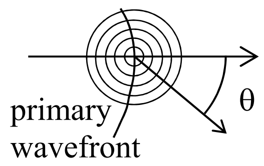
3.1 Fraunhofer Diffraction and the Fourier Transform

[K&F 6.1; S&K 7.3-5; PPP 11]

We will briefly review results from Second Year optics (PHYS2125).

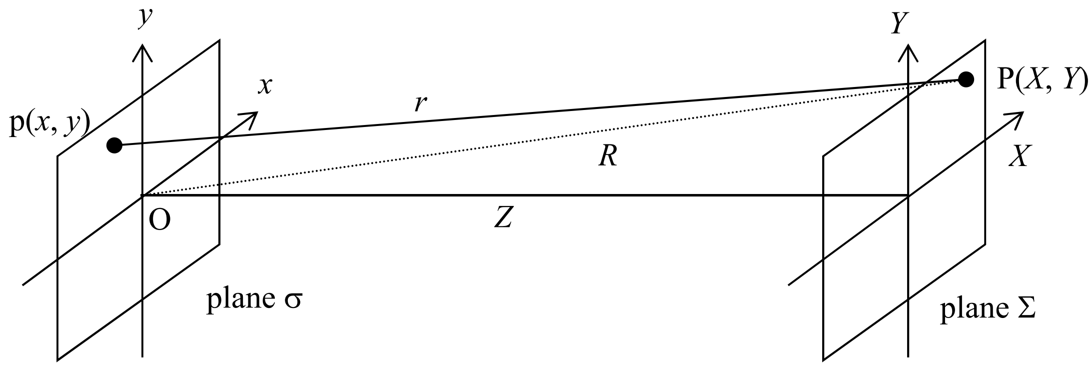
According to the ***Huygens-Fresnel Principle***, we can treat the propagation of light from one point to another in space by assuming that every point on an unobstructed wavefront is a source of secondary spherical wavelets of the same frequency as the primary wave. The amplitude of the optical field at some point beyond is the superposition of all these wavelets, including their amplitudes and phase.

To avoid inconsistencies, we also need to introduce an ***inclination factor***. The amplitude of the secondary spherical wavelets is proportional to a factor $K(\theta)$, which has the properties:



$K(0) = 1$
$K(\theta) \to 0$ as $\theta \to \pi$

It can be shown mathematically that these principles can be derived from the full Maxwell Equations, with certain approximations. The advantage of them is that they are much easier to use than the full equations, and give a better insight into the physics of diffraction.



These principles, and the geometry of the figure above, lead to the ***Diffraction Integral***:

$$\boxed{E_{\mathrm{P}}(X,Y) = C\,e^{i\omega t}\iint_{\sigma} \frac{A(x,y)K(\theta)e^{-ikr}}{r}\,dx\,dy}$$

where $E_{\mathrm{P}}(X, Y)$ is the field in plane Σ due to the field distribution $A(x, y)$ in plane σ, and C is a constant yet to be determined. The distance r is given by:

$$r = \sqrt{(X-x)^2 + (Y-y)^2 + Z^2}$$

The Diffraction Integral is not easy to evaluate in general, because of the way that r varies with x and y , so we look for further approximations. We can investigate the situation experimentally by producing some distribution of coherent light over a small area on plane σ (say, by shining laser light through an aperture), and by observing the resulting *diffraction* *pattern* on another plane Σ some distance away. What we find is that when Z is a small distance, the light pattern on Σ can change radically for small changes in Z, but as Z becomes large, the diffraction pattern settles down to a more or less stable distribution and only spreads out in size as Z increases further. Apparently, things become simpler at large Z. We therefore distinguish between **Fraunhofer** (or **far-field**) diffraction, and **Fresnel** (or **near-field**) diffraction.

When Z is large compared with the size of the aperture, we can substitute:

$$\boxed{r \approx R - \frac{xX + yY}{R}} \qquad \text{the } \textbf{Fraunhofer Approximation}$$

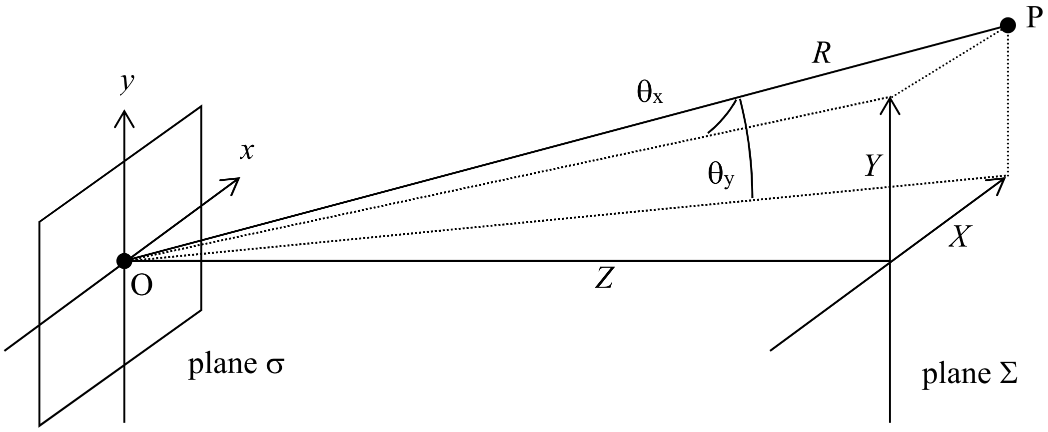
in the exponential in the numerator of the integral, and $r \approx R$ in the denominator. This will be a good approximation provided that

$$\boxed{x^2 << \lambda R; \quad y^2 << \lambda R} \qquad \text{the } \textbf{Fraunhofer Condition}.$$

We also put $K(\theta) = 1$. Then for convenience we introduce new variables to replace X and Y:

$$\boxed{\text{Let} \quad u = \frac{X}{\lambda R} = \frac{1}{\lambda}\sin\theta_x \; ; \quad v = \frac{Y}{\lambda R} = \frac{1}{\lambda}\sin\theta_y}$$

where the angles θ_x and θ_y are shown in the diagram.



Also for convenience we *normalise* the amplitude in plane σ by putting $A(x, y) = A\,\tau(x, y)$, where A is now a *constant* field amplitude and $\tau(x, y)$ is a dimensionless *transmission function* describing what is happening in plane σ.

With these definitions, we can now write:

$$\boxed{\begin{gathered} E_P(X,Y) = \frac{C\,A\exp\{i(\omega t - kR)\}}{R}T(u,v) \\ \text{where } T(u,v) = \int_{-\infty}^{\infty}\!\!\int \tau(x,y)\exp\{i2\pi(xu + yv)\}\,dx\,dy \end{gathered}}$$

This equation defines $T(u, v)$ as the *two-dimensional Fourier Transform* of the function $\tau(x, y)$. Accordingly,

3. DIFFRACTION

The electric field in the Fraunhofer diffraction pattern plane is proportional to the Fourier Transform of the electric field in the object plane.

Note that both $\tau(x, y)$ and $T(u, v)$ are generally complex-valued functions, expressing the magnitude and phase of the electric field at each point.

What a detector "sees" in the diffraction plane is the *power density*, not the field. The power density is:

$$\boxed{S_{\mathrm{P}} = \frac{1}{2} nc\varepsilon_0 \, E_{\mathrm{P}}^* \, E_{\mathrm{P}} = \frac{1}{2} nc\varepsilon_0 \frac{(CA)^2}{R^2} T^* T}$$

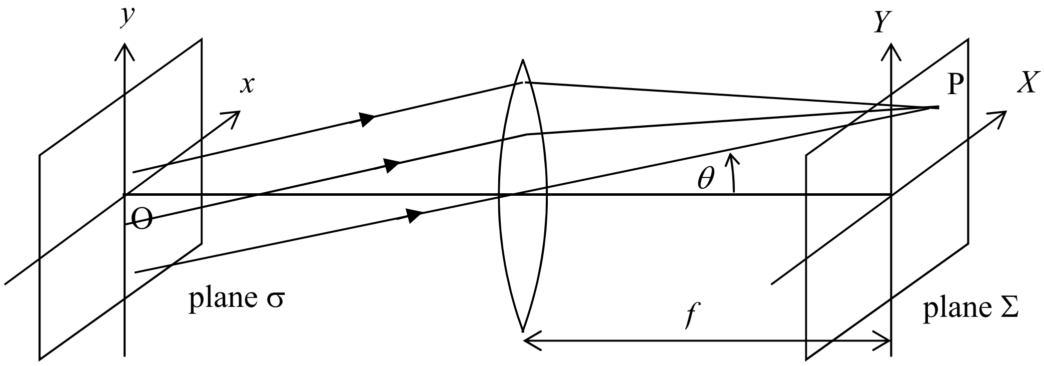
The Fraunhofer approximation uses a *linear* expression for the dependence of the phase on coordinates (x, y) and (X, Y). The phase difference relative to the path R is $\frac{2\pi}{\lambda R}(xX + yY)$.

