Lecture notes 5: Diffraction

Let us now consider how light reacts to being confined to a given aperture. The resolution of an aperture is restricted due to the wave nature of light: as light passes through any opening it is *diffracted* and the wave fronts spread out in a shape given by the envelope of Huygens's secondary wavelets which radiate spherically outwards from all points on a wave front. These waves travel along every path, from the source to the point of observation, where they are added together only giving a significant net contribution when they add coherently in phase.
(In the following I will follow Kip Thorne's lecture notes found at
`www.pma.caltech.edu/Courses/ph136/yr2004`.)

Helmholtz-Kirchhoff Integral

Let us restrict attention to the simplest (and, luckily the most useful) case: a monochromatic scalar wave

$$\psi = \psi(\mathbf{x})e^{-i\omega t}$$

with field variable ψ of frequency $\omega = ck$ which satisfies Helmholtz's equation

$$\nabla^2\psi + k^2\psi = 0 \tag{1}$$

except at the boundaries. ψ is generally a real physical value, but for convenience we will use a complex representation. The wave is monochromatic and non-dispersive and the medium is isotropic and homogeneous so that k can be treated as a constant[1]. This formalism is valid for weak sound waves in a fluid and is fairly accurate for the propagation of electromagnetic waves in vacuum or in a medium with a constant dielectric constant ϵ. In the latter case ψ can be considered as one of the Cartesian components of the electric or magnetic field vector, e.g. E_x or B_y. In vacuum Maxwell's equations imply that $\psi = E_x$ satisfies equation 1.

The Helmholtz equation 1 is an elliptic, linear, partial differential equation, and thus permits us to express the value of $\psi_{\mathcal{P}}$ of ψ at any point $\mathcal{P}$ inside some closed surface $\mathcal{S}$ as an integral over $\mathcal{S}$ of some linear combination of ψ and its normal derivative.

Let us derive this expression by augmenting the actual wave ψ in the interior of $\mathcal{S}$ with a second solution[2] of the Helmholtz equation, namely

$$\psi_0 = e^{ikr}/r.$$

[1] Each of these assumptions can be lifted at the cost of introducing technical complications.

[2] Remember that in spherical polar coordinates we write the gradient of a scalar as

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\mathbf{e}_\phi + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta$$

and the divergence as

$$\nabla \cdot A = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 A_r\right) + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi} + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(A_\theta \sin\theta)$$

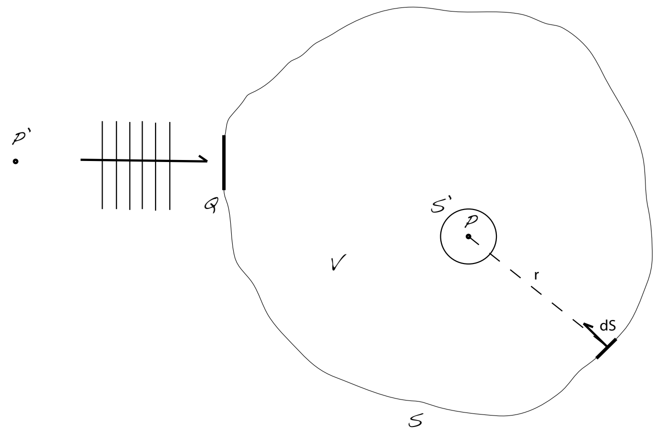

Figure 1: Surface $\mathcal{S}$ for the Helmholtz-Kirchhoff integral. The surface $\mathcal{S}'$ surrounds the observation point $\mathcal{P}$ and $\mathcal{V}$ is the volume bounded by $\mathcal{S}$ and $\mathcal{S}'$. The aperture $\mathcal{Q}$, the incoming wave to the left of it and the point $\mathcal{P}'$ are used later in the text.

