The field of a disk PhL 12.24.02

Contents of these notes: (this is a long wad of notes, far from perfect, but on the track...)

- A review of how the basic **integral equation** for diffraction is derived
- How to understand **graphically why there are rings** in the pattern of a
  (1)a slit (sinc), and
  (2) same for round hole (Airy)
- Fraunhofer limit requires that Airy amplitude be **small**, condition given
- Discussion of the **geometric limit** and Fresnel to compute it
- My derivation of **Babinet's (scalar) Principle**, with pictures
- Application to find **diffraction of a disk** in the Fraunhofer limit
  (1) First I did this incorrectly ignoring phase, implying result should have dark center
  (2) Then I did this correctly with the **phase**, causes bright center from disk
- Some **Excel plots of disk diffraction** shadow on retina assuming various distances from retina
- Estimation that **eye resolution** is about 27 μ on the retina (from halftone dot experiment)
- **Effect of the eye's lens** on floater diffraction pattern

Review of the General Scalar Theory of Diffraction

The first step is the completely general and exact Green's *Theorem* which equates a volume integral with a surface integral, as in Goodman 3-14, with two arbitrary functions in the integrands here called G and U. The second step involves first deriving the (time-independent) vector wave equation from Maxwell's equations (no charges or currents) and then computing the Green's *Function* exp(ikr)/r for the scalar wave equation, as if exp(ikr)/r could be the component of the electric field vector which solves the vector wave equation. This is an approximation in itself, see later, and this is why we are doing the "scalar theory" of diffraction. Recall that the Green's function solves the wave equation with a delta function unit impulse source on the right side. The points x and x' are called P0 and P1 here. By putting an ε ball around the point P0, and then letting P1 integrate over the volume defined in the figure on page 40, the delta function is always zero in the volume. Thus, G solves the homogeneous wave equation in there. The other function U is taken to be some scalar component of the electric or magnetic field, and it too solves the homogeneous wave equation. As a result, both del-squared operators in the volume integral can be replaced with k-squared, and the entire volume integral side of Green's Theorem is then zero, leaving us with the two-term surface integration part. The surface at this point consists of the normal outer surface and the ε ball surface. The ball surface integral is easily done in the limit ε = 0 and gives just U(P0) at the ball center, so we then end up with basic equation 3-21.

What does this equation tell us? If you know the value of U(P1) *and* its normal derivative everywhere on a bounding surface S, then this equation lets you compute U(P0) at any "observation point" inside the surface. There are two terms inside the integration, one involves U and grad G, while the other involves G and grad U. This equation gets the name Helmholtz-Kirchhoff assigned to it.

In the next section, Goodman applies the above equation to a "screen", by which he means some sort of set of apertures that are localized on a metal plate. The spherical part of the surface is then taken to infinity and it is shown that the surface integral contribution of that surface is then 0.

So now to find the field U to the right of the "screen" at point P0, we only have to integrate over the screen. At this point, the two famous **Kirchhoff Approximations** are made: (1) the field in the apertures

is the same as it would be if the apertures were not there (which is basically true if aperture size is large relative to λ), and (2) you get zero contribution from the metallic portion outside the apertures (again, true in the same limit, otherwise you get some field leaking into the shadow region). The result at this point is 3-24, where Σ refers to integration only over the apertures. Notice there are still two terms inside the integration.

At this point Goodman makes his first assumption about the "field from the left". Suppose it is provided by a unit point radiation source at some place P2 many wavelengths to the left. You then get the formula 3-27. This of course still is based on the two K assumptions noted above: the screen is not distorting that point-source wave in the apertures, and the field does not "leak" behind the screen. You can interpret the integral result here as saying that the aperture is filled with "fictitious" point sources which have the peculiar amplitudes and phases shown in 3-29. This in effect derives and makes more precise what was earlier called Huygen's Principle. That is to say, Huygen's postulated his idea in 1678 without any E&M proof.

At this point, Goodman is unhappy (as was **Sommerfeld**) with the inconsistency implied by the second K approximation because if U and gradU are really zero on the metallic screen, then U has to be zero everywhere. So we back up to 3-23 and try some *different* functions for G (recall that G is arbitrary in Green's Theorem). The first new G (called G-) is one that solves the wave equation with two delta sources phased 180 degrees out, one at the mirror image location P0~ in addition to the source at P0. The advantage of this G is that it clearly vanishes for any point on the planar screen or aperture, so you only get the U grad G term in the integration. The result is 3-36 where we end up with just the U grad G term in the original (but inconsistent) formula, but with an extra factor of 2 overall. Another choice G+ gives 3-39, which is the G grad U version, it too has a factor of 2.

So the point is that we now have formulas that require only that either U or grad U (but not both) vanish on the metal portion of the screen. I don't think the 1/r01 term needs to be dropped for result to be OK. Goodman then compares the three different formulas and they differ only in obliquity factors.

At this point we have three different fully general diffraction-from-screen formulas, and we tend to prefer one of the Sommerfeld forms. In order to really justify the K approximations, we really have to assume that the apertures are large compared to a wavelength -- this would explain why we can ignore distortion of the field by the screen (distortion is localized to areas near the aperture edges), and why we can ignore leak-through behind the screen (same reason).