We can express this fact alternatively as:

- the waves arriving at Σ from a particular point on σ are *plane* waves: Σ and σ are far enough apart that the spherical wavefronts from σ are near enough to flat over the area of Σ. Their phase then varies linearly with X and Y.

- the rays arriving at a point P on Σ all left σ in *parallel directions*: they form a plane wave component of the field on σ, since their phase varies linearly with x and y.

Rather than putting a screen a long way away from the object plane, it is possible to use a *lens* to satisfy the Fraunhofer condition. The plane Σ is found one focal length behind a converging lens.



Now we have $X \approx f\theta_x$ and $Y \approx f\theta_y$, so

$$\boxed{u = \frac{1}{\lambda}\sin\theta_x \approx \frac{X}{\lambda f}; \quad v = \frac{1}{\lambda}\sin\theta_y \approx \frac{Y}{\lambda f}}$$

We simply use f in place of R. The Fraunhofer approximation now holds *exactly*, since the focal plane is an image of a plane "at infinity".

Example 1: Rectangular aperture

A rectangular aperture with a width *a* and a height *b* has a Fourier Transform given by:

$$\boxed{\begin{aligned} &T(u,v) = ab\ \mathrm{sinc}(\pi au)\,\mathrm{sinc}(\pi bv) \\ \text{and}\quad &S(u,v) = \frac{1}{2} nc\varepsilon_0 \frac{(CA)^2}{R^2}(ab)^2\ \mathrm{sinc}^2(\pi au)\,\mathrm{sinc}^2(\pi bv) \end{aligned}}$$

where we define $\boxed{\mathrm{sinc}(w) \overset{\text{def}}{=} \frac{\sin(w)}{w},\quad \text{where } w \text{ is in } \textbf{radians}}$

Example 2: Circular aperture

A circular aperture of radius *a* has a Fourier Transform of:

$$\boxed{\begin{aligned} &T(P) = \pi\, a^2 \left[\frac{2J_1(2\pi\, aP)}{2\pi\, aP}\right] \\ &\text{where}\quad P = \sqrt{u^2+v^2} = \frac{\sqrt{X^2+Y^2}}{\lambda\, R} = \frac{L}{\lambda\, R} \end{aligned}}$$

where *L* is the radius measured from the centre of the diffraction pattern.

The power density in the diffraction pattern is:

$$S(P) = \frac{1}{2} n\varepsilon_0 c \frac{(CA)^2}{R^2}\left(\pi\, a^2\right)^2 \left[\frac{2J_1(2\pi\, aP)}{2\pi\, aP}\right]^2$$

See Second Year notes for derivations and details.

Exercise Set 3.1

1. Calculate the value of sinc(w) for values of w = 1, 2, 3, 4 and 5.

2. At what value of w does sinc(w) = 0.5? At what value does $\mathrm{sinc}^2(w) = 0.5$?

3. A square aperture with a side of length 0.5 mm is illuminated with light of wavelength 550 nm. At what distance from the aperture would the Fraunhofer diffraction pattern have a central maximum with a width also equal to 0.5 mm? What can you say about the Fraunhofer condition under these circumstances?

3.2 Applications of the Fourier Transform to Diffraction Problems

To calculate the diffraction patterns of more complicated apertures, we can often simplify the maths by using some of the properties of Fourier Transforms.

For simplicity, we consider 1-D transforms here. The results apply to higher dimensions as well.

1. Definition and inverse transform

Let $F(u)=\Im[f(x)]=\int_{-\infty}^{\infty} f(x)\exp(i2\pi\, ux)dx$; in brief, $f(x)\overset{FT}{\Rightarrow}F(u)$

Then $f(x)=\Im^{-1}[F(u)]=\int_{-\infty}^{\infty} F(u)\exp(-i2\pi\, ux)du$; in brief, $f(x)\underset{IFT}{\Leftarrow}F(u)$

provided that $f(x)$ is a "sensible" function.

This means that the inverse transform is the same as the forward transform, except for a change in sign from $-i$ to $+i$.
(NB: the integral with $-i$ is usually called the "forward" transform, and the one with $+i$ is called the "inverse" transform. It's only a matter of convention).

2. Symmetry – change of sign of coordinate

What is the Fourier transform of $f(-x)$, in terms of $F(u)$ as defined above?

$$\Im[f(-x)]=\int_{-\infty}^{\infty} f(-x)\exp(i2\pi\, ux)dx=\int_{-\infty}^{\infty} f(-x)\exp(i2\pi\,(-u)(-x))dx$$

$$=\int_{\infty}^{-\infty} f(x')\exp(i2\pi(-u)x')(-dx') \quad \text{putting } x'=-x$$

$$=\int_{-\infty}^{\infty} f(x')\exp(i2\pi(-u)x')dx'=F(-u) \qquad \text{i.e.,}\quad f(-x)\overset{FT}{\Rightarrow}F(-u)$$

This result is not surprising if we consider the physical picture of the creation of a diffraction pattern with a lens. If we reverse the object, we reverse the diffraction pattern. It also implies that if f is an even or odd function, the corresponding transform F will be even or odd respectively also.

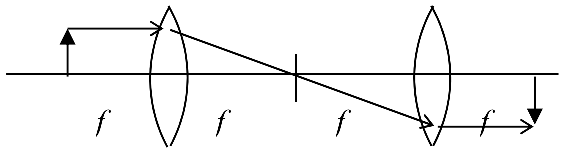
Note that if we substitute $x=-x$ in the definition of the inverse transform, we get:

$$f(-x)=\int_{-\infty}^{\infty} F(u)\exp(i2\pi\, ux)du \quad ; \text{ i.e, } \quad F(u)\overset{FT}{\Rightarrow}f(-x)$$

Combining these results, we have: $\boxed{f(x)\underset{\underset{IFT}{\Leftarrow}}{\overset{FT}{\Rightarrow}}F(u)\underset{\underset{IFT}{\Leftarrow}}{\overset{FT}{\Rightarrow}}f(-x)\underset{\underset{IFT}{\Leftarrow}}{\overset{FT}{\Rightarrow}}F(-u)\underset{\underset{IFT}{\Leftarrow}}{\overset{FT}{\Rightarrow}}f(x)\underset{\underset{IFT}{\Leftarrow}}{\overset{FT}{\Rightarrow}}\cdots}$

That is, two successive *forward* transforms on a function give back the original function, but with the direction of the x axis reversed. Two further forward transforms give the original function again.

This implies that if we use a lens system to form the Fraunhofer diffraction pattern of an original object, then use another lens to create the Fraunhofer diffraction pattern of that diffraction pattern, we get an image of the original object, but with a reversal of the direction in both x and y coordinates.



This can be understood from geometric optics. Two such lenses in series will produce an image of the original object, but the image will be inverted. In fact, in any optical system which forms an image of an object, there will be a plane somewhere between the two which corresponds to the Fraunhofer diffraction plane.

3. Symmetry – complex conjugate

Take the complex conjugate of the definition of the inverse transform:

$$f^*(x) = \left[\int_{-\infty}^{\infty} F(u)\exp(-i2\pi\, ux)du\right]^* = \int_{-\infty}^{\infty} F^*(u)\exp(+i2\pi\, ux)du \quad ; \text{ i.e., } \quad F^*(u) \overset{FT}{\Rightarrow} f^*(x)$$

Hence from previous results, $f^*(x) \overset{FT}{\Rightarrow} F^*(-u)$

A purely real function f obeys $f(x) = f^*(x)$. Then $F(u) = F^*(-u)$, meaning that the negative half of the function is the complex conjugate of the positive half. If, in addition, $f(x)$ is an even function, then we must have $F(u) = F^*(u)$, implying that F is purely real also.