This is a spherical wave originating from the point $\mathcal{P}$, and r is the distance from $\mathcal{P}$ to the point where ψ_0 is evaluated. Now apply Gauss' theorem

$$\int_{\mathcal{V}} \nabla \cdot \boldsymbol{F} dV = \int_{\mathcal{S}} \boldsymbol{F} \cdot d\boldsymbol{S}$$

to the vector field $\psi\nabla\psi_0 - \psi_0\nabla\psi$ and invoke Helmholtz equation to arrive at

$$\int_{\mathcal{S}+\mathcal{S}'} (\psi\nabla\psi_0 - \psi_0\nabla\psi) \cdot d\boldsymbol{S} = -\int_{\mathcal{V}} (\psi\nabla^2\psi_0 - \psi_0\nabla^2\psi) dV = 0$$

Where we have introduced a small sphere $\mathcal{S}'$ of radius r' surrounding $\mathcal{P}$; $\mathcal{V}$ is the volume between the two surfaces $\mathcal{S}$ and $\mathcal{S}'$; and we have made the opposite choice of direction for the integration element $d\boldsymbol{S}$ – it points into $\mathcal{V}$ instead of outwardly as is usual, changing the sign of the second expression in the equation above.

Exercise

1. Confirm that the above expression is correct.

Now let the radius r' decrease to zero. We then find that

$$\psi\nabla\psi_0 - \psi_0\nabla\psi \to -\psi(0)/{r'}^2 + O(1/r')$$

and thus the integral over $\mathcal{S}'$ becomes $4\pi\psi(\mathcal{P}) \equiv 4\pi\psi_{\mathcal{P}}$.

Exercise

2. Carry out the integration, and show that its value is $4\pi\psi_{\mathcal{P}}$.

Thus,

$$\psi_{\mathcal{P}} = \frac{1}{4\pi}\int_{\mathcal{S}}\left(\psi\nabla\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\nabla\psi\right)\cdot d\boldsymbol{S}. \tag{2}$$

This equation is the *Helmholtz-Kirchhoff formula* is the expression relating ψ at $\mathcal{P}$ to a linear combination of its value and normal derivative on a surrounding surface. If $\mathcal{P}$ is many wavelengths away from the boundary $\mathcal{S}$, then the integral is only influenced by the waves ψ as they enter through $\mathcal{S}$, and not when they are leaving.

Diffraction by an aperture

Next let us suppose that some aperture $\mathcal{Q}$ of size much greater than a wavelength but much smaller than the distance to $\mathcal{P}$ is illuminated by a distant wave source. Let $\mathcal{S}$ pass through $\mathcal{Q}$, and denote by ψ' the wave incident on $\mathcal{Q}$. Assume that the diffracting aperture has a local and linear effect on ψ': that the wave transmitted through the aperture is given by

$$\psi_{\mathcal{Q}} = t\psi',$$

where t is a complex transmission function that varies over the aperture. In practice, t is usually zero or unity. However, t can also represent a variable phase factor when, for example, the aperture comprises a medium of variable thickness and of different refractive index from that of the homogeneous medium outside the aperture.

Let us now use the Helmholtz-Kirchhoff formula 2 to compute the field at $\mathcal{P}$ due the the wave $\psi_{\mathcal{Q}} = t\psi'$ transmitted through the aperture. The surface $\mathcal{S}$ comprises the aperture $\mathcal{Q}$, a sphere of radius $R \gg r$ centered on $\mathcal{P}$, and the linear extension of the aperture to meet the sphere; and assume that the only incoming waves are those which pass through the aperture.