This is all the "classical theory of diffraction". Note that Som I contains $\cos(\theta)/r01 = z/(r01)^2$, where z is distance from screen to plane of observation. If we keep just the first two terms in a power series expansion for r01 in the phase (and just the first term in the magnitude), we quickly end up with 4-14 which is the "**Fresnel**" diffraction formula. We now have the famous exp(ikz)/z in front of the integral, and the "extra quadratic phase" inside based on the transverse coordinates. The claim is that Fresnel is accurate even in the "near field" region, even though we have expanded r01 in this manner. Equation 4-17 is an equivalent form of Fresnel.

If you now make an additional assumption on the transverse dimensions of interest and λ and z, then you can ignore the first internal phase factor in 4-17, and then you get the result that the "field to the right" is simply the 2D Fourier transform of the field at the screen, for any distance z. This is the **Fraunhofer** result 4-25. Now when you do a normal Fourier transform of some f(x) to get F(k), you have $k = 2\pi/\lambda$, where k is the "conjugate variable". Or for f(t) to get F(ω), you have $\omega = 2\pi/T$. Or you might go f(t) to F(f), with f = 1/T. Here, however, the conjugate variables to x and y have extra scaling factors! We have $fx = (x/z)/\lambda$ where x and z are observation coordinates. The factor (x/z) is what is new, the scaling factor. This is now the result varies with z, for example.

The **Airy disk** is the Fraunhofer result for a circular aperture, result is 4-31. Today I tried to create an Airy pattern with a flashlight and a small hole in cardboard, but I could not really see the first outer ring. The light us just too dim, you really need a laser source to see this stuff.

Now, in physical terms, what makes the rings of the Airy pattern?

(1) First, think of the one-dimensional box whose Fourier transform is the sinc function -- what makes the peaks and valleys in this case? It is not obvious a priori that just adding a wall of Huygen's sources is going to give a sinc function.

For the zeros, we have a nice trick here, we can divide the "fictitious sources" within a 1D slit into pairwise groups of height dy in the slot. In the Huygens arm-waving sense, we note pairwise cancellation for all differential group pairs, since they all differ say by $\lambda/2$. In more down to earth sense, you are integrating cosine and you have canceling positive and negative area contributions, so you will have perfect zeros at the usual places, namely, ka = 0, 2π, 4π and so on. Or, think of integrating the phaser exp(ikx). We know the integral is going to be (1/ik) exp(ikx) $|^{a}_{0}$ = (1/ik) [exp(ika) - 1]. Clearly this integral is zero every time ka = (2n)π. [if k = 2πf, then when fa = n].

Computing the height of the peaks is less easy! In the Huygens model, there is nothing so simple you can do in this case. For the zeros, we could talk about making pairwise cancellations between small dy height differential groups, but for the maxima pairwise adding these things is fine, but as we slide along, phase varies and we have no conclusion. The simple integral formula however suggests that maxima will be *near* values ka = nπ for odd values of n, because this is where the phasor difference is maximum, [exp(ika) - 1] = -1 -1 = -2. However, the exact locations are slightly different due to the 1/k factor outside. Differentiate the sinc function and set derivative to zero to find that maxima are at tanx = x. For large x, you get the intuitive maxima points because the k out-front varies so slowly. Finally, if you just think about integration of cosine, you know maxima will occur roughly periodically when you have cancelled all previous areas and you finish a new hump.

So, for the 1D square box case, we can explain the "rings" as follows: (1) we know from simple integration of the exponential (giving phasor difference) that there will be perfect zeros at the usual locations. (2) we know that the phasor difference is bounded by 2, so we know there will be maxima somewhere near the usual maxima points, with height roughly 1/k but not exactly. Thus, as k increases, the height of these maxima decreases. This is in fact what the sinc function does. We also can see what the sign of the sinc function changes, since area accumulates in both sign senses as you move along.

Here is an easy construction to think of: the Fourier transform of the box is roughly integral 0 to a of the function cos(kx)dx, where a is the box half-width. The idea is to keep the x-scale and position of a on that scale fixed, and to think about varying k. For small k, the cosine is just 1 and area is simply a, and this is the max value of the function sin(ka)/k. For larger k, the area is smaller, dropping I think to $(2/\pi)$a at the end of the first positive cosine hump. Then you start getting area cancellation as you increase k and hence the cosine frequency. Your zeros then occur whenever k is such that a falls at a point where all areas cancel, and these locations are exactly periodic. Your maxima will be NEAR where you have had full cancellation and you build more positive area with a new half-hump. As k increases, the area of this new half-hump is smaller and smaller as it becomes a narrower and narrower sliver. This reduction in area is exactly the 1/k factor in the sinc function.

Scaling factors. First, note from Goodman page 5 that both forward and reverse Fourier Transforms have no overall constants, and I think this follows from using the conjugate variables like f = k/(2π). Variable k is radians per meter, but f is waves per meter. If we make a box of FULL width a (which would be described as his rect(x/a)), then the integration gives the FT to be a*sinc(fa), where sinc(s) =

$\sin(\pi s)/(\pi s)$, just as he states in the table on page 14, but there he has a = 1/a just to confuse me. Note that the sinc function defined in this way has its zeros at s=1,2,3, and that s is the number of half waves.