The application is that if the phase of the electric field across the object plane is uniform, then the field across the diffraction pattern shows conjugate symmetry about the centre. The power density will be symmetrical about the centre, since it depends only on the magnitude of the field. If the object is also symmetrical about the centre, the diffraction pattern will be uniform in phase.

4. Linearity

$$a_1 f_1(x) + a_2 f_2(x) \overset{FT}{\Rightarrow} a_1 F_1(u) + a_2 F_2(u)$$

The diffraction pattern of the sum of two fields is the sum of the individual diffraction patterns. Note carefully that this rule applies to the *fields* which includes the amplitudes and phases (complex numbers). The power densities will *not* add linearly, but will show interference effects.

5. Scaling

$$f(ax) \overset{FT}{\Rightarrow} \frac{1}{|a|} F\left(\frac{u}{a}\right)$$

If the x axis of $f(x)$ is compressed by some scale factor a, then its diffraction pattern is expanded by the same factor a, and its amplitude is reduced (keeping the area under the function constant). This is the familiar reciprocal relationship between the size of an object and the size of its diffraction pattern.

6. Centre value

$$F(0) = \int_{-\infty}^{\infty} f(x)\, dx = \text{area under } f(x)$$

The field in the centre of the diffraction pattern is the integral of the field over the diffracting aperture, with no extra phase factors. If the amplitude of the field is uniform over the aperture, then if its area is doubled, the amplitude of the field in the centre of the diffraction pattern is doubled.

7. Shift

$$f(x - x_0) \overset{FT}{\Rightarrow} \exp(i2\pi\, x_0 u) F(u)$$

Shifting the aperture sideways along the x axis changes the phase of the diffraction pattern's field, but not its magnitude. Hence the power density does not change, and the diffraction pattern does not shift.

8. Convolution

This is a very important and powerful property. First we need to define what convolution is.

The convolution product of two functions $f(x)$ and $g(x)$ is a third function $h(x)$ defined by:

$$\boxed{h(x) = f(x) * g(x) \overset{\text{defn.}}{=} \int_{-\infty}^{\infty} f(x - x')\, g(x')\, dx'}$$

Convolution is a repetition of one function along the x axis with a distribution given by the other – or a "smearing out" of one function with the shape of the other.

The process can be thought of as follows (see diagram next page):

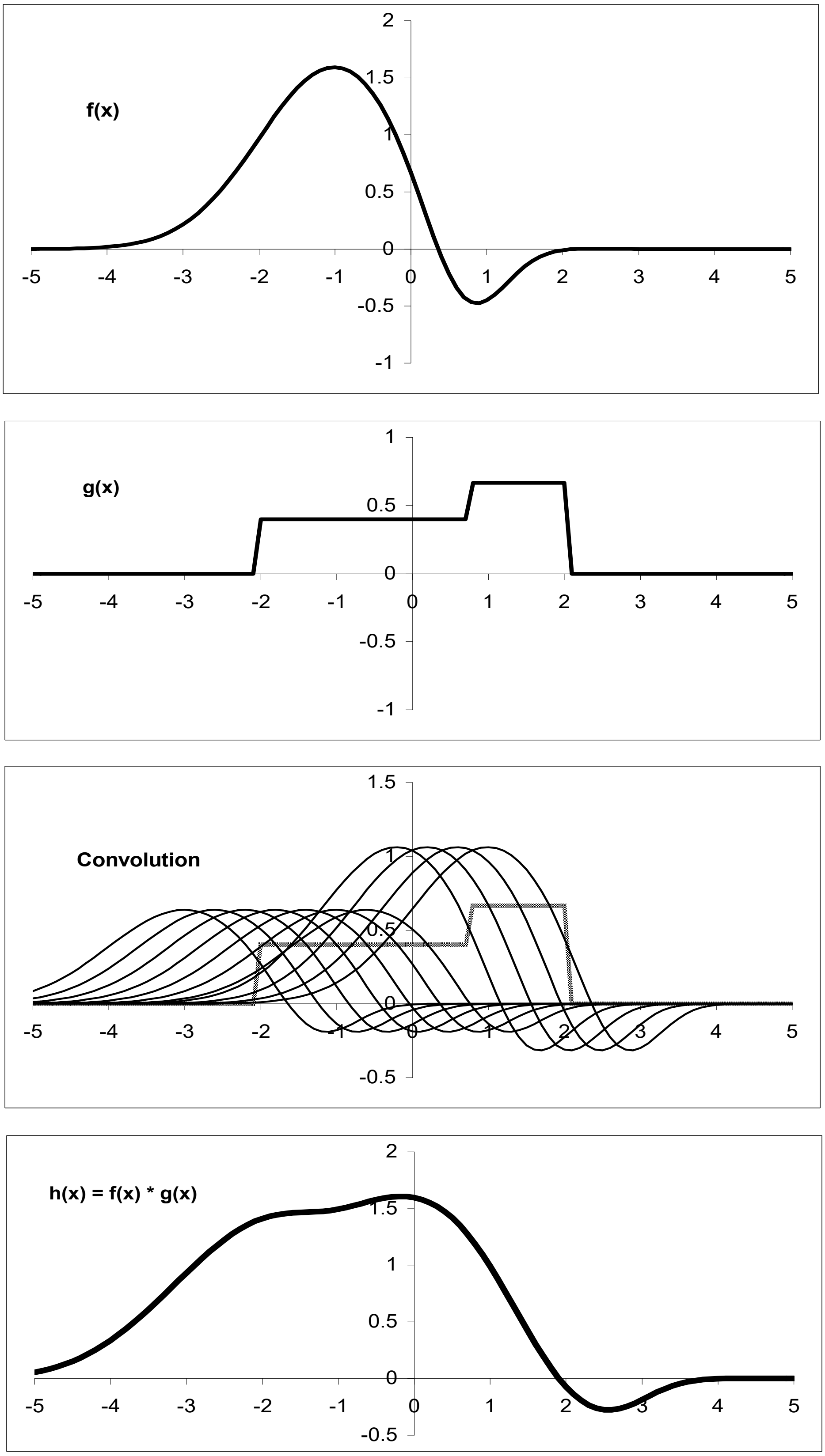
- Divide one of the functions, say $g(x')$, into many small intervals dx'.
- For each position x', shift the origin of $f(x)$ to coincide with x', giving $f(x - x')$.
- Multiply $f(x- x')$ by the quantity $g(x')dx'$.
- Add together all the shifted and scaled functions so formed.

This process actually occurs naturally in many places in physics: e.g. recording a spectrum using a spectrometer with a finite bandwidth, image blurring by an imperfect or out-of-focus lens.

The convolution product is symmetrical, in that $f(x) * g(x) = g(x) * f(x)$.

The process of convolution is illustrated with some arbitrary functions $f(x)$ and $g(x)$ on the next page.

Illustration of Convolution



The connection between convolution and the Fourier Transform is the following:

If $f(x) \overset{FT}{\Rightarrow} F(u)$ and $g(x) \overset{FT}{\Rightarrow} G(u)$ then $f(x) * g(x) \overset{FT}{\Rightarrow} F(u) \times G(u)$ and $f(x) \times g(x) \overset{FT}{\Rightarrow} F(u) * G(u)$	**The Convolution Theorem**

When two functions are *convolved* together, their Fourier Transforms are *multiplied* together, and vice versa – convolution and multiplication are *Fourier conjugate operations.* The proof follows in a straightforward way from the definitions of the Fourier Transform and the convolution operation.

Note that multiplication is much easier to calculate than convolution. In computer calculations of convolutions, particularly for large arrays of data, the standard procedure is to calculate the Fourier transforms of the functions, multiply them together, then inverse transform the result. This turns out to be more computationally efficient than calculating the convolution directly, mainly due to the existence of a very efficient method of calculating the Fourier transform, known as the Fast Fourier Transform algorithm.

We now apply some of these theorems to deduce some more Fourier Transforms/diffraction patterns.

A. Point Source

Mathematically, a point source at the origin $x = y = 0$ can be represented by a $\tau(x, y)$ which is zero everywhere except at the origin; that is, a *delta function* $\delta(0, 0)$.

Its Fourier Transform is $$T(u,v) = \iint \delta(0,0)\exp\{i2\pi(xu + yv)\}\, dx\, dy$$
$$= \exp\{i2\pi(0 + 0)\} = 1$$

This represents an electric field with a uniform amplitude and phase across the diffraction plane: i.e., a plane wave arriving at normal incidence to the *XY* plane.