On the aperture, as $kr \gg 1$, we can write $\nabla(e^{ikr}/r) \simeq -ik\boldsymbol{n}e^{ikr}/r$ where $\boldsymbol{n}$ is a unit vector pointing towards $\mathcal{P}$. Similarly we write $\nabla\psi \simeq ikt\boldsymbol{n'}\psi'$, where $\boldsymbol{n'}$ is a unit vector along the direction of propagation of the incident wave (and where our assumption that anything in the aperture varies on scales long compared to $\lambda = 1/k$ permits us to ignore the gradient of t). Inserting these gradients into equation 2 one obtains

$$\psi_{\mathcal{P}} = -\frac{ik}{2\pi}\int_{\mathcal{Q}} d\boldsymbol{S}\cdot\left(\frac{\boldsymbol{n}+\boldsymbol{n'}}{2}\right)\frac{e^{ikr}}{r}t\psi' \tag{3}$$

This equation can be used to compute the wave from a small aperture at any point $\mathcal{P}$ in the far field. It has the form of an integral transform for the incident field variable ψ', where the integral is over the area of the aperture. The kernel of the transform is the product of several factors: the factor $1/r$ ensures that the flux or energy (proportional to ψ^2) falls off as the inverse square of the distance to the aperture. The phase factor $-ie^{ikr}$ advances the phase of the wave by an amount equal to the optical path length between the element of the aperture and $\mathcal{P}$, minus $\pi/2$. The amplitude and phase of the incoming wave ψ' can also be changed by the transmission function t. Finally there is the *obliquity factor* $d\hat{\boldsymbol{S}} \cdot (\boldsymbol{n} + \boldsymbol{n}')/2$, where $d\hat{\boldsymbol{S}}$ is the unit vector normal to the aperture.

Spreading of the Wavefront

Equation 3 gives a general description for computing the diffraction pattern from an illuminated aperture. It is commonly used in two different limits, called *Fraunhofer* and *Fresnel.*

Suppose that the aperture has a linear size a and is roughly centered on the geometric ray from the the source point $\mathcal{P}'$ to the field point $\mathcal{P}$. Consider the variations of the phase ϕ of the contributions to $\psi_{\mathcal{P}}$ that come from various places in the aperture. Using trigonometry we can estimate that locations on opposite sides of the aperture produce phases at $\mathcal{P}$ that differ by $\Delta\phi = k(r_2 - r_1) \sim ka^2/2r$, where r_1 and r_2 are the distances from the two edges of the aperture to the point $\mathcal{P}$.

Exercise

3. Why can we write $(r_2 - r_1) \sim a^2/2r$? Draw a figure to show the geometry of the problem.

There are two limiting regimes depending on whether the aperture is large or small compared with the *Fresnel length*

$$r_F \equiv \left(\frac{2\pi r}{k}\right)^{1/2} = (\lambda r)^{1/2}$$

When $a \ll r_F$, the phase variation $\Delta \sim a^2/r_F^2$ is $\ll \pi$ and can be ignored; the contributions from different parts of the aperture are essentially in phase with each other – this is the *Fraunhofer* regime. When $a \gg r_F$, $\Delta\phi \gg \pi$ and the phase variation is very important in determining the observed intensity $\mid \psi_{\mathcal{P}} \mid^2$ – this is the *Fresnel* regime.

Consider a planar wave propagating perpendicular to an aperture of size a. Wave optics insists that the transverse localization of the wave into a region of size $\Delta x \sim a$ must produce a spread in its transverse wave vector, $\Delta k_x \sim 1/a$ (a momentum of uncertainty[3] $\Delta p_x = \hbar \Delta k_x \sim \hbar/a$.) This uncertain transverse vector produces, after propagating a distance r, a corresponding uncertainty $(\Delta k_x/k)r \approx r_F^2/a$ in the beam's transverse size; and this uncertainty superposes incoherently on the aperture-induced size a to produce a net transverse beam

[3] In this way we can achieve a quick and dirty "derivation" of the Rayleigh criterion using photons with momentum p impinging on a lens of diameter a which gives that the resolution $(\theta = \Delta p_x/p)$ may be found as

$$\theta = \Delta p_x/p = \frac{h/a}{h/\lambda} = \lambda/a.$$

size

$$\begin{aligned}\Delta x &\sim \sqrt{a^2 + (r_F^2/a)^2} \\ &\sim a \qquad \text{if} \quad r \ll a^2/\lambda \quad \text{Fresnel regime} \\ &\sim \left(\frac{\lambda}{a}\right) r \qquad \text{if} \quad r \gg a^2/\lambda \quad \text{Fraunhofer regime.}\end{aligned}$$

In the nearby Fresnel regime the aperture creates a beam whose edges will have the same shape and size as the aperture itself, and will be reasonably sharp but with some oscillatory blurring associated with wave-packet spreading. By contrast in the more distant Fraunhofer regime wave front spreading will cause the transverse size of the entire wave to grow linearly with the distance, and the intensity pattern will typically not resemble the aperture at all.