So far we have just commented on the fact that the Fourier transform of a box is a sinc. The connection with an actual diffraction pattern is provided by the Fraunhofer formula which says this:

$U(x,y)$ = the diffraction pattern at $z = [\exp(jkz)/(j\lambda z)] \exp[jk(x^2+y^2)/2z] \; FT(x/\lambda z, y/\lambda z)$

and for the box, we know that $FT(f_x, f_y) = a*\text{sinc}(f_x a)*b*\text{sinc}(f_y b)$, and we may conclude the actual diffraction pattern looks like this:

$U(x,y) = \exp(jkz) \bullet 1/(j\lambda z) \bullet \exp[jk(x^2+y^2)/2z] \; (ab) \; \text{sinc}(xa/\lambda z) \; \text{sinc}(yb/\lambda z)$

The diffraction zeros will be at $xa/\lambda z = 1,2,3$ etc and $yb/\lambda z$ = the same. As you would expect, the zeros are at fixed ratios of x/z, so as you back up the screen in z, the pattern just expands linearly. Notice that the overall amplitude is proportional to the area = ab of the aperture, but inversely proportional to the distance z and wavelength λ. Also, there are some strange phase factors that don't affect intensity. The intensity drops off as $1/z^2$ again as you would expect. The above agrees with Goodwin's page 75 result if we set $a = 2w_x$ and $b = 2w_y$. About the phase factors: they are really completely obvious. The true phase factor at any point is $\exp(jkr_{01})$ and in Fraunhofer we are assuming a very small aperture so we neglect its transverse dimensions, and then the second phase factor above is just a correction for the extra distance from the aperture to the flat observation screen when you move off axis. The first phase factor of course is the right one for screen center. The FT object has dimensions of area, so we are not surprised to see some multiple of the aperture area appear.

(2) The case of the **circular aperture**. In the *rectangular* aperture case, we were able to give simple arguments as to why we expect zeros and maxima, and hence why we expect the classical diffraction type pattern. How might we make those arguments in the circular case? In the 1D slit case, we explained the zeros by doing an integral of cosine, and showing that we had evenly spaced values of k where the area cancelled to give 0 -- and these were then the zeros of the diffraction pattern.

Well, in this case we end up with integral$(0,2\pi)$ of the phasor $\exp(jkr \sin\theta)$ or of $\cos[\, kr \sin\theta \,]$ as the angular integral around the azimuth. This function $f(\theta) = \cos[\, kr \sin\theta \,]$ is very hard for me to picture, I plotted it in Excel for various kr. It is of course periodic and bounded in +1 and -1, but is not a simple symmetrical function like sin is. It is a sine-like function in that it goes back and forth between +1 and -1, but the spacing between humps is uneven. Since it is sine-like, it seems reasonable that if you integrate out to some fixed distance a, as you vary kr, you will get zeros -- places where the area cancels and gives 0. But I cannot intuitively prove this is the case. It turns out that the integral of this $f(\theta)$ is in fact $J_0(kr)$, and we know this thing is extremely cosine-like but the humps are not quite spaced right -- you have to go and look up the zeros and maxima in a table somewhere.

So, what we are saying is this: if you have an annular ring slot of radius a as your aperture, the Fourier transform is going to be $J_0(ka) = J_0(2\pi fa)$, and then we replace f as usual (Fraunhofer) with $\rho/\lambda z$ where ρ is radial distance from center on the observation screen, so we get then $a \bullet J_0(2\pi\rho a/\lambda z)$, or replacing $2\pi/\lambda$ with k, we get $a^* \; J_0(k\rho a/z)$. This is a roughly cosine-like pattern with a maximum at the center, just as you would expect for an annular ring source.

Now, if you replace this ring with a disk, then you have to integrate $r\,J_0(kr)$ and that gives $r/k\,J_1(kr)$ and then you get a final result as shown on page 77 which is this:

$U(\rho) = \exp(jkz) \bullet 1/(j\lambda z) \bullet \exp[jk\rho^2/2z]\ (\pi a^2)\ 2\ J_1(ka\rho/z) / (ka\rho/z).$ // Airy

Since J1 is sine-like, this still has a peak in the center, and is then radially very sinc-like. As for the factor of 2, note that $2J_1(x)/x$ goes to 1 for small x, so the 2 is really due to the definition difference between Bessel and sinx/x or sinc(x) functions.

So the answer is that I have no simple intuitive way to explain why there are radial zeros and maxima in the Airy pattern. I see that if we put in a delta function ring aperture, we get $r\ J_0(kr)$, and I can then see that if we integrate this cosine-like function over a disk aperture dr, the areas will cancel at certain points (though not exactly periodic). This last argument is then very similar to what we said above for the sinc function as integral of a cosine. The harder problem here is understanding why the angular integral over a delta function annular ring gives a cosine-like result. The reason there are zeros in this result is this: as you move slowly away from the observation pattern center, you have a small change in the angular integral, and so you might expect to pass through points where this integral changes sign, and thus you would expect to pass through zeros.

Comments on the Airy formula above. The phase factors we understand perfectly as before. The double area is OK. On axis at $\rho = 0$ we get a value of $2\ J_1(ka\rho/z) / (ka\rho/z) = 1$ (so the two is gone there, note that unlike $\sin(x) \sim x$, we have $J_1(x) \sim x/2$.) Remember that we are always assuming an incident plane wave from the left in our computations, which is why we set $U(x,y) = 1$ inside the diffraction integral. How does this compare to the amplitude at pattern center:

amplitude at Airy center = $\text{Airy}(0) = (\pi a^2) / (\lambda z)$

As we go to very short wavelength λ, our original Kirchhoff approximation becomes better, since hole is large compared to λ so we can ignore edge effects. BUT, our original **Fresnel approximation** requires that

$\pi\ \rho^4 / (4\ \lambda\ z^3) << 1$ // says we can neglect phase error of next term in r01 expansion

