit case. However, due to the path difference between the light waves from the two slits reaching a point on the screen at an angle θ, there will be a phase difference between their contributions.

The path difference between the waves from the corresponding points in the two slits is approximately $\Delta = d \sin\theta$. This path difference leads to a phase difference δ given by:

$$\delta = \frac{2\pi}{\lambda} \Delta = \frac{2\pi}{\lambda} d \sin\theta$$

Let the electric field from the top slit at the screen be E_1 and the electric field from the bottom slit be E_2. We can represent these as phasors. If we consider the phase at the midpoint between the two slits as a reference, the phase difference for each slit relative to this midpoint will be $\pm\delta/2$. Therefore, the electric fields from the two slits at the screen can be written as:

$$E_1 = E_{single}e^{i\delta/2}$$

$$E_2 = E_{single}e^{-i\delta/2}$$

The total electric field E_{total} at the screen is the superposition of the electric fields from the two slits:

$$E_{total} = E_1 + E_2 = E_{single}e^{i\delta/2} + E_{single}e^{-i\delta/2}$$

$$E_{total} = E_{single}(e^{i\delta/2} + e^{-i\delta/2})$$

Using the trigonometric identity $e^{ix} + e^{-ix} = 2\cos x$, we can rewrite the expression for the total electric field as:

$$E_{total} = 2E_{single}\cos\left(\frac{\delta}{2}\right)$$

Substituting the expression for E_{single}:

$$E_{total} = 2E_{0,single}\frac{\sin(\beta/2)}{\beta/2}\cos\left(\frac{\delta}{2}\right)$$

The intensity of the light is proportional to the square of the amplitude of the total electric field. Let I_0 be the intensity when $\theta = 0$ for a single slit (proportional to $E_{0,single}^2$). Then the intensity at an angle θ for the double slit will be proportional to E_{total}^2:

$$I_\theta \propto (2E_{0,single})^2\left(\frac{\sin(\beta/2)}{\beta/2}\right)^2\cos^2\left(\frac{\delta}{2}\right)$$

We can relate this to the intensity of the central maximum of the double-slit pattern. When $\theta = 0$, $\beta = 0$ and $\delta = 0$, so $\frac{\sin(\beta/2)}{\beta/2} \to 1$ and $\cos(\delta/2) \to 1$. The intensity at the central maximum of the double-slit is four times the intensity of the central maximum of a single slit with the same width (because we have two slits contributing). Let $I_{0,double}$ be the intensity at the central maximum of the double-slit. Then $I_{0,double} \propto (2E_{0,single})^2$. We can write the intensity at any angle θ as:

$$I_\theta = I_{0,double}\left(\frac{\sin(\beta/2)}{\beta/2}\right)^2\cos^2\left(\frac{\delta}{2}\right)$$

If we define I_0 as the intensity of the central maximum of the double-slit pattern, then your formula is correct:

$$I_\theta = I_0\left(\frac{\sin(\beta/2)}{\beta/2}\right)^2\cos^2\left(\frac{\delta}{2}\right)$$

where $\frac{\beta}{2} = \frac{\pi D \sin\theta}{\lambda}$ represents the diffraction effect from each individual slit (the "diffraction factor" or "envelope"), and $\frac{\delta}{2} = \frac{\pi d \sin\theta}{\lambda}$ represents the interference effect due to the path difference between the waves from the two slits (the "interference factor").

The diffraction factor $\left(\frac{\sin(\beta/2)}{\beta/2}\right)^2$ modulates the finer interference fringes given by $\cos^2\left(\frac{\delta}{2}\right)$. This means that the interference maxima will have varying intensities, with the overall intensity pattern being governed by the broader diffraction envelope. The zeros of the diffraction pattern will cause the disappearance of interference fringes at certain angles.

```
interactive(children=(Checkbox(value=True, description='Show Interference Effect'), Checkbox
(value=False, desc…
```

2.12.4 Limits of resolution & circular apertures

Lenses, acting as circular apertures of diameter D, cannot image a point object as a perfect point image. This is due to diffraction and aberration. We will focus here on diffraction only and ignore aberration (discussed in previous chapter). In essence, a lense acts as a slit. Hence, light passing through a lens from a point source forms a diffraction pattern consisting of a central bright circular spot called the Airy disk, surrounded by fainter rings. The *angular half-width* θ of the Airy disk is approximately:

$$\theta \approx 1.22\frac{\lambda}{D}$$

As a consequence, the resolution of a lens, i.e. its ability to distinguish between two closely spaced objects, is limited. The **Rayleigh criterion** states that two point objects are just resolvable when the center of the diffraction pattern of one image aligns with the first minimum of the diffraction pattern of the other. The minimum angular separation θ_{min} between two just-resolvable objects is:

$$\theta_{min} = 1.22\frac{\lambda}{D}$$

A smaller θ_{min} indicates better resolution. This limit also applies to telescopes and mirrors, where D is the diameter of the objective.