Fraunhofer Diffraction

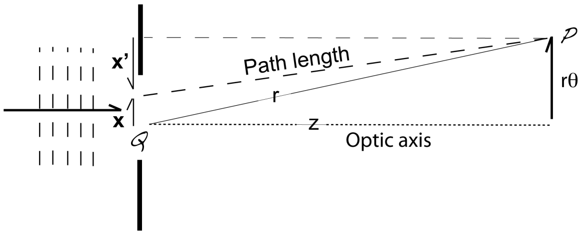

Figure 2: Geometry for computing the path length between a point $\mathcal{Q}$ in the aperture and the observation point $\mathcal{P}$. The transverse vector $\mathbf{x}$ is used to identify $\mathcal{Q}$ in the Fraunhofer analysis and in a later lecture $\mathbf{x}'$ is used for Fresnel analysis.

Consider now the Fraunehofer regime and specialize to the case of an incident plane wave with wave vector $\boldsymbol{k}$ orthogonal to the aperture plane. Regard the line along k through the center of the aperture $\mathcal{Q}$ as the "optic axis"; identify points in the aperture by their two-dimensional vectorial separation $\boldsymbol{x}$ from that axis; identify $\mathcal{P}$ by its distance r (note change in definition of r) from the aperture center and its 2-dimensional transverse separation $r\boldsymbol{\theta}$ from the optic axis. Now, restrict attention to small angle diffraction $|\boldsymbol{\theta}| \ll 1$. The geometric path length between $\mathcal{P}$ and a point $\boldsymbol{x}$ on $\mathcal{Q}$ can be expanded as

$$\text{Path length} = (r^2 - 2r\boldsymbol{x}\cdot\boldsymbol{\theta} + x^2)^{1/2} \simeq r - \boldsymbol{x}\cdot\boldsymbol{\theta} + \frac{x^2}{2r} + \ldots$$

The first term in this expression, r, just contributes an $\boldsymbol{x}$-independent phase e^{ikr} to $\psi_{\mathcal{P}}$. The third term, $x^2/2r$, contributes a phase variation that is $\ll 1$ here in the Fraunenhofer region (but will be important in the Fresnel region). Therefore, in the Fraunhofer region, we can retain just the second term $-\boldsymbol{x}\cdot\boldsymbol{\theta}$

and write equation 3

$$\psi_{\mathcal{P}}(\boldsymbol{\theta}) \propto \int e^{-ik\boldsymbol{x}\cdot\boldsymbol{\theta}} t(\boldsymbol{x}) d^2x \equiv \bar{t}(\boldsymbol{\theta})$$

Where d^2x is the surface area element in the aperture plane and we have dropped a constant phase factor and constant multiplicative factors. Thus, $\psi_{\mathcal{P}}(\boldsymbol{\theta})$ in the Fraunhofer regime is given by the two dimensional Fourier transform denoted by $\bar{t}(\boldsymbol{\theta})$, of the transmission function $t(\boldsymbol{x})$, with $\boldsymbol{x}$ made dimensionless in the transform by multiplying with $k = 2\pi/\lambda$.