This says we cannot really take the limit $\lambda = 0$. On axis, we really have to think of $\rho = a$ in this formula, so we would require that

$\pi\ a^4 / (4\ \lambda\ z^3) << 1.$ $a = 1$ cm, $\lambda = .5\ \mu$, $z = 1$ m => $1.5 \times 10^{-2} << 1$ which is OK

[Remember that the binomial next term really involves the maximum transverse distance, so even if you are on axis at the observation screen, you still have to consider positions on the aperture.] So combining the above Airy(0) expression and the Fresnel approximation requirement, we get

$\text{Airy}(0) << (2z/a)^2$. // seems to allow for a Airy(0) much larger than 1

Example: let a = 1 cm, let λ = .5 micron, let z = 1 m, then we get Airy(0) = 600 for an incident plane wave of amplitude 1. Can this really be true? (no, see below) The condition above becomes Airy(0) << 200*200 which is certainly met.

HOWEVER, the Airy formula also makes the **Fraunhofer approximation** which is this:

$(\pi\rho^2) << (\lambda z)$ / where ρ here refers to the aperture transverse

Thus, Fraunhofer tells us this:

$$\text{Airy}(0) = (\pi a^2) / (\lambda z) << 1$$

My conclusion here is that Fraunhofer diffraction really only applies in cases where the observation screen pattern is going to be very weak compared to the incident field. But Fresnel diffraction is more general. Unfortunately, Fresnel does not give the simple Fourier Transform. In my 1 mm pinhole experiment in the bathroom, the Fraunhofer condition requires z >> 18 feet, which was not really achieved. That is why you have to make very small pinholes and why you need very bright light, such as from a laser source. Also, you cannot take any limit of a Fraunhofer result (like Airy) and arrive at the geometric shadow limit. The reason is that in such a limit, the center intensity is going to be 1, but you see above that Airy(0) << 1.

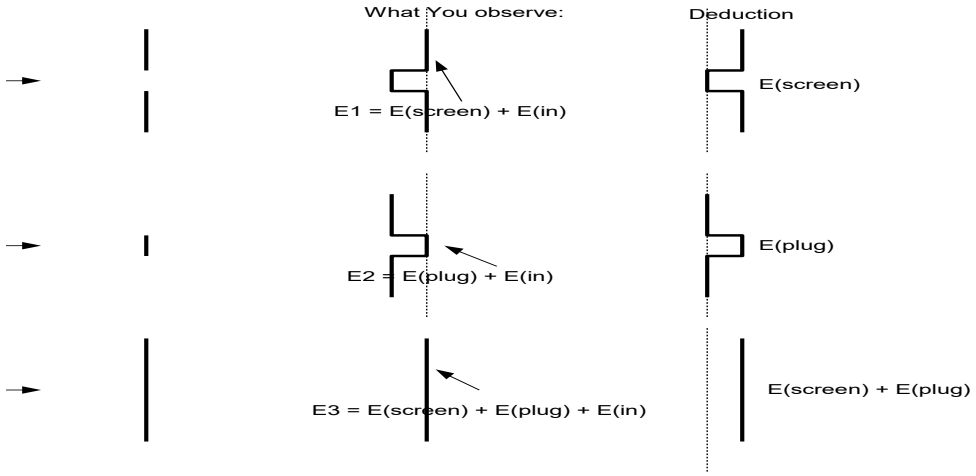
However, I can see that a geometric limit can fall within the Fresnel approximation.

The Geometric Limit. By arm-waving, Goodman claims that the kernel h(x,y) in 4-16 approaches $\delta(x)\delta(y)$ in the limit that z = 0, so that then 4-15 gives U(x,y) = the incident field. But we know this is just arm-waving and is not really true. First of all, the Fresnel formula is only correct in a large z limit so we can ignore all higher terms in the expansion, but here we are trying to apply the formula to z = 0! It is perhaps more useful to go all the way back say to Sommerfeld I as 3-41 and try to take this same limit there. At least you ought to have a better chance doing this! But even this formula is not right because we dropped a 1/r01 term to get it, but r01 will be arbitrarily small in the integration. Finally, a point in the aperture at z=0 really violates the entire derivation in terms of Green's Theorem. You certainly could ask what the result looks like for "very small z" and remain valid, and I am sure some book somewhere does this piece of work and shows you get the incident field. This sounds like one of those little ugly corners of physics that not many people want to probe into.

Nevertheless, we can accept Goodman's arm-waving and say that the Fresnel formula really can give you accurate results in the geometric shadow region. His argument is that most of the contribution to the pattern at (x,y) on the screen comes from a square patch on the aperture plane centered at (x,y) there, and of edge 4 sqrt(λz). Consider λ = .5 μ and z = 1 m so edge = 3 mm. So if you have a hole >> 3 mm, then in this case the patch fits anywhere in the hole, and you get the geometric case. In other words, if the aperture is much larger than this fictitious patch, then you get the geometric shadow everywhere except at the shadow boundary. This means that in the geometric bright central area, the amplitude will in fact be unity. The argument is this: if everything comes from that patch, then the field so generated should be the same as if there were no aperture at all (recall Kirchhoff's approx!).

In our floater case, suppose z = 1 mm just to pick a number. Then patch edge ~ 100 μ. If the floater is only 8 μ, then we are definitely NOT in the geometric limit in this case. Now at least we have a good rule of thumb for determining whether or not we are in the geometric limit.