If the point source is shifted to the point $x = x_0$, $y = 0$, then by the shift theorem the transform becomes:

$$\exp\{i\, 2\pi\, x_0\, u\}$$

This represents a plane wave with wavefronts inclined to the *XY* plane, so that its phase varies linearly across the plane. Note that the spacing Δu corresponding to one cycle of phase is just

$$\Delta u = 1/x_0.$$

B. Two point sources

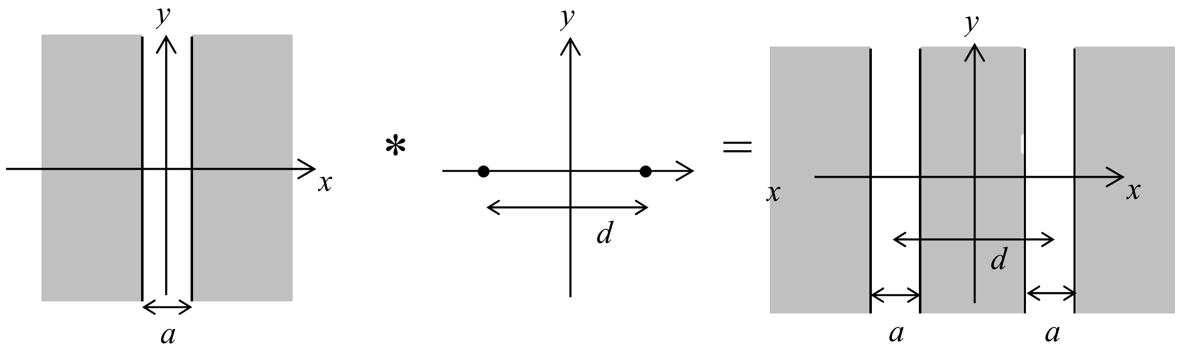
Two point sources with a separation d can be represented by a pair of delta functions $\delta(x - d/2, y)$ and $\delta(x + d/2, y)$. Using shift and linearity properties, the transform is:

$$T(u, v) = \exp\left(+i2\pi u\frac{d}{2}\right) + \exp\left(-i2\pi\frac{d}{2}\right)$$
$$= 2\cos(\pi du)$$

This is just Young's fringes, formed by interference between two plane waves. Maximum power occurs when $\pi du = m\pi$; so $u = -X/(\lambda R) = m/d$. The fringe spacing is $1/d$ in terms of u, or $(\lambda R)/d$ in terms of X.

C. Double Slit

Mathematically, we can treat the double slit aperture as the *convolution* of a single slit with a pair of point sources (delta functions).



Hence its diffraction pattern is the *multiplication* of the diffraction patterns of a single slit (sinc function) and of two point sources (Young's cosine fringes).

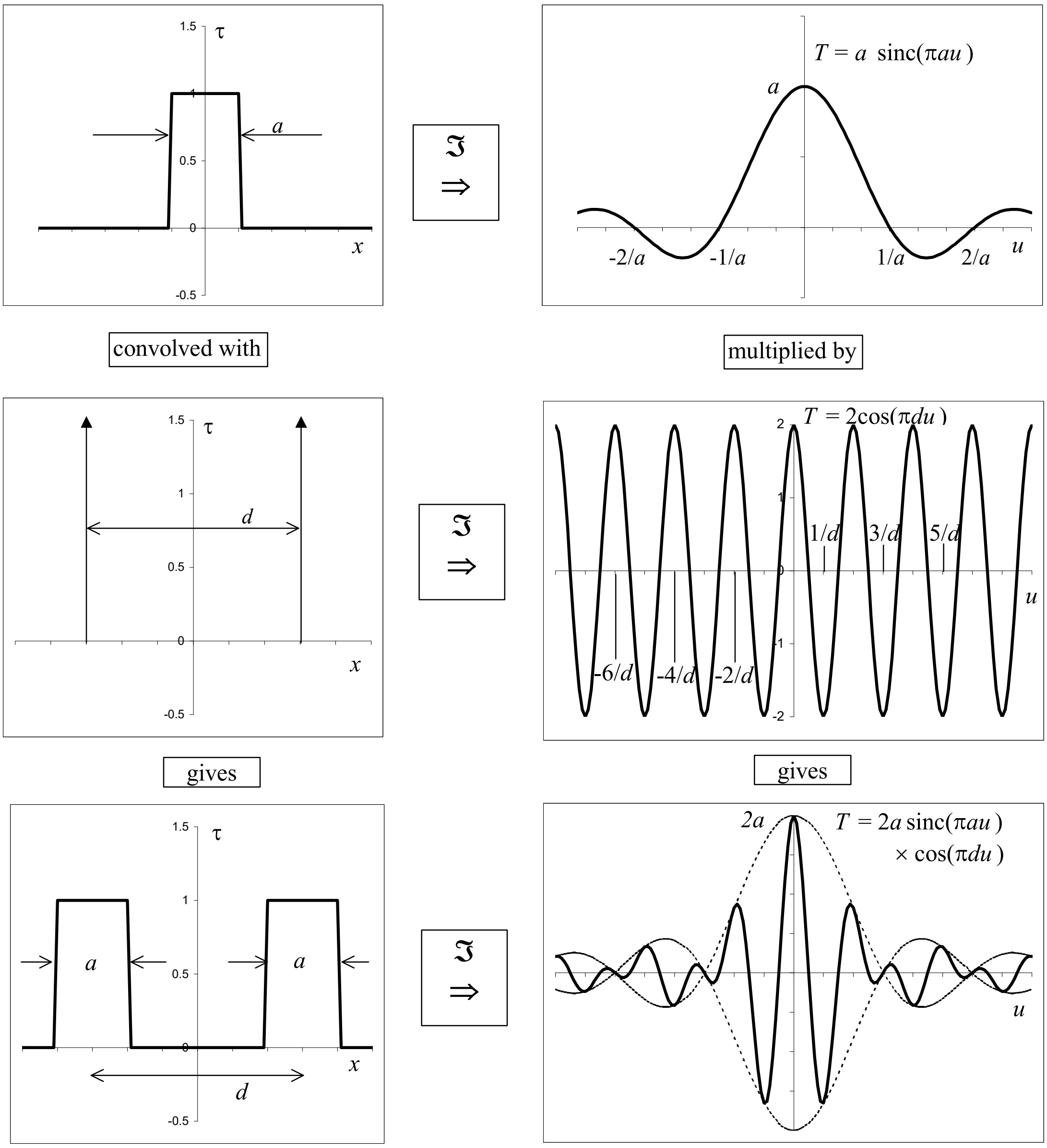
i.e., $$T(u,v) = T_{\text{Young}}(u,v) \times T_{\text{slit}}(u,v) = 2\cos(\pi\, du)a\,\text{sinc}(\pi\, au)$$

and the power density in the diffraction pattern is proportional to

$$T^{*}T(u,v) = 4a^2\,\text{sinc}^2(\pi\, au)\cos^2(\pi\, du)$$

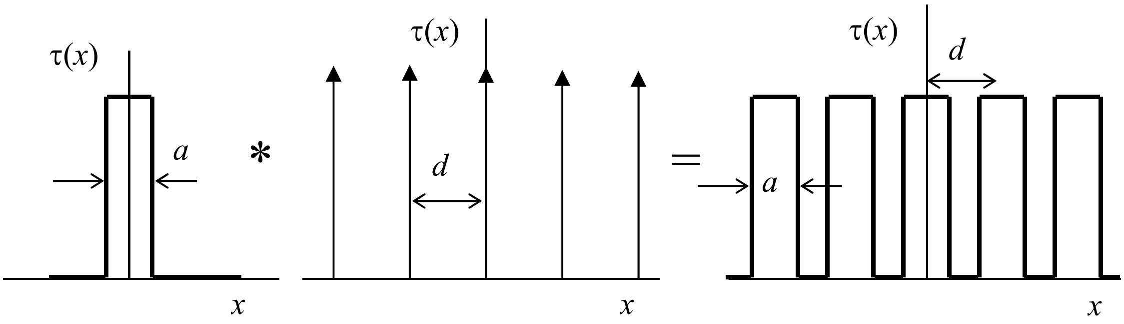
These functions and relationships are illustrated on the next page. If the slits are very narrow, then the sinc function is very wide, and Young's fringes have almost equal brightness across the pattern. As the slit width increases, the sinc function becomes narrower and the interference fringes fall off in brightness more rapidly.