The **ultimate limit of resolution** for any optical instrument is approximately half the wavelength of the radiation used: $RP \approx \frac{\lambda}{2}$ (rule of thumb).

2.12.5 Diffraction grating & spectroscopy

A **diffraction grating** consists of a large number of equally spaced parallel slits (separation d). Diffraction gratings typically have a very large number of slits per unit length, often thousands of lines per centimeter or millimeter. It is used for precise wavelength measurements. This precise relationship between the angle of diffraction and the wavelength makes diffraction gratings invaluable tools in

spectroscopy, where they are used to separate and analyze the different wavelengths present in a light source.

Using our knowledge about the double-slit experiment, we see that maxima occur at angles θ given by (m as the order):

$$\sin\theta = \frac{m\lambda}{d}, \quad m = 0, \pm 1, \pm 2, \dots$$

Like, in the double-slit experiment, the central, zero-th order maximum is the brightest, but compared to the double-slit experiment, the grating produces sharper higher-order maxima. This is because if the angle $theta$ is increased even slightly beyond the angle required for a maximum, while the waves from two adjacent slits might only be slightly out of phase, waves from slits that are hundreds of slits apart can become exactly out of phase. This leads to destructive interference across almost all the slits, causing the intensity to drop off rapidly away from the maximum.

Note that the diffraction grating we just described is a so-called **transmission grating**. Another type is the **reflection grating**, which is made by ruling fine lines on a metallic or glass surface. Light is reflected from this surface and then analyzed. The fundamental principles of analysis are the same for both transmission and reflection gratings.

Now let's use **white light instead of monochromatic light**. We will observe a sharp white peak at the center ($m = 0$ order), where all wavelengths interfere constructively at the same angle ($\theta = 0$). However, for all other orders ($m \neq 0$), the different wavelengths present in white light will be diffracted at different angles according to the grating equation:

$$\sin\theta = \frac{m\lambda}{d}$$

As a result, instead of a single sharp peak for each order, we will observe a **spectrum** of colors spread out over a certain angular width. This spreading of light into its component wavelengths is the fundamental principle behind using diffraction gratings in **spectroscopy**. Each order of diffraction will display a distinct spectrum, similar to what is observed with a prism, allowing for the analysis of the wavelengths present in the light source.

A **spectrometer** or **spectroscope** is an instrument designed for the precise measurement of wavelengths of light. It achieves this by separating the different wavelengths present in a light source using a diffraction grating or a prism. Light from the source first enters the spectrometer through a narrow slit in the **collimator**. This slit is positioned at the focal point of a lens, which then produces a parallel beam of light directed towards the diffraction grating or prism. A movable telescope is used to focus the light after it has been separated into its constituent wavelengths by the grating or prism. By carefully positioning the telescope to observe a diffraction peak (typically the first order) corresponding to a specific wavelength emitted by the source, and accurately measuring the angle θ of this peak, the wavelength λ can be calculated using the grating equation:

$$\lambda = \frac{d}{m}\sin\theta$$

where m represents the order of diffraction and d is the spacing between the slits of the grating. The observed bright line for a particular wavelength is actually an image of the entrance slit. While a narrower slit enhances the precision of angular measurements, it also reduces the intensity of the light. When the incoming light contains a continuous range of wavelengths, a **continuous spectrum** is observed in the spectroscope.

Spectrometers commonly use either a transmission grating or a reflection grating to separate light. Alternatively, some spectrometers utilize a prism. Prisms work based on the principle of **dispersion**, where different wavelengths of light are refracted (bent) at different angles. Unlike diffraction gratings, where the relationship between the diffraction angle and wavelength is linear, the dispersion of a prism is non-linear and therefore requires calibration for accurate wavelength determination.

Spectroscopy, the technique of using spectrometers, has a crucial application in identifying the composition of substances at the atomic and molecular level. When a gas is heated or subjected to an electric current, it emits light at specific, discrete wavelengths, forming a unique **line spectrum**. These emitted wavelengths act as a fingerprint for the particular element or compound. Line spectra are characteristic of gases under conditions of high temperature and low pressure or density. In contrast, heated solid objects, like the filament of a lightbulb, and dense gaseous objects such as the Sun, produce a **continuous spectrum** encompassing a broad range of wavelengths.

Interestingly, the continuous spectrum of the Sun is not perfectly uniform but contains a multitude of dark lines known as **absorption lines**. These lines occur because atoms and molecules can absorb light at the very same specific wavelengths that they are capable of emitting. The absorption lines in the Sun's spectrum are primarily caused by the absorption of light by atoms and molecules in the cooler outer layers of the Sun's atmosphere, as well as by components of the Earth's atmosphere. Through meticulous analysis of these absorption lines, scientists have been able to identify the presence of at least two-thirds of all known elements in the Sun. Spectroscopy is also an indispensable tool for determining the elemental composition of the atmospheres of planets, interstellar space, and distant stars.