Babinet and the disk. Here is the basic idea in terms of the pure geometric shadow limit:



The thing to understand is that E1, E2 and E3 are field patterns you see at the observation plane, whereas E(screen) is the field generated *by* the screen on the left having the aperture in it, and E(plug) is the field generated *by* the plug itself. The third case is the superposition of the first two. In all three cases, what you see on the observation plane includes the incoming field E(in). In Case 1, the screen must generate a field that exactly cancels E(in) away from the hole. In the second case, the plug must generate a negative square field as shown on the right, in order to give the total field observed in the center, that is, in order to give a shadow behind the plug. In the third case, we observe all black, so E(plug)+E(screen) must cancel the incoming E(in), which you can see it does. The main conclusions are these:

E(plug) + E(screen) + E(in) = 0.

E1 + E2 = Ein or E2 = Ein - E1

Note that in the geometric limit, there will in fact be a Poisson bright-spot in the middle of the shadow E2, and thus there will be a black spot in the middle of the aperture field E1, that would probably be quite hard to see. I have never heard this black spot mentioned, but I see it in the rough plots in my waveoptics pdf, at least I think I do. [see below]

Now we could redraw the above diagram in the Airy Fraunhofer limit, but Babinet would still exactly apply. We would then make this conclusion:

E1 = Airy / recall that even the central max of Airy is << Ein, see above!

E2 = Ein - Airy

We normally take Ein = 1 as unit incoming field, so then E2 = 1.0 - Airy.

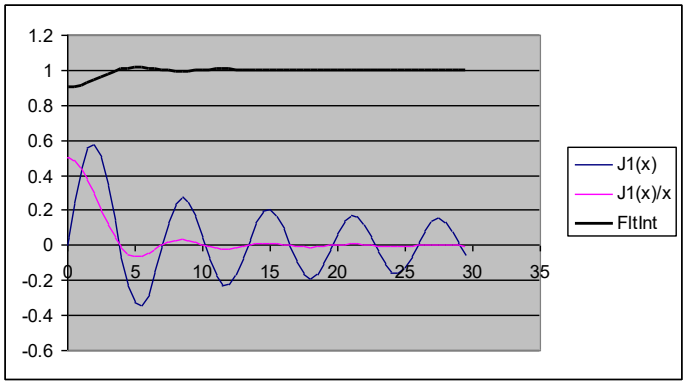
********** this little section is wrong, corrected below **********

This is the field in the circular stop case, so we expect to see a slight *diminution* of brightness in the shadow area. (This conclusion does *not* agree with the plots I have for the stop and hole in my wave optics PDF file, but maybe it is wrong too. The point of the particular slide is to illustrate Babinet, but it exactly does not do so!)

The intensity one would observe would be this:

$E2^2 = (1 - Airy)^2 = 1 - 2*Airy$ = intensity in far region behind a circular stop

where we ignore the $Airy^2$ extra term which will be very small. Here is an Excel plot of Airy:



Recall that amplitude at Airy center = Airy(0) = $(\pi a^2) / (\lambda z) << 1$, so I think we now have a real prediction for what a floater shadow should look like, assuming a floater is in fact acting like a circular opaque stop. The intensity on the retina should look like the top curve above, where I arbitrarily set Airy(0) =0 .1. You should see a slight circular disk which is darkest at the very center and the darkness gradually tapers away at the edge. You might be able to see a "white" ring around the edge since the curve does go slightly above the Ein value (which is 1 here) as shown in the plot above.

Problem with the above: I have not accounted for the PHASE of the Airy result! Continue on that tomorrow!

********** end of wrong section **********

Babinet done correctly including phase!

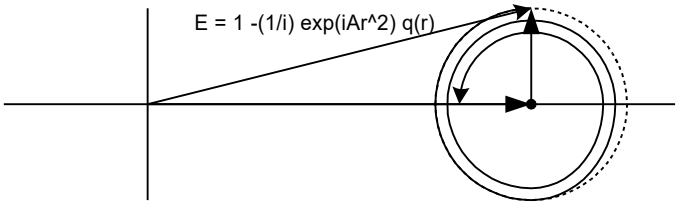
The phase of the incoming plane unit amplitude plane wave is just exp(ikz), so in the Babinet field difference we can factor this out of both Ein and Airy and we get this result:

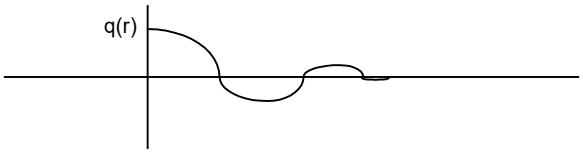
$E2 = e^{ikz} * [1 - (1/i) e^{iAr^2} q(r)]$ $A = k/2z$

where

$q(r) = B * 2 J_1(Cr)/(Cr)$ $B = (\pi a^2)/(\lambda z)$ $C = ka/z$

Remember that q(r) is a sinc like function, and B <<1 to be in the Fraunhofer limit. Looking at the form





of E2, ignoring the outside overall phase, what you see is a vector to the point "1" to which is added a second vector $i\, e^{iAr^{\wedge}2}\, q(r)$. As r increases from 0, this second vector which starts at 90 degrees starts to rotate to the left at an accelerating rate (since r^2). The magnitude of the second vector is |q(r)| which starts at value B at r=0 and shrinks and grows as that sinc-like function varies, but the magnitude never reaches this initial value of B again. So here is a Visio graphic to show the construction:

Without doing any more, you can SEE what |E(r)| is going to look like. At r = 0 it has a larger than 1 value, then it drops down to a smaller value, and then oscillates around until eventually it zeros in on value 1 because q(r) approaches 0, being sinc-like. The second smaller vector is going to follow a spiral pattern like that shown in the figure. The starting amplitude is $\sqrt{1+B^2}$ which then starts to decrease during the first turn. Depending on the values of the parameters, the amplitude could then reach an even larger value of 1 + B if it reaches the far side of the circle. So 1+B is an upper bound on the amplitude at all values of r, and 1-B is a lower bound.