3. DIFFRACTION



D. Multiple Slits

[PPP 12; S&K 10.1-4, 12.2-5]



The double slit result can be extended to the case of N slits, each of width a, and with a regular spacing d. As before, the Fourier Transform will be the multiplication of the Transform of a single slit aperture (sinc function) with the Transform of a series of N equally spaced delta functions. Let N be an odd number, for the sake of simplicity.

The transmission function for the series of delta functions can be written as:

$$\tau_{\text{del}}(x, y) = \sum_{n=-(N-1)/2}^{n=+(N-1)/2} \delta(x - nd)$$

Using the shift and linearity theorems, its transform is:

$$T_{\text{del}}(u, v) = \sum_{n=-(N-1)/2}^{n=+(N-1)/2} \exp(i2\pi ndu)$$

This is a geometric series with N terms, similar to the case of multiple-beam interference in the Fabry-Perot interferometer, but here the amplitude of each term is constant. The first term in the series is $\exp\{-i\pi u(N-1)d\}$ and the common ratio is $\exp\{i2\pi ud\}$, so the sum is:

$$\begin{aligned} T_{\text{del}}(u, v) &= \exp\{-i\pi\, u(N-1)d\}\frac{1-\exp(i2\pi\, uNd)}{1-\exp(i2\pi\, ud)} \\ &= \frac{\exp(i\pi\, uNd) - \exp(-i\pi\, uNd)}{\exp(i\pi\, ud) - \exp(-i\pi\, ud)} \\ &= \frac{\sin(\pi\, Ndu)}{\sin(\pi\, du)} \end{aligned}$$

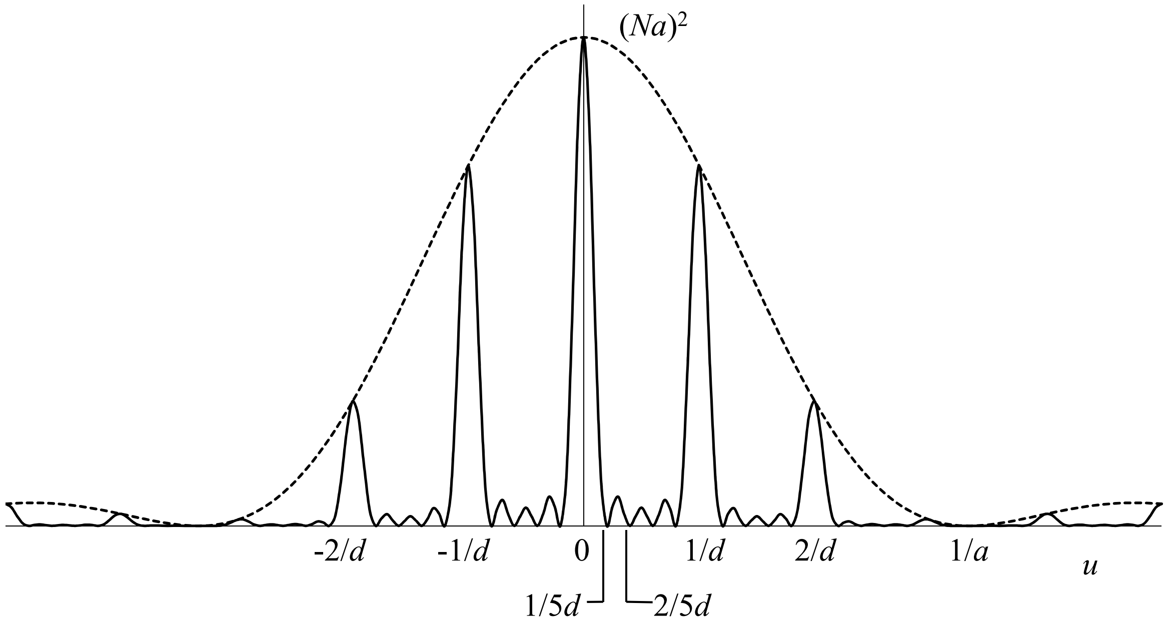
Using the convolution theorem, the total Fourier Transform is:

$$T(u, v) = a\ \text{sinc}(\pi\, \text{au})\frac{\sin(\pi\, Ndu)}{\sin(\pi\, du)}$$

and the power density is proportional to

$$\boxed{T^*T(u, v) = a^2\ \text{sinc}^2(\pi\, \text{au})\frac{\sin^2(\pi\, Ndu)}{\sin^2(\pi\, du)}}$$

The plot shows this function for the case of $N = 5$.



Whenever $\pi du = m\pi$, where m is an integer, the factor $\left(\frac{\sin(\pi\, Ndu)}{\sin(\pi\, du)}\right)$ becomes undefined because both top and bottom lines go to zero. However, using the small-angle approximation $\sin\theta \approx \theta$ when θ is small, we find that the ratio has a limiting value of N. This gives rise to a *principal maximum* at $u = m/d$, where m is the same as the *order of interference* between rays from adjacent pairs of apertures. These are the same angles that the bright fringes of a double-slit diffraction pattern would appear at, if the double slit separation was d.

The power density goes to zero whenever $\sin(\pi Ndu) = 0$ and $\sin(\pi du) \neq 0$, that is, when $u = l/(Nd)$ for integers l which are *not* multiples of N. Between each pair of principal maxima, there are $N - 1$ zeroes and $N - 2$ *secondary maxima*.

The overall diffraction pattern is modulated (multiplied) by the sinc(πau) "envelope", which goes to zero at $u = n/a$ for any integer n except $n = 0$. This is the same as the diffraction pattern for a single slit aperture of width a. Hence we can distinguish three parts to the total pattern:

1. An overall envelope which has the same dimensions as the diffraction pattern of a single slit.

2. Within this envelope, a series of equally-spaced (in terms of u) bright fringes, which occur at the same angles as would double-slit fringes for a slit spacing of d. We can associate an order of interference m with each of these maxima.

3. But instead of the cosine shape of double-slit fringes, the peaks are now much narrower and sharper. In between each pair of principal maxima, there are a number of much fainter fringes.

The distance from a peak of a principal maximum to the first zero on either side is

$$\Delta u = 1/(Nd).$$

This is also approximately the FWHM (full width at half-maximum) of each principal maximum.

As the number of apertures N increases, the principal maxima become sharper and narrower, and the secondary maxima increase in number and become smaller relative to the principal peaks.

Application: Diffraction Grating

The multiple-slit aperture is the basis of the diffraction grating, which is used in spectrometers for analysing the wavelength spectrum of light sources. Diffraction gratings can be used either in transmission or in reflection: in the latter case, light is reflected from a series of parallel rulings or ridges on a reflective surface, but the principles are the same as for transmission.

Properties of the diffraction grating were covered in the Second Year optics course (PHYS2125). Main results are summarised here.

So far we have assumed that the light is incident perpendicularly on the diffracting aperture; that is, the transmission function is purely real, with no phase variation across the aperture. Diffraction gratings are often used with the light incident at an angle θ_i to the normal. The transmission function for the multiple slit then needs to be *multiplied* by a *phase factor*

$$\exp(i\, 2\pi \sin\theta_i\, x/ \lambda)$$

expressing the fact that the phase changes periodically by one cycle for every displacement of $\lambda/\sin\theta_i$ across the aperture. Using the convolution theorem, we deduce that the Fourier Transform must be *convolved with* the Transform of this phase factor, which from the point source example above is just a delta function at $u_i = \sin\theta_i / \lambda$. Convolving any function with a delta function simply shifts the origin of the new function to the position of the delta function, so now principal maxima occur when

$$u - u_i = m/d\,.$$

Putting $u = \sin\theta_x / \lambda$, where θ_x is the angle of diffraction, the condition for principal maxima becomes:

$$\boxed{m\lambda = d(\sin\theta_x - \sin\theta_i) \quad \text{for } m = 0, \pm 1, \pm 2, \ldots}$$ **The Grating Equation**

As before, m is the order of interference. The zero order, $m = 0$, corresponds to the direct beam through the diffraction grating (or the mirror reflection for a reflection grating), which occurs at $\theta_x = \theta_i$. The first peaks on each side are the first order of diffraction, $m = \pm 1$, etc.