Diffraction of a single slit

Consider now a single transparent stripe, a slit, of width a centered on $x = 0$, and measure the scalar angle θ from the direction of the incident radiation. This slit has the transmission function

$$\begin{aligned} t_1(x) &= 1 \quad |x| < a/2, \\ &= 0 \quad |x| > a/2. \end{aligned}$$

Its diffraction pattern is

$$\begin{aligned} \psi_{\mathcal{P}}(\boldsymbol{\theta}) &\propto \bar{t}_1 \\ &\propto \int_{-a/2}^{a/2} e^{ikx\theta} dx \\ &\propto \operatorname{sinc}\left(\frac{ka\theta}{2}\right), \end{aligned}$$

where $\operatorname{sinc}(x) \equiv \sin(x)/x$.

Exercise

4. Carry out the intergration $\int_{-a/2}^{a/2} e^{ikx\theta} dx$ above, and show that the result is correct.

5. The intensity of radiation is proportional to the square of ψ. Plot, for example with IDL, the intensity pattern from the diffraction of a single slit in the Frauhofer regime.

Babinet's principle

In the previous section we have shown how to compute the Fraunhofer diffraction pattern formed by a narrow slit. We might also be interested in the pattern produced by the complementary aperture, *i.e.* a needle of width and length the same as the slit. We can derive the needles pattern by observing that the sum of the waves from the two apertures should equal the wave from a completely unaltered incident wave front. That is to say if we exclude the direction of the incident wave, the field amplitude diffracted by the two apertures are the negative of each other, and hence the intensities $|\psi|^2$ are the same. Therefore the Fraunhofer diffraction patterns from the needle and the slit — and indeed from any complementary apertures — are identical, except in the direction of the incident wave.

Diffraction by a circular aperture

Let us now compute how well a telescope can distinguish neighboring stars. We cannot expect to resolve them (or any two objects) that are closer together in the sky than the angular width of the diffraction pattern formed by the telescope's aperture. Of course, optical imperfections and pointing errors in a real telescope may degrade the image quality even further, but this is the best that can be done, limited only by the uncertainty principle.

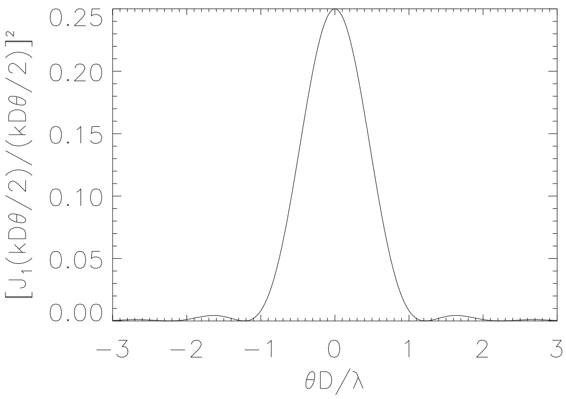

Figure 3: Airy diffraction pattern for a circular aperture. Note the first zero of the function at $\theta D/\lambda = 1.22$.

The calculation is straightforward using equation 3 and assuming a circular aperture telescope with diameter D:

$$\begin{aligned} \psi(\theta) &\propto \int_{\pi D} e^{-ik\boldsymbol{x}\cdot\boldsymbol{\theta}} d^2x \\ &\propto \text{jinc}\left(\frac{kD\theta}{2}\right) \end{aligned}$$

where $\text{jinc}(x) \equiv J_1(x)/x$ with J_1 the Bessel function of order one. The flux from an star observed at angle θ is therefore $\propto \text{jinc}^2(kD\theta/2)$. This intensity pattern is known as the *Airy pattern.* There is a central "Airy disk" surrounded by a circle where the flux vanishes, and then further surrounded by a series of concentric rings whose flux vanishes with radius. Only 16% of the tatal light falls outside the the central Airy disk. The angular radius θ_A of the Airy disk, *i.e.* the radius of the dark circle surrounding it, is determined by the first zero of $J_1(kD\theta/2)$ which is found to be $\theta_A = 1.22\lambda/D$.

Exercise

6. Write a `IDL` routine that plots a cut through the Airy disk such as figure 3 and, in addition, plots an image of the Airy disk.