Simple algebra gives this result which of course implies the same shape:

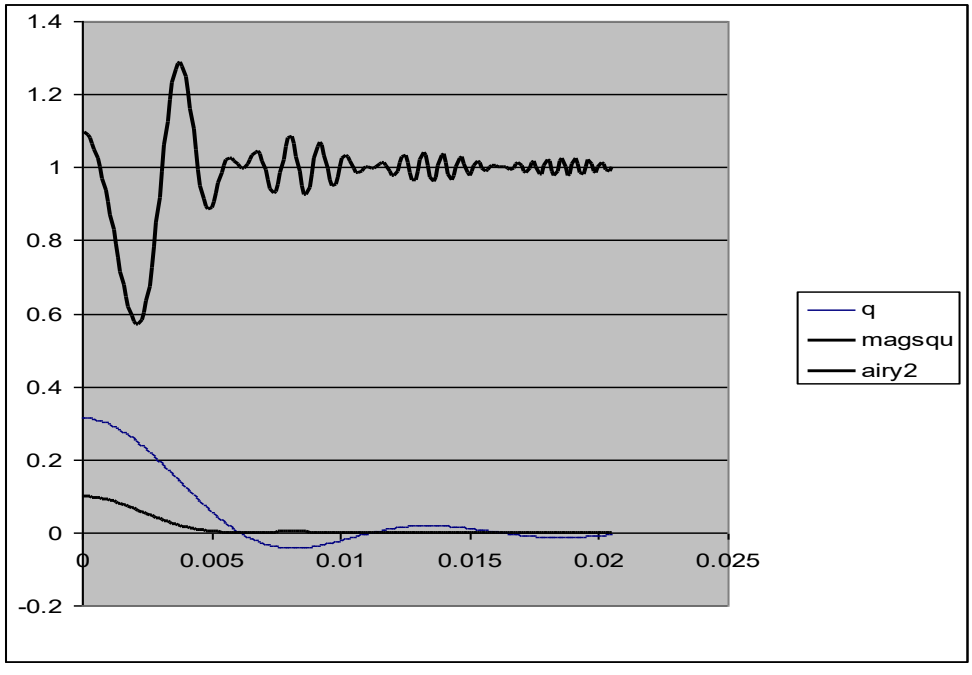
$$|E(r)|^2 = 1 - 2\sin(Ar^2)\, q(r) + q(r)^2$$

where you can see that $|E(0)|^2 = 1 + q(0)^2 = 1 + B^2$ which is in fact > 1.

So now we get a result which agrees with that PDF file. It says that in the diffraction pattern of a circular disk, the center is slightly brighter than the incident wave, the magnitude then drops off, and then oscillates toward 1, perhaps reaching a value larger than the central value. The central bright area is the remnant in the Fraunhofer limit of the Poisson bright spot in the geometric shadow in the Fresnel near-field case. Here is a specific Excel plot with the following values:

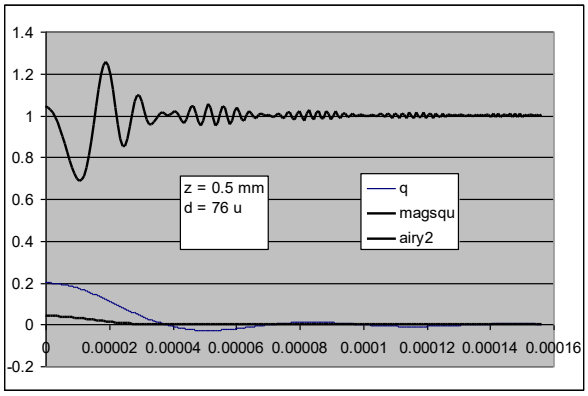
$\lambda = .5\ \mu$ a = 1 mm z = 20m

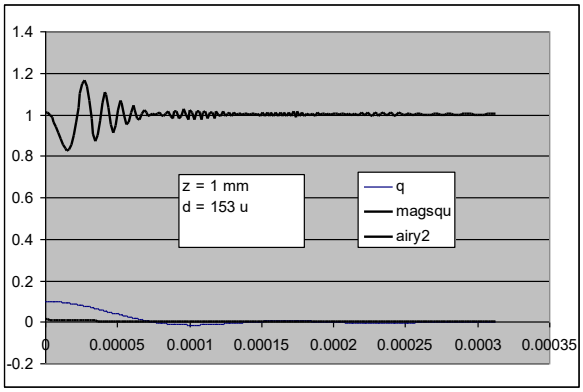
A	314159.3 minv2
B	0.314159 dimless
C	628.3185 minv
d	0.0122 lobe width