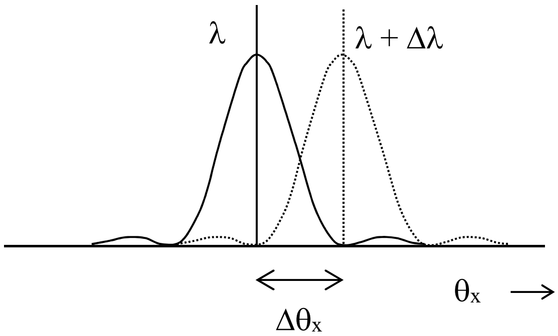
3. DIFFRACTION

The diffraction grating is used to study the spectrum of light sources containing a spread in wavelengths. If there is more than one wavelength present, then the angle at which a particular order of principal maximum occurs depends on wavelength, according to the grating equation. The "sensitivity" of the grating is measured by its *angular dispersion*, given by

$$\boxed{\text{Angular Dispersion } \mathcal{D} = \left|\frac{d\theta_x}{d\lambda}\right| = \frac{m}{d\cos\theta}}$$

Thus for good dispersion, we want large *m* and small *d*. As in the case of the Fabry-Perot interferometer, we are interested in how close together two wavelengths can be and still be distinguishable as separate.

By the Rayleigh criterion, the two wavelengths can be resolved provided that their angular separation is not less than the angular distance between the peak and first zero of one of them.



$\Delta\theta_x$ corresponding to $\Delta\lambda$ is: $\dfrac{m\Delta\lambda}{d\cos\theta_x}$

$\Delta\theta_x$ due to width of peak is: $\dfrac{\lambda}{Nd\cos\theta_x}$

Equating these, we find the *limit of resolution* or *minimum resolvable bandwidth*:

$$\boxed{(\Delta\lambda)_{\min} = \frac{\lambda}{Nm}}$$

and

$$\boxed{\textbf{Chromatic Resolving Power } \Re = \left(\frac{\lambda}{(\Delta\lambda)_{\min}}\right) = Nm}$$

It is interesting to compare the diffraction grating with the Fabry-Perot interferometer in their performance as spectrometers.

For the Fabry-Perot, we had $\Re = m\,\mathcal{F}$, where the finesse $\mathcal{F}$ was a moderate number (typically 20 - 100) and order *m* was large. For the diffraction grating, we have $\Re = m\,N$, where order *m* is small, but *N* is large.

There are also differences in light *throughput* (area-solid angle product). For a given resolving power, because the Fabry-Perot works with two-dimensional circular fringes, it can have a much higher throughput than the diffraction grating, which essentially works with one-dimensional fringes. The disadvantage of the Fabry-Perot is its greater complexity, and the overlap of orders of interference. Overlap of orders can also occur with the diffraction grating, but since *m* is very small, the effective "free spectral range" is large.

Exercise Set 3.2

1. Prove the Convolution Theorem, starting with the definition of the Fourier Transform.

2. Consider the Fraunhofer diffraction pattern that would be produced by two circular apertures of the same diameter, side by side, with a centre-to-centre separation equal to their diameter (that is, just touching). How many bright parallel interference fringes appear within the central Airy disc?

3. Find an approximate expression for the ratio of the power densities at the principal maximum to that at the first secondary maximum on either side, in the Fraunhofer diffraction pattern of an N-slit multiple aperture.

3.3 Fresnel Diffraction

[H 10.3; K&F 7.1, 7.2; PPP 13.1-6; S&K 8.1, 8.5]]

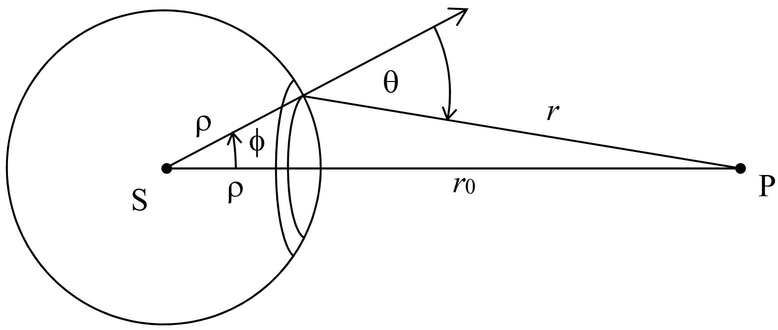
We return to a consideration of the more general case of diffraction where the Fraunhofer condition is *not* satisfied - the case of Fresnel diffraction, based on the Huygens-Fresnel Principle and the Diffraction Integral of Section 3.1.

We firstly try out these principles in a simple case to see whether they give the right answer. Suppose a point source at S generates spherical waves. From first principles, the electric field at some point P, a distance R from S, will be given by (see Section 3.1):

$$E_{\mathrm{P}} = \frac{A_{\mathrm{S}}}{R} \exp\{i(\omega t - kR)\}$$

This is just the equation of a spherical wave. The amplitude decreases in inverse proportion to the distance from the source, and the wave spreads out from the source with an angular frequency ω and a wavenumber k. The factor A_{S} is some constant, which we don't know yet, that depends on how bright the point source is, and we call it the *source strength* of S. Here we are ignoring the vector properties of the field, and just considering its scalar magnitude.

Now another way to derive the field at P is to apply the Huygens-Fresnel principle to any one of the spherical wavefronts emitted from S. According to this principle, the field at P should equal the sum of the fields due to the secondary wavelets emitted from *all over* the spherical surface of the primary wavefront.



Let the radius of the primary wavefront sphere be ρ, and the distance of P from this wavefront be r_0, so that the total distance from S to P is $R = \rho + r_0$. From the above equation, the field at the wavefront is:

$$E(\rho) = \frac{A_{\mathrm{S}}}{\rho} \exp\{i(\omega t - k\rho)\} = \left(\frac{A_{\mathrm{S}}}{\rho} e^{-ik\rho}\right) e^{i\omega t}$$

The last quantity in brackets is just the complex amplitude and phase of the field on the primary wavefront.

Now according to the Diffraction Integral, the total field at point P should equal

$$E_{\mathrm{P}} = C\, e^{i\omega t} \iint_{\text{sphere}} \left(\frac{A_{\mathrm{S}}}{\rho}\right) \frac{e^{-ik\rho}\, K(\theta) e^{-ikr}}{r} \, ds \,,$$

where $K(\theta)$ is the inclination factor, and we integrate with respect to area ds over the whole surface of the sphere. Since A_S and ρ are constants for this integration, we have:

$$E_P = C\left(\frac{A_S}{\rho}\right)\exp\{i(\omega t - k\rho)\}\iint_{\text{sphere}} \frac{K(\theta)e^{-ikr}}{r}\,ds$$

To evaluate the integral, we need to express the element of area ds in terms of length r. To do this, we divide the spherical surface up into thin *annuli*, centred on the line SP. A particular annulus contains all of the sphere surface which lies between distances r and $r + dr$ from P. What is the area ds of such an annulus?

Firstly, we know from spherical polar geometry that ds is given by $ds = 2\pi\rho^2 \sin\phi\, d\phi$, where $d\phi$ is the angle that the annulus subtends at the centre of the sphere. We need to get a relation between $d\phi$ and dr. Using the cosine rule, we have

$$r^2 = \rho^2 + (\rho + r_0)^2 - 2\rho(\rho + r_0)\cos\phi$$

Differentiating, with ρ and r_0 being constants, $\quad 2r\,dr = 2\rho(\rho + r_0)\sin\phi\, d\phi$

and so $\quad ds = \dfrac{2\pi\rho\, r\, dr}{(\rho + r_0)}$.

Substituting this back into the integral gives:

$$E_P = CA_S \exp\{i(\omega t - k\rho)\}\frac{2\pi}{(\rho + r_0)}\int_{r=r_0}^{r=r_0+2\rho} K(\theta)e^{-ikr}\,dr$$

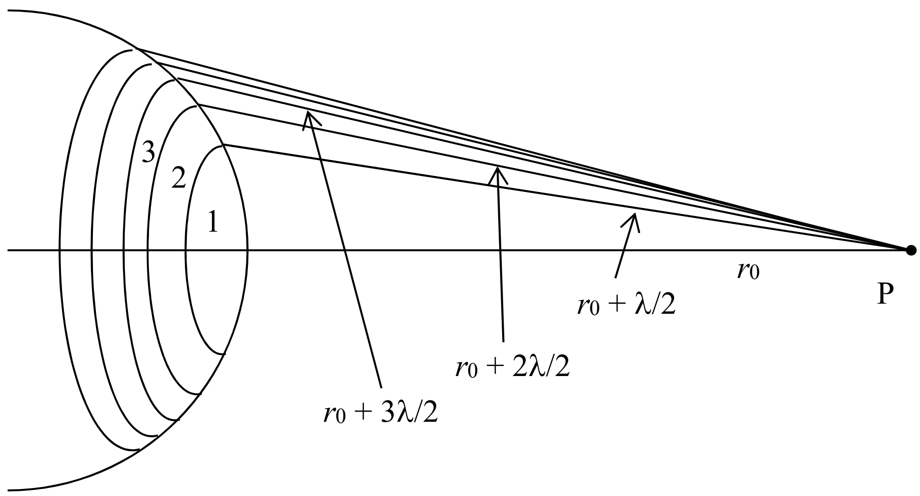
The important things here are that the factor r in the bottom line of the original integral has cancelled with the factor of r in the expression for ds, simplifying the integral, and the integral itself has been converted from two dimensions to one dimension.

In the integral, the only two factors are e^{-ikr}, a *phase factor* whose amplitude is 1 but whose phase varies rapidly as r varies, and the inclination factor $K(\theta)$, which will vary only very slowly in comparison.

To get an idea of what is going on in this integral, it is useful to divide the surface of the sphere up into particular annuli known as *Fresnel zones* or *half-period zones*, and then also to use a vibration curve to picture the result.

Firstly, we define the Fresnel zones to be annuli such that the distance r increases by half a wavelength from the edge of one zone to the next. That is,

the distance from P to the nearest point on the sphere is r_0 ;
the distance from P to the edge of the first zone is $r_0 + \lambda/2$;
the distance from P to the edge of the second zone is $r_0 + 2\lambda/2$;
the distance from P to the edge of the third zone is $r_0 + 3\lambda/2$;
the distance from P to the edge of the nth zone is $r_0 + n\lambda/2$; etc.



Because of this half-wavelength increase from one zone to the next, and since the field over the surface of the sphere all has a constant phase, the secondary wavelets arriving at P from adjacent zones will be *out of phase* with each other, and will tend to cancel.

We split the above integral up into a piece for each zone. The field at P from the lth zone will involve the integral:

$$\int_{r_0+(l-1)\lambda/2}^{r_0+l\lambda/2} K(\theta) e^{-ikr}\, dr \approx K_l \int_{r_0+(l-1)\lambda/2}^{r_0+l\lambda/2} e^{-ikr}\, dr$$

Since $K(\theta)$ will be almost constant over the small range of angles θ involved in the zone, we can replace it with its average value K_l for this zone, and take it outside the integral.

What is left is
$$\int_{r_0+(l-1)\lambda/2}^{r_0+l\lambda/2} e^{-ikr}\, dr = \left[\frac{e^{-ikr}}{-ik}\right]_{r_0+(l-1)\lambda/2}^{r_0+l\lambda/2}$$
$$= \frac{2i\lambda}{2\pi} e^{-ikr_0} (-1)^l$$

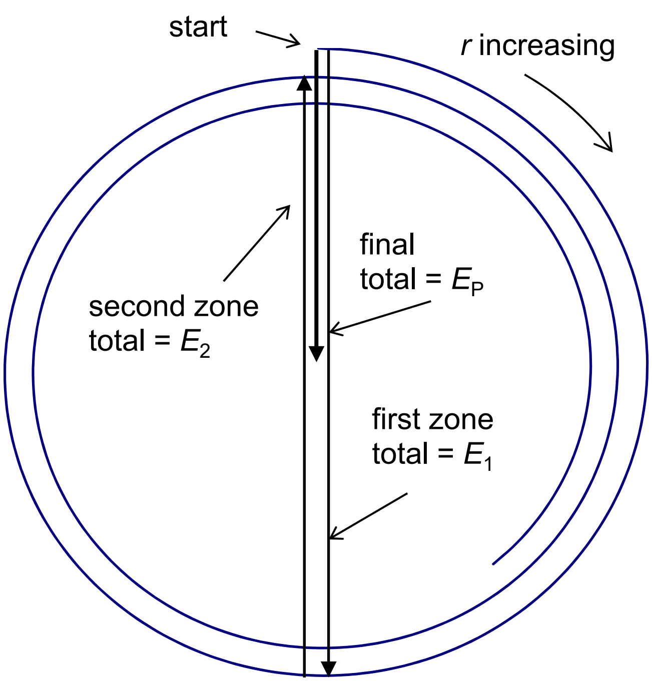
(Exercise: Fill in the missing steps!)

Putting this all together, the total field due to the single zone number l is:

$$E_l = CA_S i\lambda \frac{\exp\{i(\omega t - k[\rho + r_0])\}}{(\rho + r_0)} 2K_l (-1)^l$$

Since everything here is a constant except l, it means that the resultant total field from each zone *is the same* except for the inclination factor, which changes only slightly from zone to zone, and the sign of the field, which alternates from zone to zone. That is, the fields from adjacent zones would cancel *exactly* if it were not for the inclination factor. This is why the zone construction was carried out the way it was. It so happens that the fall-off in amplitude of the secondary wavelets reaching P, due to the $1/r$ factor, is just compensated for by the increase in surface area of the Fresnel zones as r increases (remember the cancelling factor of r).

Vibration curve for Fresnel diffraction



The vibration curve helps to explain this further. Imagine each Fresnel zone split into a large number of sub-zones, and draw the phasor diagram representing the summation of phasors from each sub-zone. The phase of the resultant light from successive subzones gradually decreases as r increases. After a whole zone has been crossed, the phase is opposite to what it was at the start, because r has increased by $\lambda/2$. After two zones have been crossed, the phase is back to the original, and the vibration curve is *almost* back where it started, but not quite, because the inclination factor has decreased slightly, and the phasor amplitudes have decreased a little bit. The vibration curve is therefore a *spiral*, which goes through half a revolution for every extra Fresnel zone. The *resultant* field for each zone alternates in sign.

The vibration curve eventually spirals into its centre as r increases, and as the inclination factor drops to zero. The final Fresnel zone is around the back of the sphere, on the other side to point P, where $K(\theta) \approx 0$. The total field at P is given by the phasor from the starting point of the vibration curve to the end point, in the centre of the circle. From the diagram, it is apparent that this is just equal to half the field due to the first Fresnel zone alone.

Hence $\boxed{E_P = \frac{1}{2} E_1}$

Substituting the expression for E_l with $l = 1$, and putting $K_1 = 1$, gives

$$E_P = CA_S i\lambda \frac{\exp\{i(\omega t - k[\rho + r_0])\}}{(\rho + r_0)}(-1) \ .$$

But this should be the same as the first-principles expression that we started with,

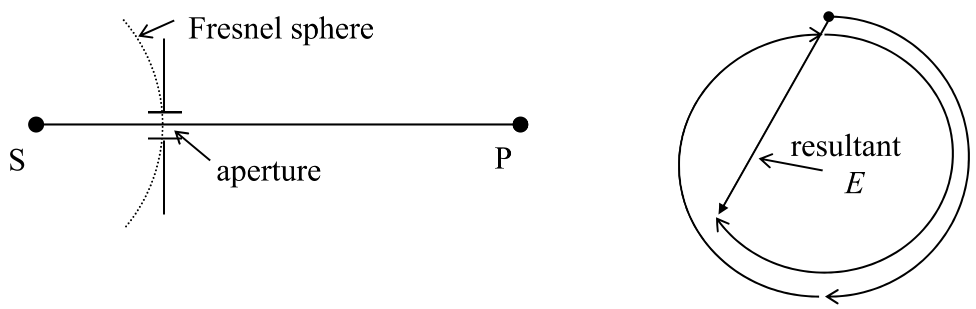
$$E_P = \frac{A_S}{R}\exp\{i(\omega t - kR)\}.$$

Comparing, we find that they do agree provided that $\boxed{C = \frac{i}{\lambda}}$.