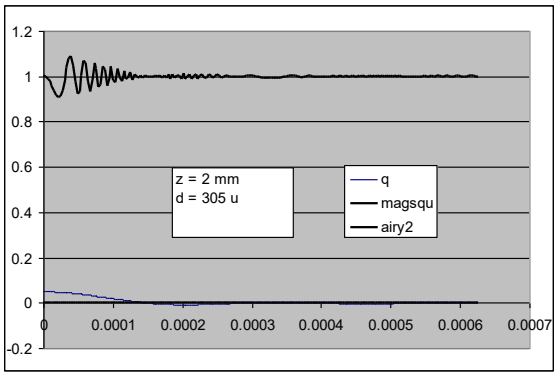


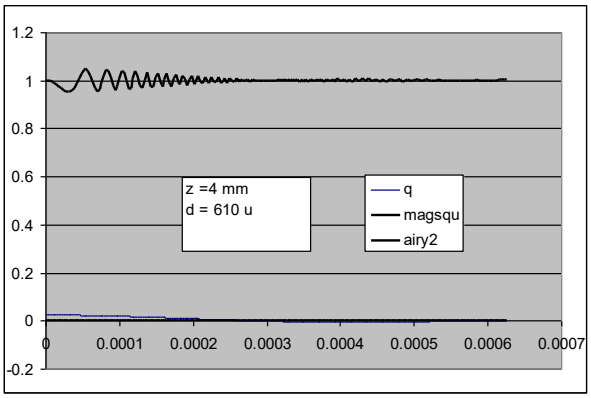
The thin curve is just the function q(r), sinc-like as noted, and q(0) = B = .3. The lower heavy curve is the Airy intensity pattern you would get from a *hole*, while the upper curve is the intensity pattern resulting from the *disk* of the same diameter according to Babinet, and these two intensities are properly scaled. We can see that the Airy pattern of the hole is fairly subdued, half-width is about 5 mm. In contrast, the pattern of the corresponding disk has a dramatic swing going from about .58 to 1.3 in intensity (this is the square of the field, so $(1\text{-}B)^2 = .49$ is the lower bound).

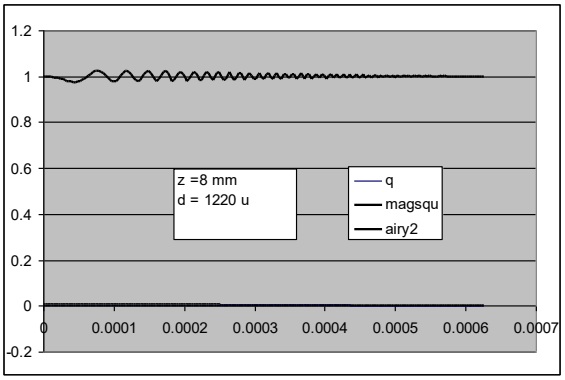
Here now are some plots using a = 4 μ, as for a bloated red blood cell of 8 μ diameter. What we vary here is the distance from the retina. The range is probably about 15 mm.











.

In the above sequence, you can see what happens as the floater moves away from the retina (we are still assuming an incident plane wave, ignoring lens effects, etc.). The diameter of the "disturbance" (which I see as a transparent shadowy thing) increases, but the amplitude of the disturbance becomes less and less. Also, the frequency slows down as z increases. I think in the z=8 mm case shown last, I would perceive nothing at all because the spacing between bright rings is about 25 μ and the amplitude is very small.
In general, you see that as the floater moves away from the retina, the amplitude decreases (as I have always imagined to be the case), and the radius increases. When the radius is only 8 μ for a super close-in floater, that is probably below spatial detection capability.

Resolution Capability of the Eye. Using my original floater picture for Dr. Call, with my left eye corrected with usual glasses, I can resolve stripes in the retina layer sketch out to about 36". There are about 5 stripes per 1/16", so 80 per inch, so spacing is 1/80 in. So, h = H x q/p = (1/80 in) x (1")/(36") = 350 μinch = 9 μ. Another experiment is to look at newspaper half-tone dots, I see them at about 1 foot away, they are also about 80 dots per inch, so this result gives 3X larger or about 27μ. A dot is harder to see that a line segment because a line segment can be picked up by some cones along its length and will perhaps be missed by others, and your brain will fill in the line. The dot really has to be seen by at least once cone cell. So let's go with this figure of about 27 μ as our best guess. A dot has to be 27 μ or larger on the retina, or we won't be able to see it, no matter now bright or dark it is.

So we can now finish our little theory. Blood cells that are less than 200 μ from the retina cast shadows too small to see. They first become visible perhaps at 250 μ, and they are quite distinct. These would be the sharp little black floaters that some of mine appear as, but today I don't seem to have any in this category (only right-eye ones are visible today). Perhaps the central white spots of some floaters are of this small size. My typical floater is about 1/16" diameter at 1 foot, which implies 132 μ as my original sheet showed, and this in turn suggests about 1 mm distant from the retina, again in agreement. I think the idea is that when floaters are 1 mm from the retina, they are perhaps maximally visible, and that is why the ones I can measure are this size. Maybe floaters that are 4 mm from the retina have insufficient amplitude to be visible. So maybe there is a soup of floaters everywhere at all distances from the retina, but I only see the ones that are in the 1-2 mm distance range.

Another twist on the theory might be that the floater cells are physically confined to a region perhaps 2 mm from the retina, only because they cannot diffuse into the vitreous gel. This is then the idea that there is a liquid layer between the vitreous and the retina due to aging.

Now where is that focal point? From the original pencil sheet, $q\text{-}f \sim q^2/p$. If something is 6 feet away, then we are talking $q\text{-}f = (1\text{ in})^2 /72\text{in} = (1/72)\text{ in} = 14$ milli-inches = 0.3 mm. So if I am looking at something at a distance of 6 feet (the floaters sloshing experiment while standing in the driveway), it could easily be the case that the floaters I see are out there at 2 mm and these are in fact "to the left" of the focal point, and that is why they slosh the way they do.