Thus the Huygens-Fresnel principle does indeed give the right answer, if we take the value of the constant C, which determines the amplitude of the secondary wavelengths, to be i/λ. The factor i implies that the secondary wavelets are 90° *ahead* in phase of the wavefront which generates them. This can also be seen from the vibration curve, which shows that the contribution from the centre of the first Fresnel zone (at the start of the spiral) is 90° ahead of the total field E_P.

3.5 Applications of Fresnel Diffraction

We will look at only one example of Fresnel diffraction, namely that involving circular apertures or obstacles, since this is easily treated with the use of the vibration curve of section 3.4.



Suppose a circular aperture is centred on a line between a point source S and a point of observation P. The total field at P will be determined by the number of Fresnel zones that the aperture lets through. The field will be given by the resultant from the starting point on the vibration curve to the point corresponding to the diameter of the aperture.

Note that as the diameter of the aperture is increased, the field at P will be alternately dark and bright as successive Fresnel zones are "uncovered". When an odd number of zones is uncovered, $E \approx 2E_P$, where E_P is the field with no obstruction. Hence the power density will be $S \approx 4S_P$, that is, four times the power density with no obstruction.

When the number of zones is even, $E \approx 0$ and $S \approx 0$.

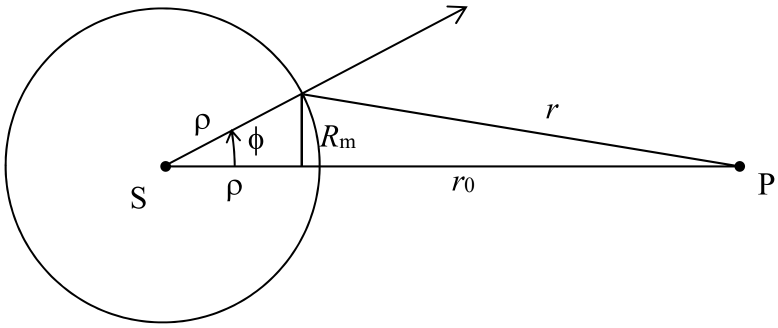
This applies just to the *central point* of the 2-D diffraction pattern at P. There will be more complicated diffraction effects surrounding P. As the diameter of the aperture increases, more and more fringes appear around the central point, but the centre itself alternates between light and dark.

Alternatively, consider a circular *obstacle* centred on the line SP (e.g. a ball bearing). Now the first part of the vibration curve is obstructed. However, the resultant phasor still ends at the centre of the vibration curve. The power density should be almost the same as if there was no obstruction at all: i.e., there will be a *bright spot in the centre of the shadow*!

This prediction was originally an obstacle to the acceptance of the wave theory for the propagation of light, until someone actually did the experiment, and verified that the bright spot was in fact there.

Radii of Fresnel zones

How do we know how many Fresnel zones are included within the diameter of the aperture or obstruction above?



Let R_m be the radius of the mth Fresnel zone. Then by definition, $r = r_0 + m\lambda/2$.

Using the cosine rule,

$$(r_0 + m\lambda/2)^2 = \rho^2 + (\rho + r_0)^2 - 2\rho(\rho + r_0)\cos\phi$$

$$\text{i.e.} \quad r_0 m\lambda + m^2\lambda^2/4 = 2\rho(\rho + r_0)(1 - \cos\phi)$$

We will only be concerned with small apertures, so that ϕ is very small, and $\cos\phi \approx 1 - \frac{1}{2}\phi^2$. Also $m\lambda/2 << r_0$, and $m^2\lambda^2/4$ can be neglected. Therefore

$$r_0 m\lambda = \rho(\rho + r_0)\phi^2 \approx \rho(\rho + r_0)\left(\frac{R_m}{\rho}\right)^2 = \left(\frac{\rho + r_0}{\rho}\right)R_m^2$$

i.e.,

$$\boxed{R_m = \sqrt{m\lambda\left(\frac{\rho r_0}{\rho + r_0}\right)}}$$

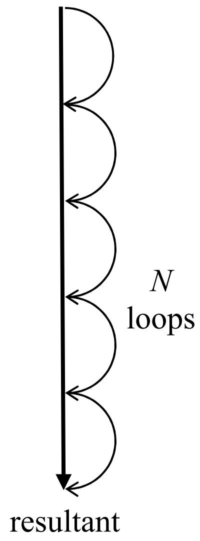
Notice that the radius of the mth zone is proportional to $\sqrt{m}$, so the zones become thinner as m increases. The pattern has a similar appearance to the Haidinger fringe pattern of a Michelson or Fabry-Perot interferometer, provided that $R_m << \rho$.

Application: Fresnel Zone Plate

Suppose an aperture is made in which every alternate Fresnel zone is opaque, and the others are transparent. A screen like this is known as a *Fresnel zone plate*. (Not the same as the "Fresnel lenses" which are used in overhead projectors). Then instead of light from adjacent zones cancelling, they will all add together with the same sign, producing a much larger total field at P.

The vibration curve for this case looks like a set of half-circles stacked on top of each other, as shown below.

3. DIFFRACTION



Suppose the total field without any obstruction is E_P as before, and that there are $2N$ zones altogether, with N clear and N opaque. The field from the first zone alone would be $E_1 = 2E_P$ as before. With the zone plate in place, the total field will be E_T, given by

$$E_T \approx NE_1 = 2NE_P \ .$$

So the final power density will be $S_T = 4N^2 S_P$. This increase in power density only occurs if the zone plate radii are correct for the combination of λ, ρ and r_0 being used.

If we rearrange the equation for R_m into:

$$\frac{1}{r_0} + \frac{1}{\rho} = \frac{m\lambda}{R_m^2} \ ,$$

we see that the Fresnel zone plate obeys the same law as a simple *lens*, of effective focal length

$$\boxed{f = \frac{R_m^2}{m\lambda}} \ .$$

Thus it forms an "image" of the source S at point P, but by a process of diffraction, not refraction.

For an incident *plane* wave, we let $\rho \to \infty$, and then

$$R_m = \sqrt{m\lambda r_0} \qquad \text{for plane wave illumination.}$$

The "focal length" in this case is just r_0.

A further improvement is possible if, instead of every second zone being opaque, it is transmitting but with a reversed phase. This can be done by etching away or evaporating a thin layer onto a transmitting film so its thickness is different by an optical path difference of $\lambda/2$ in the alternate zones. This will give an even greater field at P.

The best we can do towards getting the biggest field possible at P is to straighten out (uncurl) the vibration curve completely into a straight line, so that all rays reaching S from P arrive with the same phase. In fact there is an optical device that does this - a simple positive, converging lens!

Exercise Set 3.3

1. Plane waves impinge at normal incidence on a screen with a small circular hole in it. It is found that, when viewed from some point P on the line through the centre of the hole, half of the first Fresnel zone is uncovered by the hole. What is the electric field amplitude at P in terms of what it is when the screen is removed? What is the power density at P?

2. A Fresnel zone plate is to be produced having a focal length of 2.0 m for He-Ne laser light of wavelength 632.8 nm. An ink drawing of 20 zones is made with alternate zones shaded in, and a reduced photographic transparency is made of the drawing.

   (i) If the radius of the first zone in the original drawing is 112.5 mm, what photographic reduction factor is required?

   (ii) What is the radius of the last zone in the original drawing?

   (iii) What is the power density at the focus of the final zone plate, if the power density with no obstruction present in the beam is 50 $\mu W/cm^2$?

3. Monochromatic plane waves of wavelength 500 nm are incident normally on a screen with a circular opening of radius 5.0 mm. Let P be a point a distance 2.5 m on the other side of the opening in the centre of the geometrical bright region. The electric field amplitude at P in the absence of any screen is 4.0 V/m.

   (i) How many Fresnel half-period zones are there in the opening as seen from P?

   (ii) What is the amplitude of the electric field at P, in the presence of the screen?

   (iii) What is the field at P when a zone plate is put in the opening that blocks out every second half-period zone, and lets all the light from the other zones through?

   (iv) What is the field at P when a zone plate is put in the opening that shifts the phase of every second half-period zone by π radians, without affecting the other zones?

   (v) What is the field at P when a perfect lens is put in the opening, focusing the rays at P?

   You may neglect inclination factors.

4. A Fresnel zone plate is to be used to produce an image of a point source S at a point P, at a distance of 800 mm from S, in light of wavelength 515 nm. The zone plate is to be placed half way between P and S. Calculate the radius of the first and fifth Fresnel zones.