What is the effect of the lens on this discussion?

The difference is simple. Instead of having a plane wave impinging on our aperture or disk, we have a wave that has spherical wavefronts which are sucking down toward the focal point (assuming a plane wave is hitting the eye). The phase of these curved wavefronts has the following form:

$$\text{phase}(z,r) = e^{ikz}\ e^{-i(k/2d)r\hat{}2}$$

The lens causes this second factor! d is the distance from our aperture (floater) to the focal point -- it is the radius of the converging spheres. Just after the lens, this formula has d = f, but we want to put our screen at some arbitrary distance d from the focal point, perhaps even a negative d which would apply to the right of the focal point. But think back to d = f for the moment. As you move with z fixed transversely away from r=0 (thick part of the lens), you get *less* delay, so the phase is *earlier*. That is why there is a minus sign in the second phase. It is as if you moved to the left in z (smaller z) in the first factor. Obviously this only applies for small r and all that stuff. In other words, the lens variable thickness causes this phase delay pattern. So, if we try to do Fresnel diffraction starting with a "field on the left" that is this lens-processed plane wave, then we have to make this change in formula 4-17:

$$U(\xi, \eta) \rightarrow U(\xi, \eta) * \exp\left[-j (k/2d) (\xi^2 + \eta^2) \right] * (f/d)$$
$$= 1 * \exp\left[-j (k/2d) (\xi^2 + \eta^2) \right] * (f/d)$$

The phase factor was just discussed, and yes it is singular at the focal point d = 0. Similarly, the brightness gain factor (f/d) is singular at d=0.

The new "Fraunhofer condition" in this situation would be as follows:

$$(\pi a^2/\lambda) [1/z - 1/d] << 1$$

In our eye situation, we are interested in z ~ 2 mm or so. If the focal point is at .3 mm from the retina as in the 6 foot example above, then d = 1.7 mm -- our floaters are to the left of the focal point. For these numbers, we see that the Fraunhofer condition is "even better met" because the 1/d term cancels a bit of the 1/z term. As the eye focuses at 1 foot, perhaps focal point is then 3 mm from the eye, and now floaters are 1 mm to the right so we have d = - 1 mm. Then the two terms are additive, but they are still in the same ball park and we neglect both.

So yes, we do have a problem if the floaters get right close to the focal point -- a problem in *understanding* what happens, that is. But I think the answer is that nothing dramatic happens because as I change my eye focal point while studying floaters, nothing dramatic occurs. There is no dramatic change in floater appearance -- no magic point where floaters suddenly block the entire view, say. So my claim is that if one puts a circular aperture at some point to the right of the lens, we can really neglect the wavefront curvature in computing the diffraction field, and we will get our same Airy pattern, assuming as usual that we are in that appropriate limit.

Now when we come to the floater qua disk (as opposed to qua aperture), there will be a difference. That is because Babinet tells us in this case that:

$$E2 = e^{ikz} * \left[\underline{(f/d')*e^{iDr\hat{}2}} - (1/i)\, e^{iAr\hat{}2}\, q(r)*\underline{(f/d)} \right] \qquad A = k/2z \qquad D = k/2d'$$

valid for floaters not too close to d = 0, where

$$q(r) = B * 2 J_1(Cr)/(Cr) \qquad B = (\pi a^2)/(\lambda z) \qquad C = ka/z$$

and we have underlined the "new" term that used to be 1, and a new factor in the second term. This new first term is what the "incident" wave looks like at the retina, and this is what is going to interfere with the Airy signal. d is the distance from the disk to the focal point (positive if screen to the left of the focal

point), and d' is the distance from the focal point to the retina. With either sign of d, we have z = d + d'. So rewrite the thing like so:

$$E2 = e^{ikz} * (f/d') * e^{iDr\hat{}2} [\ 1\ -\ (1/i)\ e^{i(A-D)r\hat{}2}\ q(r)*(d'/d)\ ]$$

So at least we now have our disk field formula in a form we can understand. There are two effects of the lens on our result. First, the length of the short second vector is scaled by (d'/d). But we assumed that d is not too small already, and d' is also some reasonable number like .3 - 3 mm. Number d might range from -2 mm to + 2 mm, and we have to exclude the small-d region we don't well understand. So yes, the second factor which increases the diameter of the q(r) circle could be important. The other difference is that the second vector phasor rotates with a different speed, A-D instead of A. Both these factors will cause our results to be different than in the case of no lens! Still, we know what the general appearance of the result is going to be.

So really, the only remaining mystery is now to handle the case that d ~ 0. In this limit, the magnitude of the aperture diffraction ought to approach infinity, since the incident light is infinitely bright at the focal point. But of course it is not going to be a perfect lens and the focal point will be spread out somewhat and thereby fix this problem. The presence of a non-ignorable phase in the Fresnel integral just means that we don't get the simple Airy pattern, but we get something that Fresnel puts out. So the bottom line is that I don't think anything dramatic happens.

Let's call this now a solved problem and move on! The following subjects: Fresnel Lens, but I read about that, it is the lighthouse lens to increase capture efficiency in all directions. That is, it captures vertical light and gets it going horizontally.

The Frensnel Zones still are a mystery to me. I think they are just a way to break up a spherical wavefront surface if you want to think about approximate cancellations at an observation point. Perhaps just a curiosity like the Cornu Spiral. [see new notes on this elsewhere.]